Schätzen der Klassenzugehörigkeitswahrscheinlichkeit zur Definition des Arbeitsbereichs von chemieinformatorischen Klassifikationsmodellen

Miriam Mathea

Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Schätzen der Klassenzugehörigkeitswahrscheinlichkeit zur Definition des Arbeitsbereichs von chemieinformatischen Klassifikationsmodellen

Schätzen der Klassenzugehörigkeitswahrscheinlichkeit zur Definition des Arbeitsbereichs von chemieinformatischen Klassifikationsmodellen

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Miriam Mathea

aus Paderborn

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet
über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen: Cuvillier, 2018
Zugl.: (TU) Braunschweig, Univ., Diss., 2017

1. Referent: Professor Dr. Knut Baumann

2. Referent: Professor Dr. Hermann Wätzig

Eingereicht am: 23.10.2017

Mündliche Prüfung (Disputation) am: 21.12.2017

Druckjahr 2018

Dissertation an der Technischen Universität Braunschweig,

Fakultät für Lebenswissenschaften

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2018
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft.

ISBN 978-3-7369-9722-6
eISBN 978-3-7369-8722-7

Veröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Mathea M, Klingspohn W, Baumann K. Chemoinformatic Classification Methods and their Applicability Domain. *Molecular Informatics*. 2016; 35;160–180. doi:10.1002/minf.201501019

Klingspohn W, Mathea M, ter Laak A, Heinrich N, Baumann K. Efficiency of different measures for defining the applicability domain of classification models. *Journal of Cheminformatics*. 2017;9:44. doi:10.1186/s13321-017-0230-2

Tagungsbeiträge

Scandinavian Symposium on Chemometrics (SSC14), Sardinia, Italy, 14-17 June 2015 Calibration of Class Probability Estimates for Hedging Chemoinformatic Classifiers, M. Mathea, K. Baumann

11th German Conference on Chemoinformatics , November 8– 10, 2016, Fulda Class Probability Estimates for Defining an Applicability Domain, M. Mathea, K. Baumann

Posterbeiträge

Chemoinformatics Strasbourg Summer School 2014, 23-27 June 2014 Comparison of Different Measures for the Domain of Applicability of Classification Models, M. Mathea, W. Klingspohn, K. Baumann

11th German Conference on Chemoinformatics , November 8– 10, 2016, Fulda Class Probability Estimates for Defining an Applicability Domain, M. Mathea, K. Baumann

International PhD students/Postdocs meeting 2016 of the German Pharmaceutical Society (DPhG) Aachen, 16.-18.03.2016 Calibration of Class Probability Estimates for Defining an Applicability Domain, M. Mathea, K. Baumann

12th German Conference on Chemoinformatics, November 6– 8, 2016, Fulda Ensembles Approaches for Estimating Class Probability Estimates, M. Mathea, K. Baumann

Danksagung

Zu Beginn möchte ich Prof. Dr. Knut Baumann für die Betreuung der Arbeit, die Vergabe des spannenden Themas und für seine tatkräftige Unterstützung danken. Darüber hinaus möchte ich Prof. Dr. Hermann Wätzig für die Übernahme des Zweitgutachtens, sowie Prof. Dr. Ute Wittstock für die Übernahme der Leitung der Prüfungskommission, danken.

Gleichwohl gilt mein Dank allen Mitgliedern des „allerbesten“ Arbeitskreises. Angefangen bei Waldemar: Vielen, lieben Dank für die tolle Projektzusammenarbeit und Unterstützung! Des Weiteren bin ich Shanthy sehr dankbar für ihre mentale Unterstützung und die unermüdliche Verbreitung von guter Laune, welche die Arbeitstage sehr angenehm gestaltet hat. Außerdem möchte ich mich bei meinem Büro Fantland (Max und Frede) für die lustige, gemeinsame Zeit, die Ernennung zur Kaiserin, sowie die Duldung der Elefantensammlung, bedanken. Darüber hinaus gilt mein Dank Anke und Jessi, sie hatten immer ein offenes Ohr und viel Tee für mich. Zusätzlich möchte ich mich bei den ehemaligen AK-Mitgliedern Sabrina und Désirée ebenfalls für ihren Beistand und ihre Hilfe bedanken.

Darüber hinaus bin ich dem Team des 5. Semester Praktikums, insbesondere Annika, für die tolle Zeit und viel Pizza dankbar. Dasselbe gilt auch für Anne vom Chemometrik-Praktikum. Mein Dank gilt außerdem Frank und Britta, sowie allen Beteiligten der Frühstücksrunden, welche maßgeblich für ein gutes Arbeitsklima verantworlich waren.

Außerdem möchte ich noch der Donnerstagsrunde der Biologen sowie Anja, Dörte, Elin und Mona für die schöne Zeit an der Uni mit Kaffee und Kuchen, außerhalb von der Praktikumsbetreuung danken.

Während der gesamten Promotionszeit haben mich meine Freunde, Volker, meine Famile und Volkers Familie immer sehr unterstützt, daher gilt ihnen ebenfalls mein besonderer Dank.

Abkürzungsverzeichnis

Abkürzung	Deutsch	Englisch
Acc	Korrektklassifizierungsrate	Accuracy
AB	Anwendungsbereich	Applicability Domain
Brier	Brier-Score	Brier-Score
Brier_Log	Brier-Score nach Kalibrierung mittels logistischer Regression	Brier-Score after calibration with logistic regression
Cal	Trainingsdatenpartition zur Kalibrierung	calibration set
CP	Conformal Prediction	Conformal Prediction
CV	Kreuzvalidierung	Cross-Validation
DM	Distanz zum Modell	Distance to Model
Elastic Net	Elastic Net	Elastic Net
FN	Falsch Negativ	False Negative
FP	Falsch Positiv	False Positive
ICP	Inductive Conformal Prediction	Inductive Conformal Prediction
k-CV	k-fache Kreuzvalidierung	k-Fold-Cross-Validation
KNN	k-Nächste Nachbarn	k-Nearest Neighbor
Lasso	Least Absolute Shrinkage and Selection Operator	Least Absolute Shrinkage and Selection Operator
LDA	Lineare Diskriminanz Analyse	Linear Discriminant Analysis
LMO	Lass-mehrere-Objekte-heraus-Kreuzvalidierung	Leave-Multiple-Out-Cross-Validation
LOO-CV	Lass–ein–Objekt-heraus-Kreuzvalidierung	Leave-One-Out-Cross-Validation
MAE	Mittlerer Absoluter Fehler	Mean Absolute Error
MAE_w	Gewichteter Mittlerer Absoluter Fehler	Weighted Mean Absolute Error
MAE_Cu	Mittlerer Absoluter Fehler nach Caruana	Mean Absolute Error after Caruana
MAE_Log	Mittlerer Absoluter Fehler nach	Mean Absolute Error after calibra-

	Kalibrierung mittels logistischer Regression	tion with logistic regression
MAE_w_Log	Gewichteter Mittlerer Absoluter Fehler nach Kalibrierung mittels logistischer Regression	Weighted Mean Absolute Error after calibration with logistic regression
MAE_Cu_Log	Mittlerer Absoluter Fehler nach Caruana nachdem mittels logistischer Regression kalibriert wurde	Weighted Mean Absolute Error after Caruana after calibration with logistic regression
MICP	Mondrian off-line inductive conformal prediction	Mondrian off-line inductive conformal prediction
MOE	Molecular Operating Environment	Molecular Operating Environment
MSE	Mittlerer Quadratischer Fehler	Mean Squared Error
MSE_w	Gewichteter Mittlerer Quadratischer Fehler	Weighted Mean Squared Error
MSE_Cu	Mittlerer Quadratischer Fehler nach Caruana	Mean Squared Error after Caruana
MSE_Log	Mittlerer Quadratischer Fehler nach Kalibrierung mittels logistischer Regression	Mean Squared Error after calibration with logistic regression
MSE_w_Log	Gewichteter Mittlerer Quadratischer Fehler nach Kalibrierung mittels logistischer Regression	Weighted Mean Squared Error after calibration with logistic regression
MSE_Cu_Log	Mittlerer Quadratischer Fehler nach Caruana nachdem mittels logistischer Regression kalibriert wurde	Weighted Mean Squared Error after Caruana after calibration with logistic regression
MW	Mittelwert	Mean
NN	Neuronale Netze	Neural Networks
NBC	Naive Bayesian Classifier	Naive Bayesian Classifier
PCA	Hauptkomponentenanalyse	Principal Component Analysis
PCR	Hauptkomponentenregression	Principal Component Regression
PLS	Partial Least Squares Regression	Partial Least Squares Regression

PLSDA	Partial Least Squares Discriminant Analysis	Partial Least Squares Discriminant Analysis
PSS	Bestrafte-Summe-Quadrierter-Abweichungen-Kriterium	Penalized-Sum-of-Squares-Criterion
Pt	geeignete Trainingsdatenpartition	proper training set
QF	Quadratischer Fehler	Squared Error
QSAR	Quantitative-Struktur-Aktivitäts-Beziehung	Quantitative-Structure-Activity-Relationship
QSPR	Quantitative-Struktur-Eigen-schafts-Beziehung	Quantitative-Structure-Property-Relationship
RBF	Radial Basis Funktion	Radial Basis Function
RF	Random Forests	Random Forests
RFR	Random Forest Regression	Random Forest Regression
Ridge	Ridge Regression	Ridge Regression
RSS	Summe quadrierter Residuen	Residual Sum of Squares
Sens	Sensitivität	Sensitivity
Spec	Spezifität	Specificity
SPLS	Sparse Partial Least Squares Re-gression	Sparse Partial Least Squares Re-gression
STD	Standardabweichung	Standard Deviation
SVM	Support Vector Machines	Support Vector Machines
SVR	Support Vector Regression	Support Vector Regression
TN	Richtig Negativ	True Negative
TP	Richtig Positiv	True Positive

Inhaltsverzeichnis

1 Theoretische Grundlagen

1.1 Quantitative-Struktur-Wirkungs-Beziehungen

In der modernen Arzneistoffentwicklung werden u.a. computergestützte Verfahren für das Wirkstoffdesign verwendet. Diese Verfahren werden auch als *In-Silico*-Techniken bezeichnet. Hierzu gehört auch die Analyse von quantitativen Struktur-Aktivitäts-Beziehungen (engl.: Quantitative-Structure-Activity Relationships (QSAR)). QSAR Analysen haben das Ziel entweder die biologische Aktivität einer Substanz selbst oder bestimmte Faktoren, die die Aktivität bestimmen, vorherzusagen [1]. Hierbei wird die Struktur-Aktivitäts-Beziehung mit Hilfe einer mathematischen Funktion ausgedrückt. Wenn eine solche Analyse durchgeführt werden soll, werden zunächst die biologischen Aktivitäten von verwandten Molekülen sowie deren chemische Struktur benötigt. Damit die chemische Struktur der Analyse zugänglich gemacht werden kann, werden Deskriptoren berechnet (siehe 2.2). Danach wird der funktionelle Zusammenhang zwischen den Moleküldeskriptoren und der biologischen Aktivität modelliert.

Die QSAR Analyse nahm vermutlich ihre Anfänge mit der Publikation zweier schottischer Pharmakologen (Crum-Brown und Fraser), welche 1868 zu der Erkenntnis kamen, dass die physiologische Aktivität ϕ einer Substanz eine Funktion ihrer chemischen Konstitution C sei.

$$\phi = f(C)$$

Im Jahre 1964 knüpften Hansch und Fujita [2] sowie Free und Wilson [3] an diese Idee an, indem sie die biologische Aktivität und die physikalisch-chemischen, sowie strukturellen Eigenschaften von Molekülen korrelierten. Heutzutage ist die QSAR Analyse eine gut etablierte Methode, welche vielfältig angewendet wird, nicht nur zur Vorhersage der biologischen Aktivität, sondern auch zur Vorhersage anderer Moleküleigenschaften im Rahmen quantitativer Struktur-Eigenschafts-Beziehungen (engl.: Quantitative Structure-Property Relationships (QSPR)). Durch die Möglichkeit Eigenschaften von noch nicht vorhandenen Molekülen vorherzusagen, lässt sich gegebenenfalls der zeit- und kostenintensive Syntheseaufwand reduzieren bzw. effizienter in eine bestimmte Richtung lenken (Leitstrukturoptimierung) [4, 5,6].

1.2 Moleküldeskriptoren

1.2.1 Einleitung

Die Analyse der strukturellen Information von Molekülen ist mit der Hilfe von Moleküldeskriptoren möglich. Hierbei handelt es sich um eine numerische Repräsentation des Moleküls. Deskriptoren können einerseits das Ergebnis standardisierter Experimente umfassen, zum Beispiel physikochemische Eigenschaften repräsentieren oder sie sind das Ergebnis eines standardisierten Algorithmus. Es gibt folglich sehr viele unterschiedliche Moleküldeskriptoren, die für verschiedene Anwendungsgebiete mehr oder weniger gut geeignet sind. Eine sehr ausführliche und umfassende Übersicht wurde von Todeschini verfasst [7].

Moleküldeskriptoren lassen sich nach ihrer Dimensionalität in unterschiedliche Klassen einteilen. 1D-Deskriptoren können beispielsweise einfache Eigenschaften wie das Molekulargewicht oder aber auch die Anzahl bestimmter Atome oder Bindungen kodieren. Die populärsten Deskriptoren sind die 2D-Deskriptoren und die 3D-Deskriptoren, welche zusätzlich die Topologie bzw. die Konformation des Moleküls mit einbeziehen. Darüber hinaus gibt es auch 4D- und höher dimensionale Deskriptoren [5].

1.2.2 Fingerabdruck-Deskriptoren (engl.: Fingerprints)

Die 2D-Fingerabdruck-Deskriptoren gehören zu den am weitesten verbreiteten Deskriptoren. Es sind Vektoren [8, 3, 9–11], welche ein Molekül bezüglich der An- oder Abwesenheit und/oder der Frequenz bestimmter Substrukturen charakterisieren. So kann beispielsweise die Anwesenheit einer Hydroxylgruppe mit 1 für anwesend oder 0 für abwesend gekennzeichnet werden (Abbildung 1).

Abbildung 1: Ausschnitt der Erzeugung eines Vektors. Dieser Vektor wird auch als molekularer Fingerabdruck bezeichnet. Die chemischen Eigenschaften des Moleküls werden numerisch repräsentiert, in diesem Beispiel wird die Anwesenheit der Hydroxylgruppe durch eine 1 im Vektor gekennzeichnet.

1.2.3 Topologische Deskriptoren

Die topologischen Deskriptoren gehören ebenfalls zu den 2D-Deskriptoren. Sie sind weit verbreitet und werden vielfältig angewendet. Topologische Deskriptoren kodieren chemische Verknüpfungsinformationen von Molekülen. Diese Verknüpfungsinformationen, wie zum Beispiel die Verknüpfungsart (auch Konnektivität genannt) oder die Größe eines Rings, lassen sich aus der Strukturformel ableiten [12, 7, 13, 14].

Da 2D-Deskriptoren keine Informationen über die genaue räumliche Anordnung der Moleküle (die sog. Konformation) benötigen, sind sie oftmals beliebter als 3D-Deskriptoren. In vielen Fällen ist die Konformation der aktiven Verbindung unbekannt und somit ist eine Vielzahl an vorbereitenden und z.T. rechenintensiven Schritten notwendig, bevor mit diesen 3D-Deskriptoren gearbeitet werden kann [4].

1.3 Einführung in die Multivariate Datenanalyse

1.3.1 Einleitung

Im letzten Kapitel (1.2) wurde die Funktion von Deskriptoren beschrieben. Hierauf wird nun aufgebaut. Es wird beispielhaft angenommen, dass die Bioaktivität von verschiedenen Molekülen an einer bestimmten Zielstruktur gemessen wurde und zu jedem Molekül jeweils Deskriptoren berechnet wurden. Nun wird die Frage gestellt, wodurch die Bioaktivität beeinflusst wird und ob diese gegebenenfalls modelliert werden kann, um die Bioaktivität für ein unbekanntes Molekül vorhersagen zu können. Es gibt verschiedene Parameter von denen die Bioaktivität abhängen kann, beispielsweise die Molekülgröße oder die An- oder Abwesenheit bestimmter funktioneller Gruppen (alle Spalteneinträge des Deskriptors). In diesem Beispiel wird die Bioaktivität als abhängige oder beobachtete Variable $\boldsymbol{y}$ bezeichnet und die Spaltennamen des Deskriptors als unabhängige Variable $\boldsymbol{x}$ oder unabhängige Variablen X. Für p verschiedene unabhängige Variablen gilt: $X = (\boldsymbol{x_1}, \boldsymbol{x_2} \dots \boldsymbol{x_p})$. Wenn nun angenommen wird, dass $\boldsymbol{y}$ und X voneinander abhängen, kann dies folgendermaßen ausgedrückt werden:

$$\boldsymbol{y} = f(X) + \boldsymbol{e}.$$

Bei f handelt es sich um eine unbekannte Funktion von *X,* bei $\boldsymbol{e}$ um einen zufälligen Fehlerterm. Dieser Fehlerterm ist unabhängig von *X* und hat einen Mittelwert von Null. Die Funktion f kann auch mehr als nur einen Parameter miteinbeziehen. Das Ziel ist es *f* zu schätzen. In dieser Arbeit werden Vektoren mit einem kleinen, **fetten**, *kursiven* Buchstaben, Matrizen mit einem großen, *kursiven* Buchstaben und Skalare mit einem kleinen, *kursiven* Buchstaben gekennzeichnet.

Wenn beispielsweise nur *X* bekannt ist und $\boldsymbol{y}$ unbekannt ist und die Annahmen für den Fehlerterm (zufällig, unabhängig von *X*, Mittelwert ist Null) zutreffen, kann $\boldsymbol{y}$ durch:

$$\widehat{\boldsymbol{y}} = \hat{f}(X)$$

vorhergesagt werden. $\hat{f}$ repräsentiert die Schätzfunktion für f und $\widehat{\boldsymbol{y}}$ repräsentiert die Vorhersage für $\boldsymbol{y}$. Die Richtigkeit der Vorhersage von $\boldsymbol{y}$ hängt wesentlich von zwei Größen ab. Diese werden als reduzierbarer Fehler und nicht-reduzierbarer Fehler bezeichnet. Im Allgemeinen wird $\hat{f}$ keine perfekte Schätzfunktion für f sein. Dieser Fehler ist reduzierbar, weil die Richtigkeit von $\hat{f}$ potentiell durch unterschiedliche Techniken verbessert werden kann. Aber auch wenn f perfekt geschätzt werden würde, wären noch nicht alle Fehler beseitigt. Deshalb ist $\boldsymbol{y}$ auch eine Funktion von $\boldsymbol{e}$, welche per Definition nicht durch *X* vorhergesagt werden kann. Variabilität assoziiert mit $\boldsymbol{e}$ beeinträchtigt die Präzision der Vorhersage. Dieser Fehler wird auch nicht-reduzierbarer Fehler genannt. Die Größe $\boldsymbol{e}$ kann auch unbestimmte oder ungemessene, abhängige Variablen enthalten, welche nützlich wären um $\boldsymbol{y}$ zu bestimmen. Würden diese bekannt oder messbar sein, könnte der Fehler reduziert werden [15, 16].

$$E(\boldsymbol{y} - \widehat{\boldsymbol{y}})^2 = E[f(X) + \boldsymbol{e} - \hat{f}(X)]^2$$

$$= [f(X) - \hat{f}(X)]^2 \quad + \quad Var(\boldsymbol{e})$$

reduzierbar nicht-reduzierbar

Der nicht-reduzierbare Fehler wird immer eine obere Grenze für die Präzision von Vorhersagen vorgeben, welche in der Praxis fast immer unbekannt ist.

Es ist allerdings nicht immer das Ziel Vorhersagen für $\boldsymbol{y}$ zu tätigen, in manchen Fällen soll auch die Beziehung zwischen *X* und $\boldsymbol{y}$ untersucht werden, um zu verstehen, wie sich

$\boldsymbol{y}$ als Funktion von $\boldsymbol{x_1}, \boldsymbol{x_2} \ldots \boldsymbol{x_p}$ verändert. Es kann beispielsweise auch von Interesse sein lediglich einige wichtige Variablen zu identifizieren.

An dieser Stelle stellt sich natürlich die Frage, wie die unbekannte Funktion f geschätzt werden kann. Hierfür gibt es unterschiedliche lineare und nicht-lineare Ansätze. Generell haben diese Methoden bestimmte Charakteristiken, nach welchen sie unterschieden werden. Die meisten Methoden lassen sich entweder in die Gruppe der parametrischen oder in die Gruppe der nicht-parametrischen Methoden einteilen [15, 16].

1.3.2 Datenvorbehandlung

Häufig sind unterschiedliche Variablen nicht miteinander vergleichbar, da sie auf verschiedenen Skalen gemessen wurden. Mathematische Funktionen können aber sensibel für solche Unterschiede sein und diese mit modellieren.

Die Datenmatrix der unabhängigen Variablen *X* wird als **zentriert** bezeichnet, wenn von jedem Variablenvektor $\boldsymbol{x}$ der Mittelwert berechnet wird und der Mittelwertvektor schließlich von der Rohmatrix subtrahiert wird. Folglich wird von jedem Element von *X* sein entsprechender Spaltenmittelwert abgezogen.

Die Datenmatrix *X* wird als **autoskaliert** bezeichnet, wenn von jedem Variablenvektor $\boldsymbol{x}$ die Standardabweichung berechnet wird und jede Variable der zentrierten Matrix durch die zugehörige Standardabweichung geteilt wird. Wenn anstelle der zentrierten Matrix die Rohmatrix verwendet wird, so wird die Datenmatrix als **skaliert** bezeichnet.

Wenn die Datenmatrix sowohl zentriert als auch skaliert wurde, wird sie als autoskaliert oder **z-transformiert** bezeichnet.

1.3.3 Parametrische Methoden

Bei den parametrischen Methoden wird als erstes eine Annahme über die Gestalt von f gemacht. Beispielsweise wäre eine einfache Annahme, dass f linear in *X* ist.

$$f(X) = b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \cdots + b_p\boldsymbol{x_p}$$

Das Schätzproblem wurde durch die Annahme vereinfacht und somit müssen nur die Koeffizienten (auch Parameter genannt) $p + 1$: $b_0, b_1, b_2 \ldots b_p$ geschätzt werden. Nun

werden Trainingsdaten/Moleküle benötigt um das Modell zu trainieren. Es müssen $b_0, b_1 \dots b_p$ geschätzt werden, wobei Werte gefunden werden sollen, sodass:

$$\boldsymbol{y} \approx b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \cdots + b_p\boldsymbol{x_p}.$$

Die geläufigste Methode um das Modell anzupassen wird als "Methode der kleinsten Quadrate" bezeichnet. Dieser Ansatz reduziert das Problem f zu schätzen darauf eine Reihe von Parametern zu schätzen. Die möglichen Nachteile sind, dass das Modell welches gewählt wird, normalerweise nicht das wahre f sein wird und je weiter es davon entfernt liegt, desto dürftiger wird die Schätzung. Eine Lösung hierfür wäre es flexiblere Modelle zu wählen, welche sich verschiedenen, möglichen Formen von f anpassen können, allerdings müssten hierfür mehr Parameter geschätzt werden. Ein komplexeres Modell neigt leichter zu einer „Überanpassung". Dieses Phänomen wird später noch einmal detaillierter betrachtet [15, 16].

1.3.4 Nichtparametrische Methoden

Nichtparametrische Methoden machen keine ausdrücklichen Annahmen über die Form von f. Stattdessen streben sie eine Schätzung von f an, welche die Abweichung der Schätzwerte und der Trainingsdaten minimiert. Dieser Ansatz hat einen großen Vorteil gegenüber den parametrischen Methoden. Dadurch, dass Annahmen über die Form von f vermieden werden, können nichtparametrische Methoden, durch eine größere Spanne an möglichen Formen, potentiell eine genauere Anpassung an f ermöglichen. Ein großer Nachteil jedoch ist, dass, solange das Problem f zu schätzen nicht auf eine kleine Anzahl Parameter reduziert werden kann, eine größere Anzahl an Daten/Molekülen benötigt wird um f genau zu schätzen [15, 16].

1.3.5 Das Dilemma zwischen Vorhersagegenauigkeit und Interpretierbarkeit

Allgemein lässt sich sagen, dass bei Erhöhung der Flexibilität eines Modells, sich die Interpretierbarkeit erniedrigt. In einigen Fällen würde ein unflexibles Modell bevorzugt werden. Wenn beispielsweise Interesse am Zusammenhang zwischen $\boldsymbol{y}$ und X besteht, ist es vorteilhafter ein leicht zu interpretierendes Modell (z.B. lineares Modell mit wenig Parametern) vorliegen zu haben. Im Gegensatz dazu wären flexible Ansätze weniger geeignet, da in diesem Fall Zusammenhänge mit einzelnen Variablen und $\boldsymbol{y}$ nur schwer zu erkennen sind. Auch wenn ausschließlich Interesse an der Vorhersage besteht, wäre

es trotzdem nicht immer sinnvoll die flexibelste Methode zu wählen aufgrund der Problematik der Überanpassung [15, 16].

1.4 Regression

1.4.1 Einfache Lineare Regression

Die Einfache Lineare Regression modelliert die abhängige Variable $\boldsymbol{y}$ mit nur einer einzigen unabhängigen Variablen $\boldsymbol{x}$. Hierbei wird angenommen, dass ein linearer Zusammenhang zwischen $\boldsymbol{x}$ und $\boldsymbol{y}$ besteht.

$$\boldsymbol{y} = b_0 + b_1\boldsymbol{x} + \boldsymbol{e}$$

Bei $b_0$und $b_1$handelt es sich um die Koeffizienten und bei $\boldsymbol{e}$ handelt es sich um den Gesamtfehler des Modells. Nun wird ein Teil der Moleküle des Datensatzes, die sogenannte Trainingsdatenpartition, benutzt um die Koeffizienten $\widehat{\boldsymbol{b}}$ zu schätzen. Danach kann eine Vorhersage für ein zukünftiges ungesehenes Molekül x_0 gemacht werden.

$$\widehat{y_0} = \widehat{b_0} + \widehat{b_1}x_0$$

Das Ziel ist es, die Koeffizienten so zu schätzen, dass die resultierende Gerade möglichst nah an den vorhandenen Punkten verläuft. Es gibt unterschiedliche Verfahren um zu messen, was denn eigentlich nah ist. Die wohl geläufigste Methode ist die bereits erwähnte „Methode der kleinsten Quadrate“. Bei dieser Methode werden zunächst die Residuen berechnet ($\boldsymbol{y} - \widehat{\boldsymbol{y}}$) und danach werden diese quadriert und aufsummiert. Schließlich werden die Koeffizienten $\widehat{\boldsymbol{b}}$ so ausgewählt, dass die Summe der quadrierten Residuen (engl. Residual Sum of Squares (RSS)) minimiert wird.

1.4.2 Modellvalidierung

Im Rahmen der Regression wird meist der **Mittlere Quadratische Fehler (engl.: Mean Squarred Error (MSE))** verwendet um die Leistungsfähigkeit einer Methode zu beurteilen.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x_i}))^2$$

Falls der vorhergesagte Vektor $\boldsymbol{y} = (f(\boldsymbol{x}))$ nah an den experimentell ermittelten Vektor $\boldsymbol{y}$ ist, wird der MSE klein. Nachdem der MSE für das betrachtete Modell berechnet wurde, stellt sich die Frage, wie gut dieses Modell für zukünftige Daten/Molekülen geeignet ist, welche bisher nicht bei der Modellbildung zum Einsatz gekommen sind. In der Regel werden Modelle nicht nur erstellt um einen Zusammenhang zwischen den unabhängigen und abhängigen Variablen herzustellen, sondern auch um mit ihnen Eigenschaften (z.B. Bioaktivität) zukünftiger bisher nicht vorhandener Moleküle vorhersagen zu können. Um die Modellgüte testen zu können, wird der verwendete Datensatz in eine Trainingsdatenpartition und eine Testdatenpartition unterteilt. Mit den Trainingsdaten $\{(x_1, y_1), \dots, (x_n, y_n)\}$ wird die Schätzfunktion $\hat{f}$ erhalten, dieser Prozess wird auch als „Modelltraining“ bezeichnet. Anschließend können $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$ berechnet werden. Wenn diese Werte ungefähr gleich $y_1, y_2, \dots y_n$ sind, dann ist der MSE_{Train} (MSE der Trainingsdaten) klein. Wie bereits erwähnt, ist es zusätzlich interessant, nicht nur den MSE_{Train} zu berechnen, sondern zu wissen, ob $\hat{f}(x_0)$ ungefähr gleich y_0 ist. Bei x_0 handelt es sich um ein bisher ungesehenes Testdatum, welches bisher nicht als Trainingsdatum benutzt wurde. Am vielversprechendsten ist die Methode, welche den niedrigsten MSE_{Test} (MSE der Testdaten) aufweist. Denn wenn neue Moleküle hinzukommen, von denen beispielsweise der jeweilige experimentelle Wert für y_0 nicht bekannt ist, so kann davon ausgegangen werden, dass der MSE vergleichbar ist mit dem MSE_{Test}.

Wenn eine Methode einen niedrigen MSE_{Train} aber einen hohen MSE_{Test} aufweist, ist dies ein Anzeichen für eine Überanpassung. Dies passiert, weil die Methode nicht nur die unbekannte Funktion modelliert, sondern auch den Zufallsfehler. Unabhängig von der Überanpassung wird immer ein höherer MSE_{Test} als MSE_{Train} erwartet, da die meisten Methoden direkt oder indirekt versuchen den MSE_{Train} zu minimieren. Weniger flexible Methoden neigen weniger zur Überanpassung. Es ist oftmals deutlich schwieriger aufgrund geringer Datenlage den MSE_{Test} zu bestimmen und somit das Modell mit dem niedrigsten MSE_{Test} zu finden. Eine wichtige Methode, welche effizient die vorhandenen Daten ausschöpft um aus den Trainingsdaten den MSE_{Test} zu bestimmen, ist die Kreuzvalidierung. [15, 16].

Der MSE_{Test} ist ein Qualitätsmaß für die Vorhersagekraft von QSAR Modellen. Es gibt darüber hinaus noch weitere Qualitätsmaße. Auf eines von diesen wird im nächsten Abschnitt kurz eingegangen. Neben dieser Funktion dient der MSE_{Test} auch zur Auswahl

von Modellparametern *p.* Modelle werden beispielsweise für verschiedene Parameter *p* erstellt und untereinander verglichen. Anschließend wird dann das Modell mit dem entsprechenden Parameter *p* ausgewählt, welches die beste Vorhersagekraft mit sich bringt. Dieser Prozess wird auch als interne Validierung bezeichnet, da alle Moleküle (inklusive der Testmoleküle) die Modellauswahl beeinflussen und somit der MSE_{Test} möglicherweise verzerrt geschätzt wird [15].

Ein weiteres Qualitätsmaß ist der **quadrierte Korrelationskoeffizient R^2**, welcher auch als Bestimmtheitsmaß bezeichnet wird. Er beschreibt zu welchem Anteil das gebildete Modell die Varianz der abhängigen Variable erklären kann. Der R^2 nimmt Werte zwischen 0 und 1 an.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Das eigentliche Qualitätsmaß ist analog zum MSE_{Test} der R^2_{Test}, welcher mit bisher ungesehenen, von der Modellbildung unabhängigen, Molekülen berechnet wird. Bei $\bar{y}$ handelt es sich um den Mittelwert der Trainingsdaten. Eine wichtige Methode, welche effizient die Trainingsdaten nutzt um den R^2_{Test} zu berechnen, ist genau wie beim MSE_{Test}, die Kreuzvalidierung, welche im Folgenden noch näher erläutert wird [15].

1.4.2.1 Das Dilemma zwischen Bias (systematischer Fehler) und Varianz

Der erwartete MSE_{Test} für ein ungesehenes Molekül x_0 kann zerlegt werden in die Summe aus der Varianz von $\hat{f}(x_0)$, dem quadrierten Bias von $\hat{f}(x_0)$ und der Varianz des Fehlerterms *e*.

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(e)$$

$E(y_0 - \hat{f}(x_0))^2$ definieren den erwarteten MSE_{Test} und beziehen sich auf den gemittelten MSE_{Test} über alle Objekte x_0 aus dem Testdatensatz. Um den zu erwartenden Testfehler zu minimieren wird eine Methode benötigt, welche zugleich eine niedrige Varianz und einen niedrigen Bias erreicht. Der Bias eines Schätzers ist definiert als Differenz zwischen seinem Erwartungswert und der zu schätzenden Größe. Die Varianz und der quadrierte Bias sind positiv. Somit lässt sich erkennen, dass der zu erwartende MSE_{Test} niemals kleiner sein kann als die Varianz von *e*, dem nicht reduzierbaren Fehler. Die

$Var(\hat{f}(x_0))$ bezieht sich auf den Grad der Änderung von $\hat{f}$, wenn zur Schätzung verschiedene Trainingsdatensatzpartitionen benutzt werden. Aus unterschiedlichen Trainingsdatensatzpartitionen resultieren unterschiedliche $\hat{f}$s. Im Idealfall sollten die Unterschiede nicht zu groß sein. Falls eine Methode eine hohe Varianz aufweist, können kleine Unterschiede in den Trainingsdaten große Unterschiede in $\hat{f}$ hervorrufen. Allgemein weisen flexiblere Methoden eine höhere Varianz auf. Der Bias bezieht sich auf den Fehler, der gemacht wird, wenn ein komplexes Problem auf ein viel einfacheres Modell reduziert wird. Flexiblere Methoden weisen in der Regel einen geringeren Bias auf, aber eine höhere Varianz. Das Verhältnis in dem sich diese beiden Größen verändern entscheidet darüber, ob der MSE steigt oder sinkt. Wenn die Flexibilität von einer Methode erhöht wird, dann neigt der Bias dazu stärker zu sinken als die Varianz steigt und der MSE verringert sich. Ab einem bestimmten Punkt jedoch hat eine Steigerung der Flexibilität keinen großen Einfluss mehr auf den Bias aber die Varianz steigt signifikant, folglich vergrößert sich der MSE_{Test}. Das Ziel ist es, eine Methode zu finden, welche eine niedrige Varianz und einen niedrigen Bias aufweist. Angenommen das wahre f ist linear, dann würde die lineare Regression keinen Bias haben und flexiblere Methoden hätten Schwierigkeiten mitzuhalten. Wenn aber das wahre f hochgradig nicht-linear ist, funktionieren flexiblere Methoden vermutlich besser [15–18].

1.4.2.2 Kreuzvalidierung (engl.: Cross-Validation (CV))

Prinzipiell werden bei der Kreuzvalidierung der gesamte zur Verfügung stehende Datensatz in einen Konstruktionsdatensatz und einen Validierdatensatz aufgeteilt. Mit den Konstruktionsdaten wird ein Modell gebildet. Daraufhin werden die Eigenschaften der Validierdaten mit dem erstellten Modell vorhergesagt und ein Gütekriterium, wie beispielsweise der MSE, wird berechnet. Dieser Prozess wird mehrfach wiederholt, allerdings werden unterschiedliche Konstruktions- bzw. Validierdatenpartitionen gebildet. Je nach Vorgehensweise und Aufbau werden unterschiedliche Varianten der Kreuzvalidierung unterschieden.

Die **„Lass–ein–Objekt-heraus-Kreuzvalidierung“ (engl.: Leave-One-Out-Cross-Validation (LOO-CV))** ist eine Variante der Kreuzvalidierung, bei der einem Datensatz bestehend aus n Molekülen immer ein Molekül entzogen wird. Mit den restlichen $n - 1$ Molekülen wird das Modell gebildet und das separierte Molekül wird anschließend vor-

hergesagt. Der Vorgang wird so lange wiederholt, bis jedes Molekül einmal vorhergesagt wurde. Folglich wurden insgesamt *n* Modelle gebildet. In Abbildung 2 ist das Schema der LOO-CV graphisch dargestellt.

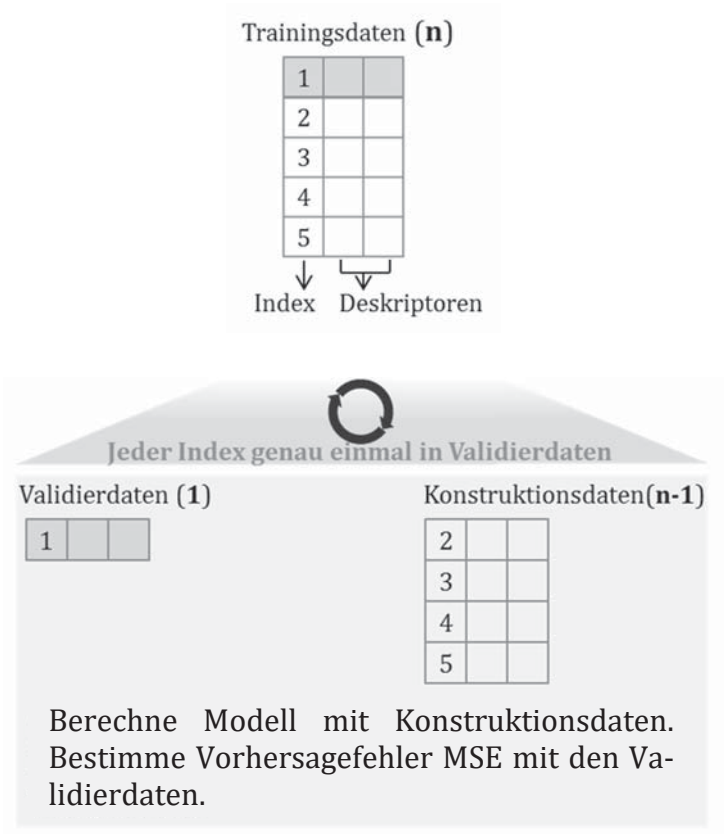


Abbildung 2: Graphische Darstellung des Prinzips der LOO-CV. Bei dieser Variante der CV wird jedes Molekül genau einmal bei der Modellerstellung ausgelassen und vorhergesagt.

Die **k-fache Kreuzvalidierung (engl.: k-Fold-Cross-Validation (k-CV))** ist eine weitere Variante der Kreuzvalidierung, bei der der Datensatz zufällig in k möglichst gleich große Teilmengen unterteilt wird. Im Gegensatz zur LOO-CV, wo immer ein Molekül ausgelassen wurde, wird bei der k-CV jede Teilmenge k genau einmal zur Validierung zur Seite gelegt. Die übrigen $k - 1$ Teilmengen werden entsprechend zur Modellerstellung verwendet. Somit werden hier insgesamt k Modelle erstellt. Im Fall $k = n$ entspricht die k-CV der LOO-CV. Die ursprüngliche Motivation, welche zur Entwicklung der k-CV geführt hatte, war die Einsparung von Rechenzeit [19]. Die Anzahl der Teilmengen k ist ein benutzerdefinierter Hyperparameter. In vielen Fällen wird dieser allerdings gleich 5 oder 10 gesetzt [15].

Die **„Lass-mehrere-Objekte-heraus-Kreuzvalidierung“ (engl.: Leave-Multiple-Out-Cross-Validation (LMO-CV))** ist ebenfalls ein weit verbreitetes Schema der Kreuzvalidierung [20], bei welchem der Datensatz wiederholt zufällig in einen Konstruktionsdatensatz und einen Validierdatensatz unterteilt wird. Der Validierdatensatz kann beliebig viele Moleküle (d) enthalten, somit besteht der Konstruktionsdatensatz, mit welchem das Modell gebildet wird, aus $n - d$ Molekülen. Die Aufteilung des Datensatzes kann beliebig oft erfolgen. Generell sollte die Anzahl an Wiederholungen eher höher gewählt werden, da somit der Einfluss der zufälligen Datenpartitionen auf den resultierenden Schätzwert des Fehlers vermindert wird [21].

Außerdem gibt es die Unterscheidung zwischen einfacher Kreuzvalidierung [15] und doppelter Kreuzvalidierung, welche bereits von M. Stone in dem Artikel „Cross-Validatory Choice and Assessment of Statistical Predictions“ 1974 beschrieben wurde. Bei der einfachen Kreuzvalidierung handelt es sich um ein einstufiges Verfahren, welches häufig zur Modelloptimierung oder Modellselektion eingesetzt wird [15, 21]. Bei der doppelten Kreuzvalidierung handelt es sich um ein zweistufiges Verfahren, welches aus zwei ineinander verschachtelten Kreuzvalidierschleifen besteht und zusätzlich die Evaluierung von Modellen gestattet.

1.4.3 Multiple Lineare Regression

Im Fall der Multiplen Linearen Regression wird das Modell der Einfachen Linearen Regression auf mehrere abhängige Variablen erweitert. Es wird davon ausgegangen, dass sich $\boldsymbol{y}$ additiv aus der Summe der Beiträge der einzelnen, unabhängigen Variablen zusammensetzt und die Regressionskoeffizienten den Einfluss der einzelnen Variablen quantifizieren:

$$\boldsymbol{y} = b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \cdots .. b_p\boldsymbol{x_p} + \boldsymbol{e}.$$

In Matrixschreibweise:

$$\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e}\,.$$

Hierbei ist $X \in \mathbb{R}^{n*(p+1)}$ (n = Anzahl der Moleküle, p = Anzahl der unabhängigen Variablen), $b \in \mathbb{R}^{p*1}$ und $e \in \mathbb{R}^{n*1}$. Die Koeffizienten-Schätzung erfolgt wieder durch Minimierung der RSS (Summe quadrierter Residuen):

$$RSS(b) = ||\boldsymbol{y} - X\boldsymbol{b}||_2^2$$

mit der L_2 Norm $|| r ||_2$:

$$||r||_2 = \sqrt{\sum_{i=1}^{n} r_i^2}.$$

Die Lösung für die Regressionskoeffizienten ergibt sich wie folgt:

$$\widehat{\boldsymbol{b}} = \underset{\boldsymbol{b}}{\operatorname{argmin}}\, RSS(b) = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Da $\widehat{\boldsymbol{b}}$ so gewählt wird, dass die Fehlerquadratsumme der Residuen minimiert wird, wird dieser Schätzer auch „kleinste Quadrate“-Schätzer (engl.: Least Squares) genannt.

Zukünftige Moleküle X_0 können nun auf diese Weise vorhergesagt werden:

$$\widehat{\boldsymbol{y_0}} = X_0 \widehat{\boldsymbol{b}} \quad [15, 16, 22].$$

Auf dem Gebiet der QSAR stehen eine große Anzahl unterschiedlicher Molekül-deskriptoren mit zum Teil hunderten von Variablen zur Verfügung. Wenn $n < p$ ist, dann ist die Schätzung der Regressionskoeffizienten $\widehat{\boldsymbol{b}}$ nicht mehr eindeutig, da unendlich viele gleichwertige Lösungen existieren [23]. Folglich kann die Methode nicht verwendet werden. Darüber hinaus ist es häufig der Fall, dass nicht alle unabhängigen Variablen mit der abhängigen Variablen in Verbindung stehen. Solche irrelevanten Variablen führen zu unnötig komplexen Modellen. Wenn diese Variablen aus dem Modell entfernt werden würden, indem die entsprechenden $\widehat{\boldsymbol{b}} = 0$ gesetzt werden würden, so würde ein besser interpretierbares Modell erhalten werden. Es gibt verschiedene Alternativen zum oben genannten Schätzer. An dieser Stelle wird nur auf zwei Gruppen eingegangen. Die erste Gruppe umfasst die Techniken, die unter den englischen Begriff „Shrinkage“ fallen. Hierbei handelt es sich um Regressionstechniken, welche die geschätzten Regressionskoeffizienten Richtung Null schrumpfen. Somit sind diese Methoden in bestimmten Fällen zusätzlich in der Lage Variablenselektion zu betreiben. Die andere Gruppe umfasst die Techniken, welche die Dimension von X reduzieren indem die p Variablen in einen m-dimensionalen Unterraum projiziert werden, welcher die Eigenschaft $m < p$ besitzt [15, 16].

1.4.4 Shrinkage-Methoden

1.4.4.1 Ridge Regression (Ridge)

Wenn $n < p$ kann die MLR-Schätzung nicht verwendet werden, da es, wie bereits beschrieben, keine eindeutige Lösung gibt. Die MLR-Schätzung ist auch nicht anwendbar, wenn X einen reduzierten Spaltenrang hat [23]. Es wird von exakter Kollinearität gesprochen, wenn die Spalten von X exakt linear abhängig sind. Die Matrix $X^T X$ ist dann singulär, die Inverse kann nicht berechnet werden und der „Kleinste Quadrate"-Schätzer existiert nicht. Sobald die abhängigen Variablen stark miteinander korrelieren, wird dies als Multikollinearität bezeichnet. Falls exakte Multikollinearität vorliegt ist die Matrix $X^T X$ ebenfalls singulär und nicht invertierbar. Multikollineares Verhalten führt zu einer nahezu singulären Matrix $X^T X$. Dieses Phänomen wird häufig bei QSAR Datensätzen beobachtet, da die verwendeten Deskriptoren häufig stark korreliert sind. In diesem Fall können die Regressionskoeffizienten zwar eindeutig geschätzt werden, allerdings möglicherweise mit einer inakzeptabel hohen Varianz der einzelnen Koeffizienten und der absolute Wert der Koeffizienten kann künstlich ansteigen [15, 24, 25]. Diese Schwierigkeiten waren die ursprüngliche Motivation für die Einführung der Ridge Regression [26]. Hierbei wird nicht die Fehlerquadratsumme (RSS) minimiert, sondern es wird zusätzlich ein Bestrafungsterm eingeführt und die Summe aus der RSS und diesem Strafterm wird minimiert. Dies wird auch als Bestrafte-Summe-Quadrierter-Abweichungen-Kriterium (engl.: Penalized-Sum-of-Squares-Criterion (PSS)) bezeichnet:

$$PSS_2(\boldsymbol{b}, \lambda) = ||\boldsymbol{y} - X\boldsymbol{b}||_2^2 + \lambda||\boldsymbol{b}||_2^2$$

An dieser Stelle wird angenommen, dass X zentriert ist und somit kein y-Achsenabschnitt vorhanden ist, sonst würde dieser Koeffizient nicht bestraft werden. Der Strafterm wird durch den Parameter $\lambda > 0$ kontrolliert. Je größer der Wert für λ gewählt wird, desto stärker wirkt sich die Bestrafung auf $\widehat{\boldsymbol{b}}$ aus. Dieser Vorgang wird als Schrumpfung der Regressionskoeffizienten Richtung Null bezeichnet. Um einen geeigneten Parameter λ auszuwählen, stehen iterative Algorithmen sowie die Kreuzvalidierung zur Verfügung. Für $\lambda = 0$ entspricht der Ridge-Schätzer dem „Kleinste Quadrate"-Schätzer. Durch bestimmte analytische Eigenschaften der L_2 Norm ergibt sich durch Ableitung die Lösung:

$$\widehat{\boldsymbol{b}}(\lambda) = \underset{\boldsymbol{b}}{\operatorname{argmin}}\ PSS_2(\boldsymbol{b}, \lambda) = (X^T X + \lambda I)^{-1} X^T \boldsymbol{y}.$$

Bei der Matrix I handelt es sich um eine Identitätsmatrix $I \in \mathbb{R}^{p*p}$. Die Diagonalmatrix der Form λI wird zu $X^T X$ addiert und es resultiert für $\lambda > 0$ eine invertierbare Matrix $X^T X + \lambda I$. Somit wird die Varianz verringert und zugleich wurden die Regressionskoeffizienten geschrumpft. Ein Nachteil der Ridge-Regression ist allerdings, dass genau wie bei dem „Kleinste Quadrate"-Schätzer keine automatische Variablenselektion durchgeführt wird, d.h. keiner der Regressionskoeffizienten wird exakt Null geschätzt [22].

1.4.4.2 Lasso

Lasso ist eine Abkürzung und steht für „Least Absolute Shrinkage and Selection Operator" [27]. Im Gegensatz zur Ridge-Regression, welche einen L_2-Strafterm zum RSS-Kriterium addiert, addiert das Lasso einen L_1Strafterm:

$$PSS_1(\boldsymbol{b}, \lambda) = ||\boldsymbol{y} - X\boldsymbol{b}||_2^2 + \lambda||\boldsymbol{b}||_1 .$$

Hierbei kennzeichnet $||b||_1$ die Verwendung der L_1-Norm:

$$||\boldsymbol{b}||_1 = \sqrt{\sum_{i=1}^{p} |b_i|}.$$

Der sich daraus ergebene Regressionsschätzer definiert sich folgenderweise:

$$\widehat{\boldsymbol{b}}(\lambda) = \underset{\boldsymbol{b}}{argmin}\ PSS_1(\boldsymbol{b}, \lambda).$$

Die Lösung ist nicht linear in $\boldsymbol{y}$, deshalb ist sie nicht in geschlossener Form darstellbar. Aufgrund der Eigenschaften der L_1-Norm können Regressionskoeffizienten in $\widehat{\boldsymbol{b}}$ auch exakt Null gesetzt werden. Dieser Fall tritt ein, wenn λ groß genug gewählt wird. Falls $\lambda = 0$ ist, dann ergibt sich der „Kleinste Quadrate"-Schätzer. Durch die Möglichkeit Koeffizienten exakt Null setzen zu können und damit die entsprechenden Variablen aus dem Modell entfernen zu können, kann das Lasso auch als Variablenselektionsmethode angesehen werden. Denn die wichtigsten Variablen werden in das Modell integriert. Die Variablenselektion führt zu Modellen, welche einfacher zu interpretieren sind. Rauschen (Variablen mit wenig Vorhersagekraft) wird unterdrückt, somit kommt es in der Regel zu einer Verbesserung des Modells und dessen Vorhersagequalität [22].

1.4.4.3 Elastic Net

Diese Technik wurde zuerst von Zou und Hastie eingeführt und kann als Kompromiss zwischen Ridge Regression und Lasso betrachtet werden [28]. Es schrumpft die Koeffizienten wie die Ridge Regression und selektiert Variablen wie das Lasso. Die Zielfunktion ist wie folgt definiert:

$$PSS_{1,2}(\boldsymbol{b}, \lambda_1, \lambda_2) = ||\boldsymbol{y} - X\boldsymbol{b}||_2^2 + \lambda_1||\boldsymbol{b}||_1 + \lambda_2||\boldsymbol{b}||_2^2.$$

Der sich daraus ergebene Regressionsschätzer definiert sich folgendermaßen:

$$\widehat{\boldsymbol{b}}(\lambda_1, \lambda_2) = (1 + \lambda_2)\underset{\boldsymbol{b}}{argmin}\ PSS_{1,2}(\boldsymbol{b}, \lambda_1, \lambda_2).$$

Er wird durch numerische Optimierung erhalten. Das Elastic Net ist flexibler als die beiden anderen Techniken. Für den Fall $\lambda_2 = 0$ entspricht das dem Lasso. Neben der Flexibilität besitzt das Elastic Net noch einen anderen Vorteil gegenüber dem Lasso, wenn $n < p$ ist, kann das Lasso nur höchstens n Variablen auswählen. In einigen Fällen, wenn n beispielsweise sehr klein ist, kann es aber vorteilhaft sein mehr Variablen auswählen zu können. Das Elastic Net ist durch die Kombination (Ridge Regression und Lasso) in der Lage dazu [22]. Darüber hinaus ist ein weiterer potentieller Vorteil des Elastic Nets, dass es, wenn Variablen eine hohe paarweise Korrelation aufweisen, die ganze Gruppe auswählen kann und dann mittelt, wohingegen das Lasso in diesem Fall nur eine Variable auswählen würde [15].

1.4.5 Dimensions-Reduktions-Methoden

Die zuvor betrachteten Schrumpf-Methoden wurden unter Verwendung der unveränderten, unabhängigen Variablen definiert. Nun werden Methoden betrachtet, welche zunächst die unabhängigen Variablen transformieren [16].

1.4.5.1 Hauptkomponentenregression (engl.: Principal Component Regression (PCR))

Die Kernidee der PCR ist es, Hauptkomponenten der ursprünglichen unabhängigen Variablen zu berechnen und diese danach zur Regression zu verwenden. Hauptkomponenten können die Dimensionalität des ursprünglichen Regressionsproblems reduzieren

und orthogonale, unabhängige Variablen generieren. Somit wird das Problem der Multikollinearität vermieden [15, 22].

Die **Hauptkomponentenanalyse** (engl.: Principal Component Analysis (PCA)) [29] sucht nach orthogonalen Richtungen $\boldsymbol{a}$, für die die Varianz der projizierten Daten $X\boldsymbol{a}$ maximiert wird. Hierbei ist die Kovarianz Matrix von $X\,\mathbb{V}ar(X) = \hat{\Sigma}$ und daraus folgt:

$$\mathbb{V}ar(X\boldsymbol{a}) = \boldsymbol{a}^T\mathbb{V}ar(X)\boldsymbol{a} = \boldsymbol{a}^T\hat{\Sigma}\boldsymbol{a}.$$

Das Kriterium für die j-te Hauptkomponente ist:

$$\widehat{\boldsymbol{a}}_j = \underset{\boldsymbol{a},||\boldsymbol{a}||_2=1}{\operatorname{argmax}}\, \boldsymbol{a}^T\hat{\Sigma}\,\boldsymbol{a} \qquad s.t.\, \boldsymbol{a} \perp \widehat{\boldsymbol{a}}_{\boldsymbol{i}}\ \forall i\ \in \{1, \dots, j-1\}.$$

Die Regression auf $\boldsymbol{y}$ wird danach auf den ersten k Hauptkomponenten durchgeführt, bzw. genauer gesagt auf den sogenannten „Scores“ oder „Score-Vektoren“ $X\widehat{\boldsymbol{a}}_i, i \in \{1, \dots, k\}$. Die optimale Anzahl an k ist unbekannt und kann beispielsweise über die Kreuzvalidierung ausgewählt werden. Häufig ist die ausgewählte Anzahl k deutlich kleiner als p (Lösung des Problems der Invertierbarkeit singulärer Matrizen) [22]. Hierdurch wird die Varianz der geschätzten Regressionskoeffizienten reduziert. Allerdings bringt diese Nichtberücksichtigung auch einen systematischen Fehler mit sich (siehe Bias-Varianz-Dilemma Kapitel 1.4.2.1). In der Praxis überwiegt in der Regel der Vorteil einer starken Varianzreduktion gegenüber dem Nachteil einer kleinen Erhöhung des systematischen Fehlers [15].

Mit der PCA werden „Score-Vektoren“ $X\widehat{\boldsymbol{a}}_{\boldsymbol{j}}$ erhalten, welche danach für die Regression benutzt werden, ohne die Information von $\boldsymbol{y}$ zu verwenden. Dies ist ein offensichtlicher Nachteil, welcher schließlich bei der Partial Least Squares Regression (PLS) berücksichtigt wird [22].

1.4.5.2 Partial Least Squares Regression (PLS)

An dieser Stelle wird erneut vom multiplen linearen Modell ausgegangen:

$$Y = XB + E\,.$$

Jedoch wird die abhängige Variable $\boldsymbol{y}$ erweitert zu einer abhängigen Variablen $\in \mathbb{R}^{n*q}$ Matrix Y, $X \in \mathbb{R}^{n*p}$, $B \in \mathbb{R}^{p*q}$und $E \in \mathbb{R}^{n*q}$. Es wird angenommen, dass X und Y zentriert vorliegen.

Ähnlich zu der PCR führt auch die PLS [30, 31] eine Dimensionsreduktion auf den ursprünglichen, unabhängigen Variablen durch und sucht nach „Richtungen“ $\boldsymbol{w}$. Die Zielfunktion der PLS ist, im Gegensatz zur PCR, die Kovarianz zwischen den Score-Vektoren $X\boldsymbol{w}$ und einer linearen Projektion auf den abhängigen Variablen Y zu maximieren. Somit wird sichergestellt, dass die neu erhaltenen, unabhängigen Variablen die wichtigen Informationen für die Vorhersage von den abhängigen Variablen enthalten [22].

Für die PLS Regression gibt es eine Vielzahl unterschiedlicher Modelle und Schätzer. Eine Übersicht ist in [32] zu finden. An dieser Stelle wird der Ansatz von Chun und Keles erläutert, da diese Autoren auch eine Version der Sparse Partial Least Squares Regression (SPLS) beschreiben, mit welcher fortgefahren werden soll [33].

Die Kernidee der PLS Regression beinhaltet die Zerlegung der Matrix der unabhängigen Variablen X und der abhängigen Variable(n) Y:

$$X = TP^T + E_x \ (1)$$

$$Y = TQ^T + E_Y \ (2),$$

bei T handelt es sich um die Score Matrix $T = XW \in \mathbb{R}^{n*k}$ (Scores) und bei W handelt es sich um die Matrix der „Richtungen“ (engl.: Loadings Vectors (Loadings)) $W = (\boldsymbol{w_1}, \dots, \boldsymbol{w_k}) \in \mathbb{R}^{p*k}$. Die Gleichungen (1) und (2) können als „Kleinste-Quadrate“ Schätzer Problem aufgefasst werden. Somit sind $P \in \mathbb{R}^{p*k}$ und $Q \in \mathbb{R}^{q*k}$ Matrizen der Regressionskoeffizienten und $E_x \in \mathbb{R}^{n*p}$ und $E_y \in \mathbb{R}^{n*q}$ sind Matrizen des Zufallsfehlers. K steht wie im letzten Abschnitt für die Anzahl an Komponenten, wobei $k \leq \min\{n, p, q\}$ ist [22].

An dieser Stelle kann nun die Gleichung (2) neu formuliert werden:

$$Y = TQ^T + E_Y = XWQ^T + E_Y.$$

Bei $WQ^T \in \mathbb{R}^{p*q}$ handelt es sich eine Matrix von Regressionskoeffizienten, welche Y mit den ursprünglichen, unabhängigen Variablen X in Beziehung setzt [32, 22].

Um die Vektoren $\boldsymbol{w}$ zu finden, welche die Kovarianz zwischen den unabhängigen Variablen und den abhängigen Variablen maximieren, wird das von de Jong vorgestellte SIMPLS Kriterium verwendet [34]:

$$\widehat{\boldsymbol{w}_j} = \underset{\boldsymbol{w},||\boldsymbol{w}||_2=1}{argmax}\, \boldsymbol{w}^T N \boldsymbol{w} \qquad s.t.\ \boldsymbol{w}^T \hat{\Sigma} \widehat{\boldsymbol{w}_i} = 0\ \forall i \in \{1, \dots, j-1\}.$$

Bei $\hat{\Sigma}$ handelt es sich um die Kovarianzmatrix von X und $N = X^T Y Y^T X$ [22]. Sobald $(\widehat{\boldsymbol{w_1}}, \dots, \widehat{\boldsymbol{w_k}})$ erhalten wurden, kann Q mit Hilfe des „Kleinste-Quadrate" Schätzers geschätzt werden und es wird folgende Gleichung erhalten:

$$\widehat{\boldsymbol{b}} = \widehat{W} \hat{Q}^T [22].$$

1.4.5.3 Sparse Partial Least Squares Regression (SPLS)

An dieser Stelle steht das englische Wort sparse (zu Deutsch „spärlich") nicht für die Art der Daten (sparse data enthalten viele Nullen als Einträge), sondern für den geschätzten Koeffizienten Vektor, welcher dazu gebracht wird viele Nullen zu beinhalten, indem ein Strafterm zur Zielfunktion hinzugefügt wird. Beispielsweise die zuvor beschriebene Shrinkage-Methode Lasso zählt somit auch zu den sparse Methoden [22].

Falls ein Strafterm, auf die bei den Shrinkage-Methoden bereits beschriebene Weise, bei der PLS hinzugefügt wird, dann sind, wenn ein hoher Anteil erklärter Varianz benötigt wird, die erhaltenen Schätzer für die Richtung $\boldsymbol{w}$ nicht sparse genug [35]. Um dem Rechnung zu tragen wurde von Chun und Keles die Zielfunktion verändert [33], angelehnt an die Ansätze zur Sparse PCA und dem Elastic Net [36, 28]. Anstatt die „Spärlichkeit" dem ursprünglichen Vektors $\boldsymbol{w}$ aufzuerlegen, wird sie einem Ersatz, dem Vektor $\boldsymbol{c}$ auferlegt.

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{c},\boldsymbol{w},||\boldsymbol{w}||_2=1}{argmin}\ -\kappa \boldsymbol{w}^T N \boldsymbol{w} + (1-\kappa)(\boldsymbol{c}-\boldsymbol{w})^T N(\boldsymbol{c}-\boldsymbol{w}) + \lambda_1 ||\boldsymbol{c}||_1 + \lambda_2 ||\boldsymbol{c}||_2^2 \quad .$$

Diese Formel besteht aus verschiedenen Teilen. Der erste Term $\boldsymbol{w}^T N \boldsymbol{w}$ ist, wie bereits beim ursprünglichen SIMPLS Kriterium, verantwortlich für eine hohe Kovarianz zwischen den abhängigen und unabhägigen Variablen. Der zweite Term $(\boldsymbol{c}-\boldsymbol{w})^T N(\boldsymbol{c}-\boldsymbol{w})$ stellt sicher, dass $\boldsymbol{c}$ und $\boldsymbol{w}$ nahe zusammen gehalten werden. Der Parameter κ kontrolliert den Kompromiss zwischen diesen beiden Termen. Der L_1 Strafterm legt dem Vektor $\boldsymbol{c}$ die „Spärlichkeit" auf und der L_2 Strafterm ist verantwortlich für die Schrumpfung der Parameter. Die Lösung der obrigen Gleichung für $\boldsymbol{c}$ und ein fixes $\boldsymbol{w}$ erfolgt analog wie

beim Elastic Net. In der Praxis kann das Problem 4 Parameter zu finden auf eine 2 Parameter Suche reduziert werden, da für λ_2 gewöhnlich ein sehr hoher Wert ausgewählt wird und dieser somit auf Unendlich gesetzt werden kann und eine univariate Lösung für *Y* nicht von κ abhängig ist [22].

Die Regressionstechniken Random Forest Regression (RFR), Support Vector Regression (SVR) und Neuronale Netze zur Regression sind eng verwandt mit den entsprechenden Klassifikationstechniken. Aus diesem Grund werden die Grundlagen dieser Techniken im Kapitel 1.5 Klassifikation für beide Varianten (Klassifikation und Regression) gemeinsam erläutert.

1.5 Klassifikation

1.5.1 Einleitung

Variablen können sowohl qualitativ als auch quantitativ sein. Bei quantitativen abhängigen Variablen wird meist von Regressionsproblemen gesprochen, wie bereits im vorherigen Kapitel beschrieben. Bei qualitativen abhängigen Variablen wird meist von Klassifikationsproblemen gesprochen. Das Ziel ist es, auf Basis der unabhängigen Variablen eines bisher ungesehenen Moleküls, dessen Klasse vorherzusagen [37, 38]. Wenn es nur zwei unterschiedliche Klassen gibt, wird von einem binären Klassifikationsproblem gesprochen [16].

Bisher wurde die Regression betrachtet, aber viele Konzepte wie das Bias-Varianz-Dilemma lassen sich mit ein paar Modifikationen (weil y nicht mehr numerisch ist) auf die Klassifikation übertragen. Das Ziel ist es, f auf Basis der Trainingsdatenpartition $\{(\boldsymbol{x_1},y_1),\ldots,(\boldsymbol{x_n},y_n)\}$, wobei $y_1,\ldots,y_n$ qualitativ sind, zu schätzen und danach auf Basis der unabhängigen Variablen X_0 zukünftige abhängige Variablen $\boldsymbol{y_0}$ vorherzusagen.

1.5.2 Beurteilung der Klassifikationsgüte des Modells

Während der Modellentwicklung muss die Vorhersagekraft des Klassifikationsmodells bewertet werden. Diese Bewertung wird bei der Klassifikation analog zur Regression durchgeführt. Der geläufigste Ansatz um die Güte der Schätzung von $\hat{f}$ zu quantifizieren ist die Fehler-Rate der Trainingsdaten. Hierbei handelt es sich um das Verhältnis der

Fehler, die gemacht werden, wenn die Schätzfunktion $\hat{f}$ auf die Trainings-Beobachtungen anwendet wird [16]:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{f}(X)).$$

In diesem Fall ist $\hat{y}_i$ die vorhergesagte Klasse für die i-te Beobachtung, wenn die Schätzfunktion $\hat{f}$ angewendet wird. $I(y_i \neq \hat{y}_i)$ ist eine Indikatorvariable, welche 1 ergibt, wenn $(y_i \neq \hat{y}_i)$, und welche 0 ergibt, wenn $(y_i = \hat{y}_i)$. Wenn $I(y_i \neq \hat{y}_i) = 0$, dann wurde die *i*-te Beobachtung richtig durch die Klassifikationsmethode klassifiziert. Folglich wird die Fraktion der falsch klassifizierten Moleküle berechnet. Viel interessanter als die Training-Fehler-Rate ist aber, wie bei der Regression, die Fehler-Rate der Testdaten, welche den mittleren Fehler für bisher ungesehene Moleküle angibt, von denen allerdings die Klassenzugehörigkeit bekannt ist [16]:

$$Ave(I(\boldsymbol{y_0} \neq \hat{\boldsymbol{y}}_0)).$$

An dieser Stelle soll noch einmal in Erinnerung gerufen werden, dass diese Testmoleküle weder für die Modellbildung, noch für die Modellselektion sondern lediglich zur Beurteilung der Leistung verwendet werden dürfen [39]. Die Fehler-Rate der Testdaten sollte möglichst klein sein. Wenn bisher ungesehene Moleküle hinzugenommen werden und vorhergesagt werden sollen, von denen möglicherweise die Klassenzugehörigkeit nicht bekannt ist, so wird erwartet, dass die Fehler-Rate der Testdaten vergleichbar sein wird [40].

Im Zwei-Klassen-Fall leiten sich die gebräuchlichsten Leistungskriterien von den vier möglichen Ausgaben des Klassifikations-Algorithmus ab. Hierfür wird einmal angenommen, dass es sich bei der ersten Klasse um die sogenannten „Positiven“ und bei der zweiten Klasse um die „Negativen“ handelt. Wenn ein Molekül „Positiv“ ist und auch richtig als „Positiv“ klassifiziert wurde, dann wird es als „Richtig Positiv“ (engl.: True Positive (TP)) bezeichnet. Ein Molekül, dass eigentlich „Positiv“ ist, aber fälschlicherweise als „Negativ“ klassifiziert wurde, wird als „Falsch Negativ“ (engl.: False Negative (FN)) bezeichnet. Analog dazu wird ein Molekül, dass als „Negativ“ klassifiziert wurde und auch tatsächlich „Negativ“ ist, als „Richtig Negativ“ (engl.: True Negative (TN)) bezeichnet und wenn ein Molekül der „Negativen“ Klasse fälschlicherweise als „Positiv“ klassifiziert wurde, dann wird es als „Falsch Positiv“ (engl.: False Positive (FP)) bezeich-

net. Die soeben beschriebenen Größen werden häufig in Form einer Wahrheitsmatrix (engl.: Confusion matrix) dargestellt [41, 42, 40].

Die Korrektklassifizierungsrate (engl.: Accuracy (Acc)) einer Klassifikationstechnik gibt den prozentuellen Anteil der richtig klassifizierten Objekte an:

$$Acc = \frac{TP + TN}{(TP + FN + TN + FP)}.$$

Wenn die Klassen sehr unausgeglichen besetzt sind, kann die Korrektklassifizierungsrate (Acc) einen falschen Eindruck über die Leistung erwecken, da aus der Korrektklassifizierungsrate (Acc) nicht hervorgeht wie die einzelnen Klassen getrennt wurden. Beispielsweise könnte immer nur eine Klasse vorhergesagt werden, was, wenn der Datensatz zu 80% aus dieser Klasse besteht, in 80% der Fälle richtig wäre. Es würde der Eindruck erweckt werden, die Klassifikation war recht erfolgreich, obwohl sie völlig uninformativ ist [43]. Aus diesem Grund werden zusätzlich die Sensitivität (engl.: Sensitivity (Sens)), welche den prozentualen Anteil der richtig klassifizierten „Positiven" Moleküle angibt und die Spezifität (engl.: Specificity (Spec)), welche den prozentualen Anteil der richtig klassifizierten „Negativen" Moleküle angibt, mit aufgeführt [40]:

$$Sens = \frac{TP}{(TP + FN)}$$

$$Spec = \frac{TN}{(TN + FP)}.$$

1.5.3 k-Nächste Nachbarn (engl.: k-Nearest Neighbor (KNN))

Der KNN-Algorithmus gehört aufgrund seiner einfach nachvollziehbaren Funktionsweise zu den mit am häufigsten angewendeten Techniken des Maschinellen Lernens. Die Methode beruht auf der Annahme, dass ähnliche Moleküle vermutlich auch zur selben Klasse gehören [44, 16]. Es handelt sich um einen nicht-linearen, nicht-parametrischen, distanz-basierten Ansatz, bei welchem auf Basis der unabhängigen Variablen für ein neues Molekül $\boldsymbol{x_0}$ die k-nächsten Moleküle der Trainingspartition gesucht werden. Die Klassenzuweisung erfolgt auf Basis eines Mehrheitsentscheids der Klassenzugehörigkeiten der k-nächsten Nachbarn (siehe Abbildung 3). Die k-nächsten Nachbarn sind diejenigen Moleküle der Trainingspartition, die die kürzeste Distanz zum bisher ungesehenen

Molekül $\boldsymbol{x_0}$ aufweisen [45–47]. Das am häufigsten verwendete Distanzmaß ist die Euklidische Distanz:

$$d_{i,j} = d(\boldsymbol{x_i}, \boldsymbol{x_j}) = ||\boldsymbol{x_i} - \boldsymbol{x_j}||_2$$

Bei $(\boldsymbol{x_i}, \boldsymbol{x_j})$ handelt es sich entweder um zwei Moleküle aus der Trainingsdatenpartition oder ein Molekül aus der Trainingsdatenpartition und ein ungesehenes Molekül $\boldsymbol{x_0}$ [40]. Wenn zwei unabhängige Variablen korreliert sind, wird die Distanz innerhalb dieses Pärchens, im Vergleich zu der Distanz innerhalb eines unkorrelierten Pärchens, überschätzt. Bei Benutzung der Euklidischen Distanz sollten deshalb die unabhängigen Variablen möglichst unkorreliert und vergleichbar skaliert sein [48]. Korrelationen zwischen Variablen können behoben werden, wenn anstelle der Euklidischen Distanz, die Mahalanobis Distanz verwendet wird [49]. Die Mahalanobis Distanz kann als eine speziell „skalierte“ Form der Euklidischen Distanz aufgefasst werden [40]. Darüber hinaus gibt es die Manhattan Distanz [50], welche sich hauptsächlich nur von der Euklidischen Distanz in der Gewichtung größerer Entfernungen unterscheidet. Durch die Berechnung mittels L_1-Norm, werden bei der Manhattan Distanz alle Entfernungen gleich gewichtet, wobei die Euklidische Distanz große Entfernung durch Quadrieren bestraft. Darüber hinaus gibt es sowohl für die Manhattan Distanz als auch für die Euklidische Distanz gewichtete Versionen [40].

An dieser Stelle soll noch einmal in Erinnerung gerufen werden, dass die Fingerabdruck-Deskriptoren vorwiegend binäre Variablen beinhalten. Bei binären Variablen wird in vielen Fällen die sogenannte Tanimoto- oder auch Jaccard-Ähnlichkeit berechnet [51]:

$$s_{i,j} = \frac{n_{11}}{n_{i1} + n_{j1} + n_{11}}$$

Hierbei steht n_{11} für die Anzahl an Variablen, die sowohl in dem Molekül *i* als auch in dem Molekül *j* 1 betragen, n_{i1} symbolisiert die Anzahl an Variablen, die in Molekül *i* 1 sind und n_{j1} symbolisiert die Anzahl an Variablen die in Molekül *j* 1 sind. Es existiert allerdings auch eine Erweiterung für kontinuierliche Variablen [52]. Um ausgehend von dem Ähnlichkeitsmaß das entsprechende Distanzmaß zu erhalten wird das Ähnlichkeitsmaß noch von 1 subtrahiert [40]:

$$d_{i,j} = 1 - s_{i,j}.$$

Die Auswahl von *k* bestimmt die Flexibilität der Entscheidungsebene, ein zu kleines *k* kann leicht zu einer Überanpassung des Modells führen. Eine gute Methode um ein geeignetes *k* zu finden ist die Kreuzvalidierung. Diese kann natürlich ebenfalls verwendet werden um ein geeignetes Distanzmaß auszuwählen [16]. Im Regressionsfall wird aus den kontinuierlichen, abhängigen Variablen der k-nächsten Nachbarn des betrachteten Moleküls der Durchschnitt berechnet und als Vorhersagewert herausgegeben.

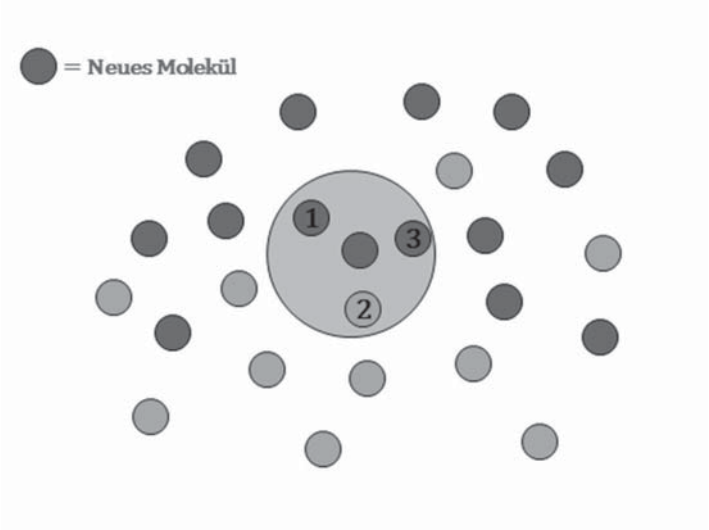


Abbildung 3: KNN Klassifikation mit *k* = 3 Nachbarn. Das rot gefärbte neue Molekül wird der blauen Klasse zugeordnet, da 2 von 3 nächsten Nachbarn dieser Klasse angehören und die Klassenzuweisung auf einer Mehrheitsentscheidung beruht.

1.5.3.1 Der „Fluch der Dimensionalität“

In niedrig-dimensionalen Räumen ist die Funktionsweise distanzbasierter Methoden leicht nachvollziehbar. Sobald allerdings die Daten in hoch-dimensionalen Räumen vorliegen, ist dies nicht mehr der Fall. In hoch-dimensionalen Räumen kann sich die Leistung distanzbasierter Methoden verschlechtern. Dieses Phänomen wird auch als der „Fluch der Dimensionalität“ bezeichnet [53–58]. Es kann unter bestimmten Umständen gezeigt werden, dass es mit zunehmender Dimensionalität zu einem Kontrastverlust zwischen nahe gelegenen und weit entfernten Nachbarn kommt [59]. Genauer gesagt nimmt die Varianz der Werte der Distanz-Methode ab, bis schließlich fast alle Punkte equidistant sind. Folglich kann nicht mehr so einfach zwischen nah gelegenen und weit entfernten Nachbarn unterschieden werden. Zudem können distanzbasierte Methoden

in hohen Dimensionen instabil sein, denn selbst kleine Änderungen im Deskriptorraum können zu großen Änderungen in den k-nächsten Nachbarschaften führen [40].

Dieser Kontrastverlust hat eine Reihe unterschiedlicher Gründe. Ein Grund ist die spärliche Besetzung der Daten in einem hoch-dimensionalen Deskriptorraum und ein weiterer Grund können irrelevante Variablen sein. Gewöhnlicher Weise sind nicht alle unabhängigen Variablen im hoch-dimensionalen Deskriptorraum wichtig für die Unterscheidung von neuartigen und gewöhnlichen Molekülen, sondern nur wenige Variablen. Wenn nun über viele irrelevante Variablen gemittelt wird und die Entfernung zwischen den Molekülen hauptsächlich von Rauschen dominiert wird, kommt es somit zum Kontrastverlust. Dieses Problem könnte durch den Ausschluss irrelevanter Variablen oder die Projektion der Daten in einen niedriger-dimensionalen Unterraum gelöst werden, was nicht trivial ist [57, 40].

Nicht alle distanzbasierten Methoden sind gleichermaßen anfällig für den „Fluch der Dimensionalität“. Die Anfälligkeit der ungewichteten Euklidischen und Manhattan Distanz konnte gezeigt werden [57]. Ähnlichkeits-/ Distanzmaße wie der Tanimoto Koeffizient, welcher alle abhängigen Variablen nicht gleich behandelt, sondern nur die Variablen beurteilt, für die beide Moleküle Einträge ungleich Null haben, ist weniger anfällig für den „Fluch der Dimensionalität“ [60–62, 40].

1.5.4 Random Forests (RF)

RF sind eine sehr beliebte und leistungsstarke, nicht-parametrische Technik des Maschinellen Lernens und können sowohl bei Klassifikations- als auch bei Regressionsproblemen angewendet werden [19]. Sie bestehen aus einem Ensemble (siehe Kapitel 1.5.10) von Entscheidungsbäumen. Entscheidungsbäume teilen die Trainingsdatenpartition X rekursiv in disjunkte Teilmengen von X auf. Hierbei wird mit X als Wurzel gestartet. Von da aus werden die Knoten über Kanten/ Ecken erreicht. Die Knoten, von denen Kanten ausgehen, werden Nicht-terminale Knoten genannt. Die Idee ist es für bestimmte unabhängige Variablen an jedem nicht-terminalen Knoten Kriterien auszuwählen, sodass die Teilmengen absteigend reiner im Hinblick auf die Klassenzusammensetzung sind [19]. Es können Kriterien vorgegeben werden, bei denen die Aufteilung gestoppt wird, z.B. wenn die Klassen rein sind, oder eine gewisse Anzahl an Molekülen den Knoten erreicht hat. In diesen Fällen entstehen sogenannte terminale Knoten (ohne Aus-

gang). Diese werden auch als Blatt bezeichnet. Jedem terminalen Knoten ist eine Klasse zugeordnet. Es wird angenommen, dass es sich bei $n_{l,c}$ um die Anzahl an Molekülen der Trainingsdatenpartition der Klasse c im Terminalen Knoten l handelt, welchem ein bisher ungesehenes Molekül $\boldsymbol{x_0}$ zugeordnet wurde. Des Weiteren wird angenommen, dass es sich bei $n_{l,t}$ um die Gesamtmenge an Trainings-Molekülen handelt, welche diesem Terminalen Knoten zugeordnet wurden. Vergleichbar wie beim KNN wird nun das Verhältnis der beiden Größen berechnet ($q_c = n_{l,c}/n_{l,t}$). Die Klassenzuweisung erfolgt basierend auf dem maximalen Wert q_c für eine bestimmte Klasse $c \in \{1, \dots, m\}$, wobei m die Anzahl der Klassen darstellt [40].

Wie bereits beschrieben nutzt der RF ein Ensemble an Entscheidungsbäumen [63]. Hierbei wird jeder Baum auf Basis einer Bootstrap (siehe Kapitel 1.5.10) Stichprobe der Trainingsdatenpartition gebildet, sodass diverse Teilmengen der Trainingsdaten resultieren. Für jeden Knoten wird das beste Aufspaltungs-Kriterium inmitten einer zufällig gezogenen Teilmenge an unabhängigen Variablen ausgewählt. Hierdurch kommt es zu einer höheren Randomisierung des Trainings-Vorgangs. Jedes neue Testobjekt $\boldsymbol{x_0}$ wird von jedem Ensemble-Mitglied einmal vorhergesagt. Die Klassenzuweisung erfolgt basierend auf einer Mehrheitsentscheidung der Ensemble-Mitglieder [40]. Im Fall der Regression basiert die Vorhersage auf der Mittelung der Einzelvorhersagen der Ensemble-Mitglieder [63–65].

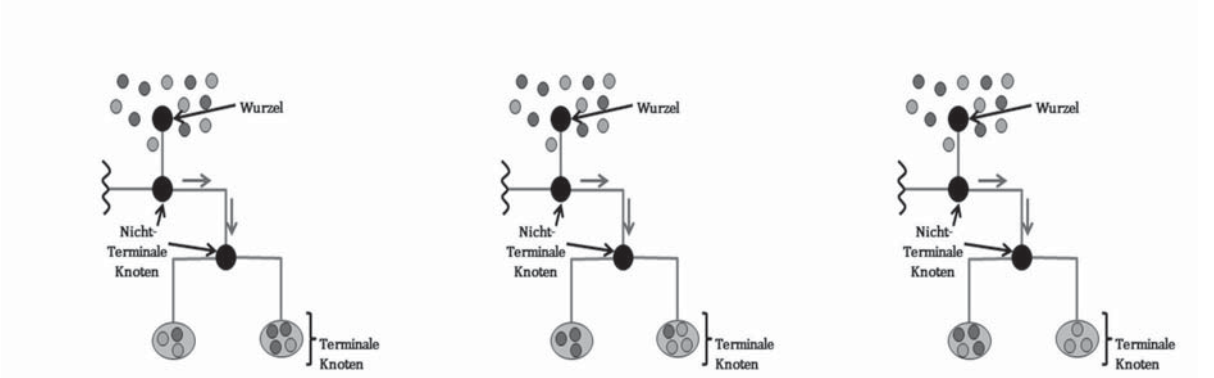


Abbildung 4: Drei einzelne Entscheidungsbäume. Die Entscheidungsbäume sollen den RF als ein Ensemble bestehend aus einer Vielzahl unterschiedlicher Entscheidungsbäume symbolisieren. In der Regel werden 100-300 Bäume verwendet.

Der RF im Regressionsmodus (**Random-Forest-Regression (RFR)**) ähnelt sehr der entsprechendenen Klassifikationstechnik. Ein entscheidender Unterschied ist allerdings,

dass der Mittlere Quadratische Fehler (MSE) beim Konstruieren der einzelnen Entscheidungsbäume als Kriterium herangezogen wird. Außerdem wird die vorhergesagte Variable $\hat{y}$ als Mittelwert über alle Entscheidungsbäume berechnet [66].

1.5.5 Support Vector Machines (SVM)

SVM gehören ebenfalls mit zu den leistungsfähigsten Techniken aus dem Bereich des Maschinellen Lernens, welche sowohl bei Klassifikationsproblemen als auch bei Regressionsproblemen angewendet werden können [67, 68].

Prinzipiell suchen SVM eine Hyperebene (auch Entscheidungsebene genannt), welche zwei Klassen voneinander trennt, so dass die kleinste Distanz aller Moleküle zur Hyperebene, die sogenannte Margin (engl.: Margin) maximiert wird [69, 70]. Es wird angenommen, dass die Moleküle entweder der Klasse -1 oder der Klasse $+1$ angehören. Die Hyperebene wird wie folgt definiert:

$$f(\boldsymbol{x}) = h(\boldsymbol{x})^T \cdot \boldsymbol{b} + b_0 = 0,$$

wenn die Moleküle nicht linear klassifizierbar sind, dann wird die Funktion $h(\cdot)$ benötigt um die unabhängigen Variablen zu transformieren. Bei $\boldsymbol{b}$ handelt es sich um den Normalvektor der Hyperebene und bei b_0 um den y-Achsenabschnitt [40].

Wie bereits erwähnt, trennt die Hyperebene die beiden Klassen, dabei wird das Vorzeichen von $f(\boldsymbol{x_0})$ benutzt um ein bisher ungesehenes Molekül einer Klasse zu zuordnen. Als „Support Vektoren“ werden die Vektoren bezeichnet, welche die folgende Bedingung erfüllen:

$$f(\boldsymbol{x}) = \pm 1,$$

sie liegen genau auf der Margin, denn sie definieren die Hyperebene. Die Inverse der Norm von $\boldsymbol{b}$ definiert die Weite der Margin. Der absolute Wert von $f(\boldsymbol{x})$ ist ein Vielfaches der Weite der Margin und spiegelt somit die Entfernung zur Hyperebene wieder. Je größer dieser Wert ist, desto weiter ist die Distanz zur Hyperebene [40].

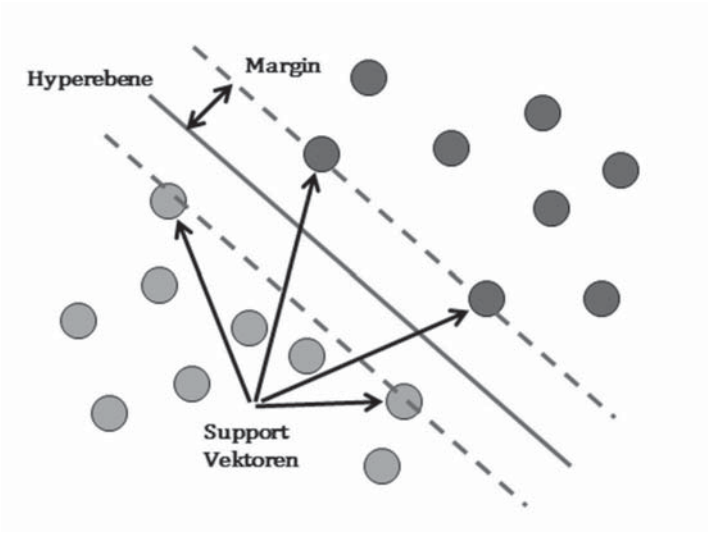


Abbildung 5: Definition einer Hyperebene im linearen Fall durch SVMs. Die Moleküle, die sich auf der gestrichelten Linie befinden, entsprechen den Support Vektoren. Sie sind der Hyperebene am nächsten und bestimmen deren Lage.

Wenn die Moleküle nicht voneinander trennbar sind, würden sich Moleküle der Trainingspartition innerhalb der Margin befinden oder sogar auf der falschen Seite der Trennebene, in diesem Fall müssen sogenannte Schlupfvariablen (engl.: Slack Variables) eingeführt werden und ebenfalls in den Optimierungsschritt integriert werden. Hieraus resultieren dann sogenannte Soft-Margin Hyperebenen [68], bei denen es erlaubt ist, dass sich Moleküle der Trainingspartition nicht nur innerhalb der Margin befinden, sondern auch auf der falschen Seite der Hyperebene. Der Regularisierungsparameter C (engl.: Misclassification Cost) kontrolliert den Kompromiss zwischen einer Minimierung des Vorhersagefehlers und der Maximierung der Weite der Margin [40].

Um die Hyperebene möglichst effizient berechnen zu können, benutzen SVM eine äquivalente Entscheidungsfunktion, welche auch als dual bezeichnet wird:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \cdot \alpha_i \cdot k(\boldsymbol{x}, \boldsymbol{x_i}) + b_0.$$

In diesem Fall steht α_i für einen Lagrange-Multiplikator (Erhaltung durch Lösung des Optimierungsproblems) und $k(\mathbf{a}, \mathbf{b}) = h(\mathbf{a})^T \cdot h(\mathbf{b})$ repräsentiert eine Kernelfunktion. Es ist hierbei wichtig sich für eine geeignete Kernelfunktion zu entscheiden, denn diese Auswahl ist einfacher als sich mit Transformationen $h(\cdot)$ zu beschäftigen. Eine sehr weit

verbreitete Kernelfunktion, welche bei nicht-linearen soft-margin SVMs breitflächig Anwendung findet , ist die Radial-Basis-Funktion (RBF) [68, 40].

Bei der **Support Vector Regression (SVR)** handelt es sich um eine Regressionsvariante der SVM. Die SVR basiert, mit ein paar kleinen Ausnahmen, auf denselben Grundlagen (Minimierung des Fehlers, Maximierung der Margin, Toleranz eines gewissen Fehleranteils) wie die SVM. Zum einen werden kontinuierliche abhängige Variablen verwendet und zum anderen wird mit dem Hyperparameter ε ein Toleranz-Bereich definiert, in welchem sich alle wahren Vorhersagen befinden müssen.

1.5.6 Neuronale Netze (engl.: Neural Networks (NN))

Bei den sogenannten Feed Forward Neuronalen Netzen handelt es sich um eine häufig angewendete Klassifikations- und Regressionsmethode. Die Modelle sind zweistufig und sie bestehen aus einem sogenannten Input-Layer, einem Hidden-Layer und einem Output-Layer [15]. Das Ergebnis des Hidden-Layers passiert eine nicht-lineare Aktivierungs-Funktion, bei welcher es sich für gewöhnlich entweder um eine logistische Funktion oder um eine tanh Funktion handelt. Ohne Aktivierungsfunktion würde das Modell zu einem großen linearen Regressionsmodell herunterbrechen [71]. NN sind eine sehr leistungsfähige Technik, allerdings neigen sie zur Überanpassung der Modelle, aus diesem Grund ist eine sorgfältige Regularisierung wichtig [72, 73, 40].

Der Input-Layer wird bestimmt durch die Dimension der unabhängigen Variablen, der Hidden-Layer muss angepasst werden und der Output-Layer wird je nach Fragestellung ausgewählt. Neben diesen Layern gibt es noch andere Netzwerkparameter, welche angepasst werden müssen. Im Fall der Klassifikation erfolgt die Klassenzuweisung auf Basis des Ausgabe-Knotens mit dem größten Wert, das bedeutet es gibt einen Knoten für jede Klasse. Für den Fall, dass in einem Feed-Forward-NN mit Multilayer mit der Back-Propagation Regel [74] die Summe quadrierter Fehler minimiert wird und die Klasse entsprechend kodiert ist [0 1] [1 0] und weiterhin ein 2 Klassen-Fall vorliegt, dann schätzt diese Ausgabe (engl.: Output) Klassenzugehörigkeits-Wahrscheinlichkeiten [75, 40] (Näheres zu Klassenzugehörigkeits-Wahrscheinlichkeiten siehe: Kapitel 3.1.2). (Eine gute Schätzung hängt natürlich davon ab wie gut die Funktion $\hat{f}$ schätzt).

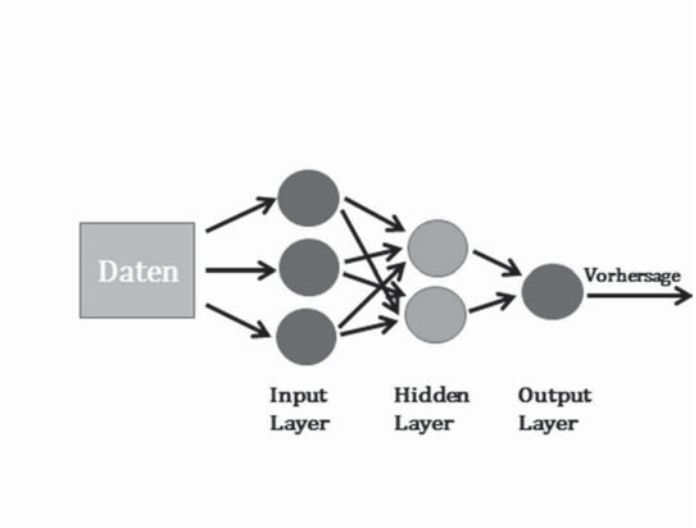


Abbildung 6: Schematische Darstellung des Aufbaus eines Feed-Forward-NN-Modells. Es besteht aus einem Input-Layer, einem Hidden-Layer und einem Output-Layer.

Feedforward-NN werden häufig auch als Ensemble verwendet (siehe 2.10.5). Hierbei wird erwartet, dass durch Mehrheitsentscheid vieler verschiedener NN (Modelle) der Klassifikationsfehler reduziert wird [76]. Die Ensemble Mitglieder werden hierbei häufig durch Bagging generiert [77, 40].

1.5.7 Bayes-Klassifikator (engl.: Naive Bayesian Classifier (NBC))

Unter Benutzung des Bayes Theorem wird die Wahrscheinlichkeit, dass ein bisher ungesehenes Molekül $\boldsymbol{x_0}$ zur Klasse c gehört, folgendermaßen berechnet [78]:

$$p(C = c|X = \mathbf{x_0}) = \frac{p(C = c) \cdot p(X = \mathbf{x_0}|C = c)}{p(X = \mathbf{x_0})} \text{ [39]}$$

Hier steht $p(C = c)$ für die A-priori-Wahrscheinlichkeit der Klasse c und $p(X = \mathbf{x_0}|C = c)$ für die klassenbedingte Dichte. Aus dem Produkt dieser beiden Größen dividiert durch einen Skalierungsfaktor wird die Klassenzugehörigkeits-Wahrscheinlichkeit $p(C = c|X = \mathbf{x_0})$ erhalten. Die klassenbedingte Dichte kann auch folgendermaßen ausgedrückt werden, wenn angenommen wird, dass alle Variablen bei gegebener Klasse c bedingt unabhängig sind:

$$p(X = \mathbf{x_0}|C = c) = \prod_{j=1}^{p} p\big(X_j = x_{0,j}\big|C = c\big).$$

Unter diesen Annahmen wird der sogenannte Bayes-Klassifikator erhalten. Da die Annahme der Unabhängigkeit häufig nicht zutreffend ist, kommt es zu einer Vereinfachung der klassenbedingten Dichteschätzung. Die Klassenzugehörigkeits-Wahrscheinlichkeits-Schätzungen sind allerdings nur korrekt, wenn die Unabhängigkeits-Annahme zutreffend ist. Jedoch ist die Klassifikations-Leistung oftmals trotzdem gut, obwohl die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen nicht so gut sind [79]. Denn für die Klassenzuweisung ist ausschließlich die maximale Klassenzugehörigkeits-Wahrscheinlichkeit wichtig. Sofern die maximale Klassenzugehörigkeits-Wahrscheinlichkeit der richtigen Klasse entspricht, ist die Zuweisung folglich korrekt [40].

1.5.8 Lineare Diskriminanz Analyse (engl.: Linear Discriminant Analysis (LDA))

Bei der LDA wird jede Klasse als eine multivariate Normalverteilung mit der gleichen Kovarianzmatrix aber einem unterschiedlichen Mittelwertvektor modelliert. Die Dichte $\hat{f}_c(\boldsymbol{x_0})$ für ein neues ungesehenes Molekül kann bei gegebenem Datensatz X und der Klassenzugehörigkeit $C = c$ $(c = 1, \dots, m)$ wie folgt ermittelt werden [40]:

$$\hat{f}_c(\boldsymbol{x_0}) = \frac{1}{(2\pi)^{\frac{p}{2}} \cdot |C|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x_0}-\bar{x}_c)^T \Sigma^{-1} (\boldsymbol{x_0}-\bar{x}_c)},$$

bei C handelt es sich um die kombinierte (engl.: pooled) Kovarianzmatrix der Moleküle $\Sigma = 1/(n_1 + \dots + n_m - 1) \cdot (X_1^T X_1 + \dots + X_m^T X_m)$. Die Moleküle der Klassen 1 bis m bilden jeweils sogenannte Submatrizen (X_1 (n_1* p); X_m (n_m*p)) von der ursprünglichen Matrix Xund es wird ferner angenommen, dass diese zentriert sind. Wenn angenommen wird, dass es sich bei $p(C = j)$ um die A-priori-Wahrscheinlichkeit für die Klasse j handelt ($\sum_{j=1}^{m} p(C = j) = 1$), dann ergibt sich für ein bisher ungesehenes Molekül $\boldsymbol{x_0}$ zugehörig einer Klasse c folgende Klassenzugehörigkeits-Wahrscheinlichkeit [40]:

$$p(C = c | X = \boldsymbol{x_0}) = \frac{\hat{f}_c(\boldsymbol{x_0}) \cdot p(C = c)}{\sum_{j=1}^{m} \hat{f}_j(\boldsymbol{x_0}) \cdot p(C = j)}.$$

Hieraus ergibt sich eine Entscheidungsebene zwischen den Klassen, welche linear in x ist. Bei dieser handelt es sich um eine Hyperebene, welche sich aus p Dimensionen „zusammensetzt“. Ein bisher ungesehenes Molekül wird nun der Klasse zugewiesen, für die die Klassenzugehörigkeits-Wahrscheinlichkeit am höchsten ist. Im Grunde genommen hängt die Klassenzugehörigkeits-Wahrscheinlichkeit im Kern von der Mahalanobis Dis-

tanz eines neuen Moleküls $\boldsymbol{x_0}$ zum Klassenmittelwerts-Vektor $\overline{\boldsymbol{x}}_{\boldsymbol{c}}$ bei gegebener kombinierter Kovarianz-Matrix Σ ab [40, 15, 71, 48].

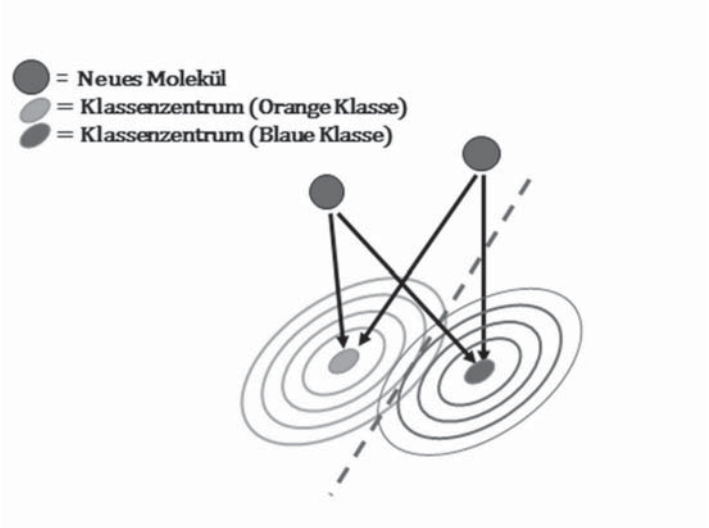


Abbildung 7: Ein neues Molekül wird etwas vereinfacht ausgedrückt in Abhängigkeit von der Mahalanobis Distanz zum Klassenmittelwerts-Vektor einer Klasse (entweder der orangen Klasse oder der blauen Klasse), bei gegebener kombinierter Kovarianz-Matrix, zugeordnet.

1.5.9 Partial Linear Discriminant Analysis (PLSDA)

Die PLSDA [80] ist die entsprechende Klassifikationstechnik zur PLS (Regression) welche bereits beschrieben wurde [32]. Die Methode eignet sich besonders, wenn hochdimensionale Daten vorliegen und klassische Klassifikationstechniken wie beispielsweise die LDA aufgrund von Problemen mit der Singularität Schwierigkeiten haben [22].

Für die PLSDA wird wie bei der PLS von folgendem Modell ausgegangen:

$$Y = XB + E\ .$$

Danach kommt es wie bereits bei der PLS beschrieben zur Zerlegung von X und Y in Scores und Loadings.

$$X = TP^T + E_x\ (1)$$

$$Y = TQ^T + E_Y\ (2)$$

Allerdings ist hier der Unterschied, dass Y nun die Klasse (Y besitzt m Spalten (m ist die Anzahl der Klassen)) enthält [22].

Grundsätzlich gibt es zwei Möglichkeiten, wie neue Moleküle den Klassen zugeordnet werden. Die erste Möglichkeit ist mittels PLS für jede Klasse eine Vorhersage zu machen und dann die Klasse auszuwählen, für die der vorhergesagte Wert am größten war. Die zweite Möglichkeit ist mittels PLS die Scores zu schätzen und dann die LDA auf den Scores anzuwenden [22, 32].

1.5.10 Ensemble Methoden

Ensemble Methoden nutzen eine Reihe von unterschiedlichen Klassifikationstechniken mit verschiedenen Stärken und Schwächen. Die Ausgaben der einzelnen Mitglieder eines Ensembles für ein bisher ungesehenes Molekül werden gewichtet und zu einer Vorhersage zusammengeführt. Das Ziel ist es, hierbei eine Klassifikationstechnik zu schaffen, welche besser ist als die einzelnen Mitglieder des Ensembles [76]. Es gibt unterschiedliche Möglichkeiten diese Mitglieder zu generieren [40].

Eine sehr populäre Möglichkeit nennt sich **Bagging** [77]. Der Begriff Bagging ist ein Akronym aus den Worten Bootstrap und Aggregation. Das **Bootstrapping** ist neben der Kreuzvalidierung ein Verfahren zur wiederholten Stichprobenziehung [81], bei welchem so viele Moleküle aus dem Datensatz zufällig mit Zurücklegen gezogen werden, wie Moleküle in dem Datensatz enthalten sind. Durch das Zurücklegen sind einige Moleküle wahrscheinlich mehrfach in der Stichprobe enthalten und andere überhaupt nicht. Beim Bagging werden viele Bootstrap-Stichproben der Trainingsdaten gezogen und mit jeder Stichprobe wird anschließend eine Klassifikationstechnik trainiert. Dann stimmen alle Klassifikationstechniken über die Klassenzugehörigkeit ab und dem bisher ungesehenen Molekül wird per Mehrheitsentscheid die Klasse zugewiesen. Hierbei ist es wichtig, dass die Klassifikationstechniken unterschiedliche Stärken und Schwächen haben, da andernfalls keine Verbesserung der Vorhersage zu erwarten ist. Diversität kann auch hineingebracht werden, indem unterschiedliche Variablen-Submengen (der unabhängigen Variablen) verwendet werden [82, 83]. Diese Technik in Kombination mit dem Bagging wird beispielsweise beim RF angewendet [63, 40].

1.6 Arbeitsbereich (AB) (engl.: Applicability Domain)

1.6.1 Einleitung

Sowohl bei der Bewertung von Regressions- als auch von Klassifikationsmodellen wird die Annahme gemacht, dass der MSE_{Test} bei der Regression und die Test-Fehler-Rate bei der Klassifikation für die Vorhersage zukünftiger ungesehener Moleküle vergleichbar sind. Allerdings ist es niemals möglich ein Modell zu trainieren welches den kompletten potentiell möglichen Datenraum (alle möglichen Chemotypen) berücksichtigt. Somit ist es nur nachvollziehbar, dass zukünftige Moleküle, welche nicht gut in den Trainingsdatenraum eingebettet sind, vermutlich einen höheren Fehler aufweisen können als Moleküle, welche den Trainingsdaten ähnlich sind. Wenn nun Vorhersagen mit einer deutlich höheren Fehlerwahrscheinlichkeit vermieden werden sollen, sollte der Bereich in dem das Klassifikationsmodell angewendet werden kann, auf den Bereich beschränkt werden, welcher von den Trainingsdaten abgedeckt wird [40].

Dieser Bereich wird gewöhnlich als Arbeitsbereich (AB) (engl.: Applicability Domain) bezeichnet [84]. Vom Prinzip definiert der AB den Bereich im chemischen Strukturraum indem sich die „gewöhnlichen“ Moleküle befinden. Sobald die Grenzen des AB definiert werden, kann dies entweder als Ausreißer-Erkennung oder als Erkennung ungewöhnlicher Moleküle oder als Neuartigkeits-Erkennung betrachtet werden [40]. Auf dem Gebiet der Statistik oder dem Gebiet des Maschinellen Lernens gibt es eine Vielzahl an Methoden für die Entdeckung von Ausreißern [85], außergewöhnlichen Objekten [86] oder neuartigen Objekten [87, 88], welche auch auf dem Gebiet der QSAR/ Chemometrik/ Chemieinformatik Anwendung finden.

In den zuletzt genannten Bereichen findet die Definition von Netzeva [84] für den AB breite Akzeptanz, welche den AB eines (Q)SAR Modells als den chemischen Struktur- und Eigenschaftsraum (abhängige Variable) definiert, indem das Modell Vorhersagen mit einer gegebenen Wahrscheinlichkeit machen kann. Netzeva hat diese Definition für Regressionsmodelle formuliert, in denen die abhängige Variable kontinuierlich ist. Im Regressionsfall kann folglich die Vorhersage eines zukünftigen Moleküls außerhalb des Bereichs liegen, indem sich sonst die y-Werte der Trainingsdaten befinden. Somit kann auch der y Datenraum genutzt werden um den AB zu begrenzen. Bei der Klassifikation hingegen ist dies nicht möglich, denn es kann keine Klasse vorhergesagt werden, welche

nicht zuvor zum Training des Modells verwendet wurde. Hieraus folgt, dass lediglich der chemische Strukturraum benutzt werden kann um den AB zu definieren. Da die chemische Struktur der Moleküle einfach durch den Deskriptor beschrieben werden kann, wird der AB in der Regel im Deskriptorraum definiert [40].

Im Folgenden wird, bis auf ein paar wenige Ausnahmen, nur auf die Definition des AB im Klassifikationsfall eingegangen. Es gibt zwei unterschiedliche Situationen, die dazu führen, dass Vorhersagen von Klasseninformationen unzuverlässig sind. Im ersten Fall könnten sich zukünftige Moleküle weit weg von den Molekülen des Trainingsdatensatzes in spärlich bevölkerten Regionen des Deskriptorraums befinden. Im zweiten Fall könnten sich zukünftige Moleküle in einer Region im Deskriptorraum befinden, in der sich die Klassen überlappen. Folglich sind diese Moleküle gut eingebettet im Deskriptorraum, aber sie sind ungewöhnlich bezüglich ihrer Klassenzugehörigkeit und können deshalb auch nur schwer vorhergesagt werden. In beiden Fällen greift die Definition des AB, denn sie beinhaltet sowohl den Deskriptorraum als auch die Vertrauenswürdigkeit der Vorhersage. Folglich gibt es zwei unterschiedliche Möglichkeiten die Grenzen für den AB zu setzten [40].

Die erste Möglichkeit ist, den AB nur im Deskriptorraum zu definieren. Hierbei wird angenommen, dass der Deskriptorraum, welcher durch die Moleküle der Trainingspartition abgedeckt wird, mit einer bestimmten Wahrscheinlichkeit vorhergesagt werden kann. Diese Wahrscheinlichkeit wird bestimmt, indem bisher ungesehene Testmoleküle zur Validierung verwendet werden. Moleküle, die sich außerhalb des Bereichs der Trainingsmoleküle befinden, können größere Vorhersagefehler hervorrufen (siehe Abbildung 8). Dieser Ansatz, welcher nach neuen oder ungewöhnlichen Molekülen sucht, oder nach sogenannten Ausreißer-Molekülen, wird von nun an als Neuartigkeits-Detektion bezeichnet. Üblicherweise werden neuartige Moleküle nach ihrer Detektion in das Modell mit aufgenommen und der AB wird entsprechend erweitert; beispielsweise wenn neue Moleküle synthetisiert werden [40].

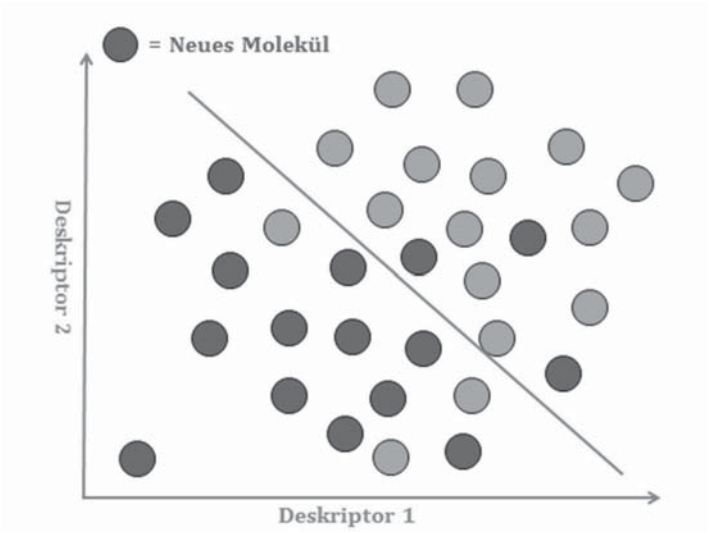


Abbildung 8: Abbildung der Situation der Neuartigkeits-Detektion am Beispiel eines Deskriptors. Dieser besteht zur Veranschaulichung aus lediglich zwei Spalten. Moleküle, die weit entfernt von den anderen Molekülen lokalisiert sind, werden als neuartig eingestuft. Die Detektion neuartiger Moleküle basiert ausschließlich auf der Information der unabhängigen Variablen. Die Information über die Klassenzugehörigkeit fließt nicht mit ein.

Die zweite Möglichkeit ist, die zu erwartende Zuverlässigkeit der Vorhersagen zu benutzen um den AB zu definieren. Im Gegensatz zu dem Ansatz der Neuartigkeits-Detektion führt dieser Ansatz nicht zwingend zu einem zusammenhängenden Deskriptorraum des AB, aber es werden ungewöhnliche Moleküle erfasst, deren Gemeinsamkeit es ist, ungewöhnlich im Bezug auf ihre Klassenzugehörigkeit zu sein. Diese Moleküle können sehr gut im molekularen Deskriptorraum eingebettet sein, sind aber trotzdem schwierig vorhersagbar. Dieser Ansatz wird von nun an als Vertrauens-Schätzung bezeichnet, da versucht wird, die Zuverlässigkeit einer bestimmten Vorhersage zu schätzen (siehe Abbildung 9) [89, 40].

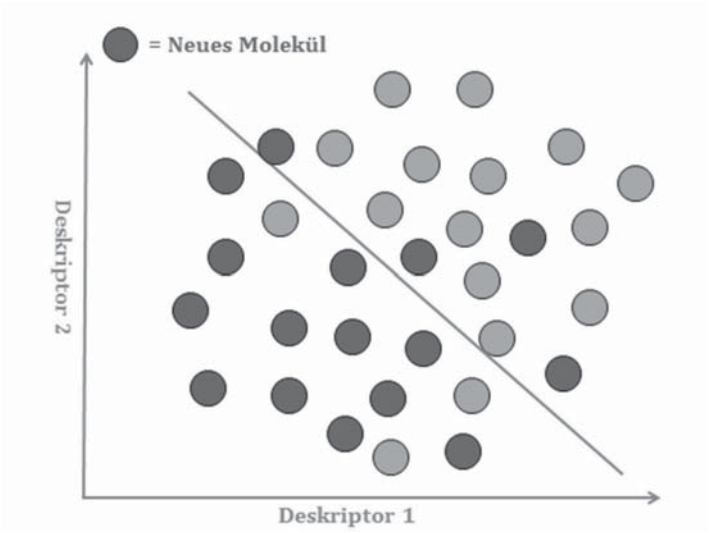


Abbildung 9: Veranschaulicht ist der Ansatz der Vertrauens-Schätzung am Beispiel eines Deskriptors mit zwei Spalten. Mit diesem Ansatz sollen Moleküle identifiziert werden, dessen Vorhersagen wahrscheinlich unzuverlässig sind. Hierbei wird im Gegensatz zum Ansatz der Neuartigkeits-Detektion, die Information über die Klassenzugehörigkeit mit einbezogen. In vielen Klassifikationsproblemen können die Klassen nicht vollständig getrennt werden, hierbei werden oftmals die meisten Fehler nahe der Entscheidungsebene gemacht. Deshalb gelten Vorhersagen für Moleküle, welche sich nah der Entscheidungsebene befinden als weniger zuverlässig als Vorhersagen für Moleküle, die weit entfernt von der Entscheidungsebene lokalisiert sind.

Beide beschriebenen Ansätze sind gebräuchlich um den AB von Klassifikationsmodellen zu definieren. Es gibt allerdings einige unterschiedliche Bezeichnungen/ Benennungen [90–92]. Beide Ansätze verbindet das gemeinsame Ziel, die Zuverlässigkeit von Vorhersagen zu erhöhen, indem Moleküle zurückgewiesen werden, die wahrscheinlich zu fehlerhaften Vorhersagen führen würden [40, 93].

1.7 Kalibrierung von Wahrscheinlichkeitsschätzern

1.7.1 Einleitung

In vielen Bereichen, wie beispielsweise im Bankwesen bei der Vergabe von Krediten, bei der Analyse vom Kaufverhalten verschiedener Kunden oder in der frühen Wirkstoffentwicklung bei der Auswahl von möglichen Wirkstoffen oder Eliminierung von potentiell toxischen Substanzen, werden Entscheidungen basierend auf Wahrscheinlichkeitsschätzungen von Klassifikationstechniken getroffen. Ein Beispiel wäre, wenn nur mit Molekülen weitergearbeitet werden soll, welche mit einer Wahrscheinlichkeit von mindestens

80% an einer bestimmten Zielstruktur aktiv sind. In diesen Fällen ist es wichtig, gut kalibrierte Wahrscheinlichkeiten vorhersagen zu können. Das bedeutet, dass die vorhergesagten Wahrscheinlichkeiten möglichst nah an der wahren Wahrscheinlichkeit liegen sollen. Bei starken Abweichungen der vorhergesagten Wahrscheinlichkeit vom wahren Wert kann es sonst zu ungerechtfertigtem Verwerfen oder fälschlichem Annehmen von z.B. potentiellen Wirkstoffkandidaten kommen, wobei der Anwender von einer hohen Wahrscheinlichkeit für die Richtigkeit seiner Schätzungen ausgeht. In diesen Fällen wird von schlecht kalibrierten Klassifikatoren gesprochen.
Kalibrierung wird definiert als den Grad der Approximierung der vorhergesagten Wahrscheinlichkeiten an die wahre Wahrscheinlichkeit [94]. Wie bereits erwähnt, geben die meisten Klassifikationstechniken eine Art Wahrscheinlichkeitsschätzer aus, welcher die Zuverlässigkeit der Vorhersage quantifizieren soll. Allerdings sind diese Wahrscheinlichkeitsschätzungen oftmals schlecht kalibriert und liegen somit nicht unbedingt nah an der wahren Wahrscheinlichkeit. Beispielsweise geben viele der bisher vorgestellten Klassifikationstechniken Klassenzugehörigkeits-Wahrscheinlichkeiten heraus. Wenn nun aber die Annahmen der zugrunde liegenden Klassifikationstechnik verletzt werden, kann dies zu schlecht kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeiten führen. Folglich sollten sie rekalibriert werden.

1.7.2 Kalibriermethoden

Die beiden in der Literatur am häufigsten eingesetzten Kalibriermethoden um Vorhersagen aus Modellen auf Klassenzugehörigkeits-Wahrscheinlichkeiten abbilden zu können, sind die Platt-Kalibrierung und die Isotonische Regression [95–98].

Die **Platt Kalibrierung** wurde im Jahre 1999 von John Platt vorgestellt um Vorhersagen aus SVMs in Klassenzugehörigkeits-Wahrscheinlichkeiten zu transformieren.

Die ausgegebenen, rohen Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der SVM (Distanzen zur Hyperebene), werden mit f dargestellt. (Auf diese unkalibrierten Schätzer wird in Kapitel 3.1 näher eingegangen). Um an die kalibrierten Wahrscheinlichkeiten zu gelangen wird die Ausgabe der SVM sigmoid transformiert:

$$P(y = 1|f) = \frac{1}{1 + \exp(af + b)}.$$

Dabei werden die Parameter a und b durch eine Minimierung der negativen log Likelihood auf Basis einer Trainingsdatenpartition (f_i, y_i) angepasst. Hierbei handelt es sich um eine Kreuzentropie-Fehler-Funktion:

$$min - \sum_i y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

dabei ist

$$p_i = \frac{1}{1 + \exp(af_i + b)}.$$

Um eine Überanpassung an die Trainingsdatenpartition zu vermeiden, wird eine unabhängige Kalibrierdatenpartition benötigt. Nur so können gute Klassenzugehörigkeits-Wahrscheinlichkeiten geschätzt werden. Dies ist insofern ein Nachteil, denn diese Daten können somit weder für die Modellbildung noch für die Modellselektion (inklusive der Parameterauswahl) benutzt werden. Platt erstellt sich, um eine Überanpassung zu vermeiden, so genannte out-of-sample data, welche vergleichbar sind mit einer klassischen externen Testdatenpartition. Neben diesem Ansatz gibt es noch andere Ansätze wie zum Beispiel die Kreuzvalidierung (siehe Kapitel 1.4.2.2). [95, 96]. Die Anwendung der Platt-Kalibrierung entspricht der Durchführung einer logistischen Regression.

Die **Logistische Regression** verwendet als Aktivierungsfunktion eine Sigmoidfunktion:

$$\phi(z) = \frac{1}{1 + e^{-z}},$$

bei z handelt es sich in diesem Fall um die Nettoeingabe (Linearkombination aus Gewichtungen und Variablen):

$$z = \boldsymbol{w}^T \boldsymbol{x} = w_0 x_0 + w_1 x_1 + \cdots + w_m x_m$$

Als Eingabe nimmt die Sigmoidfunktion reelle Zahlen entgegen und bildet sie auf das Intervall [0,1] ab. Die Werte, welche die Sigmoidfunktion herausgibt werden als Wahrscheinlichkeit $\phi(z) = P(y = 1|\boldsymbol{x}; \boldsymbol{w})$ interpretiert, dass ein bestimmtes Objekt bei gegebenen, durch Gewichtungen **w** parametrisierten Variablen **x**, zur Klasse 1gehört.

Anschließend kann die vorhergesagte Wahrscheinlichkeit mittels einer Sprungfunktion wieder zurück konvertiert werden in ein binäres Ereignis [66].

Die Anwendung der Platt-Kalibrierung bzw. die logistische Regression ist nicht allein auf die SVM beschränkt, sondern kann auch auf die ausgegebenen Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer anderer Klassifikationstechniken angewendet werden. Bei den rohen Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern handelt es sich um interne Ausgabekriterien, wie beispielsweise die Distanz zur Hyperbene bei der SVM, auf Basis dessen die Einteilung der einzelnen Objekte in eine bestimmte Klasse erfolgt.

Die **Isotonische Regression** wurde 1988 von Robertson vorgestellt [99] und ist eine nicht-parametrische Regressionstechnik, bei der angenommen wird, dass die Funktion aus einer Klasse von isotonen (monoton steigenden) Funktionen ausgewählt wird. Auf Basis dieser Technik haben Zadrozny und Elkan [97, 98] Vorhersagen aus SVM, NBC und ein paar weiteren Techniken kalibriert. Diese Methode ist im Vergleich zur Platt-Kalibrierung viel allgemeiner, da die Mapping-Funktion die einzige Einschränkung hat monoton zu steigen [95]. Angenommen bei f handelt es sich um die Vorhersagen eines Modells, genauer betrachtet um die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer und bei $\boldsymbol{y}$ um die dazugehörigen wahren Klassen, dann ist die Basis Annahme der isotonen Regression:

$$\boldsymbol{y} = m(f) + \varepsilon,$$

wobei es sich bei m um eine monoton steigendende (isotone) Funktion handelt. Bei gegebener Trainingsdatenpartition (f_i, y_i), wird die Isotone Regressionsfunktion $\hat{m}$ gefunden, sodass:

$$\hat{m} = argmin_z \sum (y_i - z(f_i))^2$$

Ein sehr verbreiteter Algorithmus namens PAV (engl.: pair-adjacent violator) [100] zur Berechnung der isotonen Regression findet eine schrittweise, konstante Funktion, welche am besten nach dem Kriterium des mittleren quadratischen Fehlers (MSE) an die Daten angepasst ist [98]. Genau wie bei der Platt-Kalibrierung muss auch bei der isotonen Regression bedacht werden, eine unabhängige Kalibrierdatenpartition zu benutzen, um eine mögliche Überanpassung an die Trainingsdatenpartition zu vermeiden. Alternativ wird hierfür ebenfalls auch die Kreuzvalidierung verwendet.

1.7.3 Zuverlässigkeits-Diagramme

Um die Effizienz der Kalibriermethoden zu visualisieren, können Zuverlässigkeits-Diagramme (engl.: Reliability Diagrams) eingesetzt werden. Für die Erstellung solcher Zuverlässigkeits-Diagramme werden die Ausgaben aus einem Modell, bzw. die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer, in eine bestimmte Anzahl an Segmenten unterteilt. In dieser Arbeit wurden immer zehn Segmente verwendet. Unter dieser Voraussetzung und der Annahme, dass die Ausgaben zwischen 0 und 1 liegen, würden beispielsweise alle Ausgaben zwischen 0 und 0.1 in das erste Segment sortiert werden. Alle Ausgaben zwischen 0.1 und 0.2 würden in das zweite Segment sortiert werden, etc. Für jedes Segment wird der Mittelwert der Ausgaben und der Anteil an Ausgaben, der bezogen auf eine Klasse, richtig vorhergesagt wurde, berechnet (In der Regel wird der Anteil bezogen auf die Klasse 1 berechnet.).

An dieser Stelle kommt ein kleines Beispiel: Angenommen die Ausgabe einer Technik/ eines Modells liegt zwischen 0 und 1 und die Entscheidungssgrenze der Klassifikationstechnik bei 0.5. Wenn die Klassenzugehörigkeits-Wahrscheinlichkeit eines Objektes größer gleich 0.5 ist, dann wird dieses Objekt der Klasse 1 zugeteilt, im anderen Fall der Klasse 0. Die wahren Klassen der Objekte sind ebenfalls bekannt. Es wird angenommen, dass sich zwischen 0 und 0.1 10 Ausgaben {0.0, 0.0, 0.0, 0.05, 0.05 ,0.05, 0.05, 0.1, 0.1, 0.1} befinden. Deren Mittelwert beträgt 0.05. Allen Objekten würde folglich die Klasse 0 zugewiesen werden, da die Ausgaben kleiner als 0.5 sind. Es wird weiter angenommen, dass es sich hierbei {0, 0 ,0 ,0 ,0 ,0 ,0 ,0 ,0 ,1} um die wahren Klassen der Objekte handelt. Der Anteil an Ausgaben, der bezogen auf die Klasse 1, richtig vorhergesagt wurde, beträgt folglich 0.1. (Bezogen auf die Klasse 0 würde dieser Anteil 0.9 betragen).

Dann wird der Mittelwert gegen den Anteil wahrer Vorhersagen aufgetragen. Durch diese Art der Darstellung ergibt die wahre Wahrscheinlichkeit die Winkelhalbierende. Wenn das Modell gut kalibrierte Wahrscheinlichkeiten hervorgebracht hat, dann werden sich die Punkte bzw. dann wird sich der Graph nahe der Winkelhalbierenden befinden [101, 95].

Basierend auf diesen Zuverlässigkeits-Diagrammen wurde von Caruana und Niculescu-Mizil noch eine weitere Variante vorgestellt, welche anstelle der zehn Mittelwerte der Segmente, immer 100 Objekte in ein Segment zusammenfasst. Anschließend werden

von den einzelnen Segmenten (S_x) {S_1=1-100, S_2=2-101,S_3=3-102 etc.} die Mittelwerte berechnet und der Anteil an Vorhersagen aufgetragen, der bezogen auf eine Klasse, richtig vorhergesagt wurde [102].

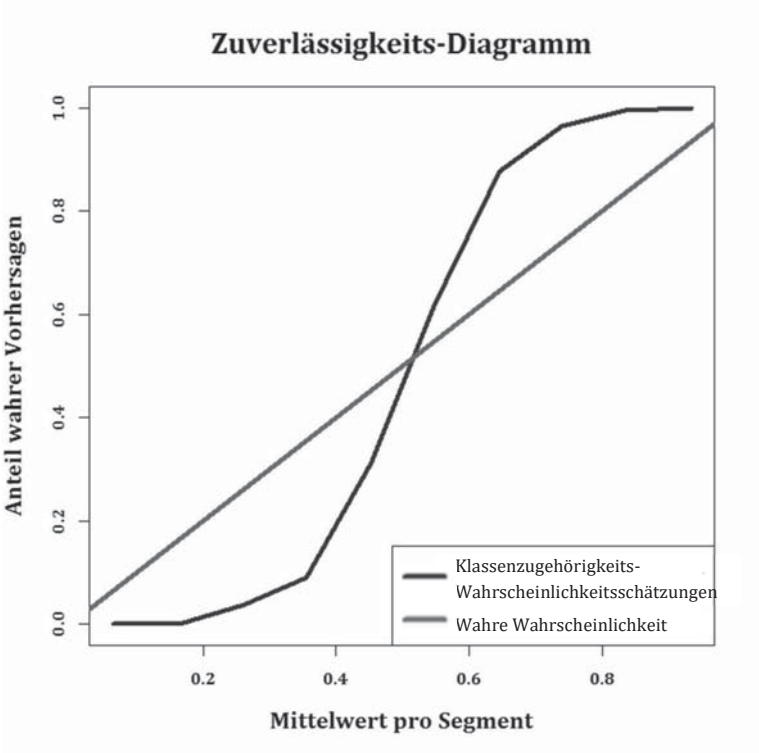


Abbildung 10: Abbildung eines Zuverlässigkeits-Diagramms. Auf der x-Achse ist der Mittelwert jedes Segmentes aufgetragen und auf der y-Achse der Anteil wahrer Vorhersagen bezogen auf eine Klasse. Die rot gekennzeichnete Winkelhalbierende entspricht der wahren Wahrscheinlichkeit. Der blaue Graph repräsentiert beispielhaft die Vorhersagen eines Modells. Der blaue Graph ist nicht sehr nah an der roten Winkelhalbierenden, das lässt die Schlussfolgerung zu, dass dieses Modell keine gut kalibrierten Wahrscheinlichkeiten hervorbringt.

1.8 Conformal Predicton (CP)

Vertrauensschätzer haben das Ziel, die Zuverlässigkeit von Vorhersagen aus Klassifikationsalgorithmen zu quantifizieren. Wie im letzten Abschnitt bereits erwähnt, sind die Vertrauensschätzungen, welche von Klassifikationstechniken herausgegeben werden, oftmals schlecht kalibriert. Was bedeutet, dass die Werte nicht zwingend nah an der wahren Wahrscheinlichkeit liegen müssen. Deshalb ist es schwierig, die unkalibrierten Werte zu benutzen um Entscheidungen zu treffen. Der CP liefert einen Rahmen um nachweislich valide die Zuverlässigkeit von Vorhersagen messen zu können unter der Annahme, dass die Daten unabhängig und identisch verteilt sind [103]. Es gibt unter-

schiedliche Varianten des CP [104]. Die Variante, die bereits erfolgreich in die Chemieinformatik eingeführt wurde, nennt sich Inductive Conformal Prediction (ICP) [105, 106]. Allgemein benutzt der CP eine Vertrauens-Bewertung, welche in der Literatur Nichtkonformitäts-Score genannt wird. Dieser Nichtkonformitäts-Score wird dann reskaliert, sodass ein valides Vertrauens-Maß entsteht. Nach der Reskalierung entstehen so genannte p-Werte, die einen bestimmten Grad an Nichtkonformität der Trainingsdaten beschreiben. Zeigt ein vorherzusagendes Molekül denselben Nichtkonformitäts-Score wie ein Trainingsobjekt, dann liegt derselbe Grad an Nicht-Übereinstimmung mit dem Gesamtdatensatz vor wie derjenige des Trainingsobjektes. Daraus folgt, dass die p-Werte hoch sind, wenn das Molekül typisch (übereinstimmend) ist bezogen auf die betrachtete Klasse und niedrig sind, wenn das Molekül ungewöhnlich (nicht übereinstimmend) ist. Die p-Werte können nun auf zwei verschiedene Arten verwendet werden. Im ersten Fall können sie zusammen mit der vorhergesagten Klasse als Vertrauens-Maß für das betrachtete Molekül ausgegeben werden. Im zweiten Fall können sie benutzt werden um eine Klassifikationspartition für das betrachtete Molekül zu erstellen, welche die wahre Klassenzugehörigkeit mit einem vorgegebenen Vertrauensniveau $1 - \delta$ enthält, hierbei ist $\delta > 0$ das Signifikanzniveau. Die Klassifikationspartition enthält alle Klassen, dessen p-Werte größer als δ sind. Hieraus folgt, dass die Klassifikationspartition auch mehr als eine Klasse enthalten kann, was auch nicht informativ sein kann. Durch diesen Mechanismus kann ein bestimmtes Vertrauensniveau $1 - \delta$. garantiert werden [40].

Im Folgenden wird ein Überblick über den Algorithmus basierend auf Papadopulus [103] gegeben. Im Fall des ICP wird die Trainingsdatenpartition der Größe n in eine geeignete, so genannte geeignete Trainingsdatenpartition (engl.: proper training set (pt)) der Größe n_{pt} und eine Kalibrierdatenpartition (engl.: calibration set (cal)) der Größe n_{cal} unterteilt. ($n = n_{pt} + n_{cal}$). Die pt wird nun benutzt um die Klassifikationstechnik zu trainieren, während die *cal* benutzt wird um die p-Werte für ein neues Molekül für jede mögliche Klasse zu berechnen. Zu diesem Zweck wird ein Nichtkonformitäts-Maß (engl.: Measure) definiert. Hierfür können alle bisher beschriebenen Vertrauens-Maße und noch einige weitere Maße verwendet werden [104]. Wenn Nichtkonformität gemessen wird, schließt dies mit ein, dass extreme Moleküle einen hohen Nichtkonformitäts-Score haben und typische Moleküle eine niedrige Nichtkonformitäts-Bewertung. Somit sind Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen als Konformitäts-Maße zu

bezeichnen. Wenn diese Schätzwerte von 1 abgezogen werden, dann wird das entsprechende Nichtkonformitäts-Maß erhalten. Die Berechnung des ICP ist folgendermaßen:

1) Zuordnung einer Nichtkonformitäts-Bewertung γ_i zu jedem Molekül $\boldsymbol{x_i}$ des cal $(i = 1, \dots, n_{cal})$.
2) Für jedes neue Molekül $\boldsymbol{x_0}$ wird für jede mögliche Klasse c $(c = 1, \dots, m)$ eine Nichtkonformitäts-Bewertung γ_0^c berechnet.
3) Berechnung der p-Werte $\tilde{p}_0(c)$ für jedes neue Molekül $\boldsymbol{x_i}$ für jede mögliche Klasse c :

$$\tilde{p}_0(c) = \frac{\#\{i = 1, \dots, n_{cal}: \gamma_i \geq \gamma_0^c\}}{n_{cal} + 1} = \frac{1}{n_{cal} + 1} \sum_i^{n_{cal}} \mathbb{I}(\gamma_i \geq \gamma_0^c)$$

Die Zuweisung der Klasseninformation erfolgt auf Basis der maximalen $\tilde{p}_0(c)$ Werte

$$\hat{c} = \underset{c}{\operatorname{argmax}}(\tilde{p}_0(c)), \qquad c = 1, \dots, m.$$

Das Vertrauens-Maß für diese Vorhersage ist die Differenz zwischen dem zweitgrößten p-Wert und eins. Die so genannte Glaubwürdigkeit (engl.: Credibility) der Vorhersage ist der größte p-Wert $\tilde{p}_0(c)$. Wenn der p-Wert für eine alternative Klasse relativ hoch ist, dann sinkt das Vertrauen in die Vorhersage. Wenn der p-Wert für eine bestimmte Vorhersage sehr niedrig ist, dann ist die Glaubwürdigkeit der Vorhersage fraglich. Darüber hinaus kann der p-Wert benutzt werden, um eine Klassifikationspartition zu bilden, die die wahre Klasseninformation mit einem vorgegebenen Vertrauensniveau $1 - \delta$ enthält. In diesem Fall gibt der ICP folgende Partition heraus:

$$\{c : \tilde{p}_0(c) > \delta\}, \quad c = 1, \dots, m.$$

Das bedeutet, dass alle Klassen in die Klassifikationspartition aufgenommen werden, die einen höheren p-Wert haben als das Signifikanzniveau. Der ICP, wie er oben definiert ist, besitzt uneingeschränkte Gültigkeit. Hieraus folgt, dass die im Mittel erzeugten Fehler das Signifikanzniveau δ nicht überschreiten, sofern die Annahmen (unabhängig und identisch verteilt) erfüllt sind [107]. Die Fehler dürfen allerdings ungleich über die Klassen verteilt sein. Um nun auch hinsichtlich der Klassen bedingt gültig zu sein, das bedeutet, dass die bisher lediglich globale Gültigkeit nun auch für jede einzelne Klasse erreicht wird, wurden der Mondrian und weitere bedingte Versionen des ICP entwickelt [104,

107]. Im Bereich der Chemieinformatik sind bereits einige erfolgreiche Anwendungen von bedingten ICPs bekannt [105, 106, 40].

Der ICP errechnet die p-Werte mit Hilfe der Datenpartition *cal*, welche nicht benutzt wurde um die Klassifikationstechnik zu trainieren. Dies ist ein möglicher Nachteil, da *cal* nun nicht länger zum Training zur Verfügung steht, während die Datenpartition *pt* nicht zur Verfügung steht um die p-Werte zu berechnen. Um die vorhandenen Daten effektiver zu nutzen, sind Ansätze mit Kreuzvalidierung entwickelt worden [108]. Allerdings gibt es für keine dieser Verfahren theoretische Ergebnisse zur Gültigkeit. Aus empirischen Studien geht aber hervor, dass diese Eigenschaft häufig erfüllt werden kann [108].

Wenn die Annahmen (unabhängig und identisch verteilt) erfüllt sind, liefern CPs und ICPs gültige Vertrauens-Maße. Häufig ist allerdings neben der Gültigkeit auch die Effektivität für viele Anwendungen entscheidend. Indem ein ungeeignetes Nichtkonformitäts-Maß verwendet wird, steigt der Anteil nicht informativer Klassifikationspartitionen und somit nimmt die Effektivität ab. Folglich ist es von großer Bedeutung ein effektives Nichtkonformitäts-Maß zu verwenden [40]. Darüber hinaus ist es interessant zu wissen wie hoch der Anteil an Molekülen ist, der verloren geht um ein bestimmtes Signifikanzniveau zu halten (Effizienz). Bei einem recht hohen Anteil wäre es für die praktische Anwendung besser auf die Gültigkeit zu verzichten und dafür mehr Moleküle vorhersagen zu können. In der frühen pharmazeutischen Entwicklung würden auf diese Weise mehr potentielle Wirkstoffkandidaten zur Verfügung stehen, was ökonomisch betrachtet von größerem Vorteil sein könnte als die strikte Gültigkeit des Verfahrens.

2 Zielsetzung der Arbeit

2.1 Einleitung

Es gibt eine ganze Reihe von Maßen aus dem Bereich der Neuartigkeits-Detektion und aus dem Bereich der Vertrauens-Schätzung zur Definition eines AB. Diese Maße werden im Bereich der Chemieinformatik auch als „Distanz zum Modell“ (DM) Maße bezeichnet [91]. Bei diesen steht eine kurze Distanz für eine zuverlässige Vorhersage. Obwohl so viele unterschiedliche Maße erfunden und erforscht wurden, fehlte eine groß angelegte Vergleichsstudie. Die bisher einzig verfügbare Vergleichsstudie basiert nur auf einem einzigen Datensatz [91]. Aus dieser Studie ging bereits hervor, dass Methoden, welche nur die Information der unabhängigen Variablen (Neuartigkeits-Detektion) verwenden, weniger leistungsstark sind als diejenigen Methoden, welche zusätzlich die Information der abhängigen Variablen (Vertrauens-Schätzung) nutzen. Allerdings wurden in dieser Studie, mit einer Ausnahme, keine Vertrauens-Maße miteingeschlossen, welche bereits in den betrachteten Klassifikationstechniken miteingebaut sind (Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer).

Diese Lücke wurde geschlossen [109]. In einer groß angelegten Studie wurden unterschiedliche DM-Maße miteinander verglichen, um herauszufinden, welche Maße am besten für eine individuelle Vorhersage die Wahrscheinlichkeit einer Fehlklassifikation charakterisieren. Da es immer ein Zusammenspiel zwischen Klassifikationstechnik und DM-Maß ist und da nicht alle DM-Maße für jede Klassifikationstechnik berechnet werden können, wurde die optimale Kombination (Klassifikationstechnik und DM-Maß) gesucht. Des Weiteren wurden Neuartigkeits-Maße und Vertrauens-Maße zur Definition eines AB miteinander verglichen. Zusätzlich wurden innerhalb der Gruppe der Vertrauens-Maße die bereits eingebauten Vertrauens-Maße der jeweiligen Klassifikationstechniken näher untersucht. Ein Nebeneffekt dieser Studie war, dass die Ergebnisse dieses Vergleichs auch von Interesse sind um einen CP aufzusetzten. Das Herzstück eines CP ist der sogenannte Nichtkonformitäts-Score, welcher auch als DM-Maß gedacht werden kann. Je besser die Nichtkonformitäts-Bewertung die Wahrscheinlichkeit einer Fehlklassifikation einer individuellen Vorhersage charakterisiert, desto effizienter ist der daraus resultierende CP. Folglich würde das beste DM-Maß auch zum besten CP führen. In dieser Studie wurden RF, KNN, SVM, Ensembles von NN, Ensembles von „Boosted Classification

Stumps“ und LDA mit verschiedenen DM-Maßen auf zehn unterschiedlichen Datensätzen miteinander verglichen.

Die Studie kam u.a. zu folgenden Ergebnissen: Wenn ein AB mit dem Ziel der Reduzierung der Fehlerrate definiert werden soll, müssen Vertrauens-Maße verwendet werden. Denn diese identifizieren Moleküle, welche sich nahe der Entscheidungsebene befinden, und weisen die Vorhersage dieser Moleküle zurück. Somit kommt es zu einer Reduzierung der Fehlerrate. Von allen Vertrauens-Maßen zeigen die bereits eingebauten Wahrscheinlichkeits-Schätzer, unabhängig von der Schwierigkeit des zu Grunde liegenden Klassifikationsproblems, die beste Leistung. Die alternativen Vertrauens-Maße sind in keinem Fall besser, in einigen Fällen sogar unterlegen. In dem in dieser Arbeit untersuchten Zweiklassenproblem konnten keine Unterschiede zwischen dem Lernen eines Klassifikationsproblems und dem Lernen eines Regressionsproblems mit dichotomen abhängigen Variablen gefunden werden. Die abschließende stark vereinfachte Empfehlung zur möglichst effizienten Definition eines AB lautet eine leistungsstarke Klassifikationstechnik zu trainieren und den bereits eingebauten Wahrscheinlichkeits-Schätzer zu verwenden. (In dieser Studie wahr der RF führend) [109].

Aus den Resultaten dieser Studie ergaben sich u.a. die Fragestellungen für diese Arbeit. Durch die führende Rolle der miteingebauten Vertrauens-Maße der Klassifikationstechniken, bei welchen es sich um Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer handelt, sollten dessen Eigenschaften näher untersucht werden.

2.2 Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern

Bei den in der Einleitung erwähnten eingebauten Vertrauens-Schätzern von Klassifikationstechniken handelt es sich, wie bereits erwähnt, um Klassenzugehörigkeits-Wahrscheinlichkeits-Schätzer. In vielen Bereichen, wie zum Beispiel bei der Beurteilung der Kreditwürdigkeit von Bankkunden, bei der Analyse vom Kaufverhalten von Supermarktkunden oder eben in der frühen Arzneistoffentwicklung bei der Vorhersage bestimmter Eigenschaften potentieller Arzneistoffkandidaten, werden bzw. können Entscheidungen auf Basis von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern getrof-

fen werden. Um diese treffen zu können, sollten sie möglichst nah an der wahren Wahrscheinlichkeit liegen. Das Ziel dieser Studie ist es, unterschiedliche Regressions- und Klassifikationstechniken, hinsichtlich ihrer Fähigkeit Klassenzugehörigkeits-Wahrscheinlichkeiten möglichst exakt schätzen zu können, zu evaluieren. Eine Auflistung der Techniken, welche untersucht werden sollen, ist in Tabelle 1 zu finden.

Tabelle 1: Übersicht über die zu untersuchenden Klassifikations- und Regressionstechniken.

Klassifikation	**Regression**
RF	RFR
SVM	SVR
KNN	SPLS
PLSDA	Ridge
NBC	Elastic Net
NN	Lasso

Die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen werden sowohl unkalibriert als auch kalibriert, durch Verwendung der logistischen Regression, betrachtet und der Effekt der Kalibrierung wird beurteilt. Darüber hinaus werden unterschiedliche Faktoren, die die Wahrscheinlichkeitsschätzung beeinflussen, untersucht. Zusätzlich zu Simulationsstudien werden auch reale Datensätze verwendet. Des Weiteren wird der Effekt von Hetero-Ensembles auf die Wahrscheinlichkeitsschätzung analysiert und es werden abschließend grobe Regeln aufgestellt, mit welcher Klassifikations- und Regressionstechnik auf welche Weise die besten Wahrscheinlichkeitsschätzungen möglich sind und von welchen (äußeren) Bedingungen diese abhängig sind. Abschließend wird die Nützlichkeit dieser Ergebnisse zur Definition eines AB untersucht.

2.3 Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern versus CP

In dieser Studie sollen die in der vorherigen Studie analysierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer, sofern sie gut kalibriert sind, dafür benutzt werden, den AB eines Klassifikationsmodells zu definieren. Dieser Ansatz soll mit dem CP verglichen

werden, welcher denselben Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer als Nichtkonformitäts-Bewertung verwendet. Hierfür wird die Implementierung des CP aus dem R package „conformal“ verwendet. Die beiden Methoden sollen hinsichtlich ihrer Leistungsstärke (Einhaltung des vorgegebenen Signifikanzlevels) und Effizienz untersucht werden. In dieser Studie wird zunächst nur mit einer Auswahl an realen Datensätzen gearbeitet. Es wird sich auf die Klassifikationstechnik RF und ein DM-Maß beschränkt. Motiviert ist diese Arbeit durch Ergebnisse aus Voruntersuchungen, welche darauf hingedeutet haben, dass der CP zwar sehr leistungsstark aber eventuell weniger effizient ist hinsichtlich der Anzahl an Vorhersagen, die er tätigt. Das bedeutet, dass der CP, um das jeweilig definierte Signifikanzlevel zu halten, relativ viele Moleküle nicht vorhersagt bzw. uninformativ vorhersagt. Gerade in der frühen Wirkstoffentwicklung ist die Effizienz der Methode allerdings sehr wichtig, da sonst womöglich potentielle Wirkstoffkandidaten fälschlicherweise aussortiert werden würden, was keinesfalls erwünscht wäre.

3 Methoden

3.1 Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeits-schätzern

3.1.1 Übersicht über die verwendeten Klassifikations- und Regressionstechniken sowie deren Hyperparametereinstellungen

Die nachfolgende Tabelle 2 listet die verwendeten R-Pakete sowie Hyperparameter auf, welche, abweichend von den Standardparametern, eingestellt wurden. Sofern Abweichungen von dieser Auflistung vorgenommen wurden, wurde dies an gegebener Stelle kenntlich gemacht.

Tabelle 2: Auflistung der verwendeten Pakete sowie relevanter Parameter, welche abweichend von den Standardeinstellungen eingestellt wurden.

Methode	Parameter 1	Parameter 2	Paketname	Version
RF	/	/	randomForest	4.6.10
RFR	/	/	randomForest	4.6.10
SVM	Cost=100	Gamma=1/Anzahl Kolonnen (X)	e1071	1.6.4
SVR	Cost=10	Gamma=1/Anzahl Kolonnen (X)	e1071	1.6.4
KNN	k=15		FNN	1.1
LDA	/	/	MASS	7.3.37
NBC	/	/	e1071	1.6.4
PLSDA	ncomp=30	probMethod=softmax	caret	6.0.41
SPLS	K=1	eta=0.2	spls	2.2.1
Lasso	Lambda.min (10-CV)	alpha=1	glmnet	1.9.8
Ridge	Lambda.min (10-CV)	alpha=0	glmnet	1.9.8
Elastic Net	Lambda.min 10 :V)	alpha.min i.d.R 0.5)	glmnet	1.9.8

3.1.2 Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer

Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit. Der Klassifikationsfehler kann reduziert werden, wenn die Klassifikationstechnik diejenige Klasse ausgibt, der ein neues Molekül mit höchster Wahrscheinlichkeit angehört.

$$\hat{c}(\boldsymbol{x_0}) = \underset{j}{argmax}(\hat{p}(j|\boldsymbol{x_0})) \quad j \in \{1,2\}$$

Hierbei ist $\hat{p}(j|\boldsymbol{x_0})$ definiert als die geschätzte bedingte Wahrscheinlichkeit, dass ein bisher ungesehenes Molekül $\boldsymbol{x_0}$, mit gegebenem molekularem Deskriptor (unabhängigen Variablen), zu einer Klasse j gehört. Wie genau die Klassenzugehörigkeits-Wahrscheinlichkeiten berechnet werden, hängt von der Klassifikationstechnik ab. Einige Klassifikationstechniken, wie beispielsweise die LDA, machen bestimmte Annahmen über die Verteilung der Daten. Hieraus resultiert die Klassenzugehörigkeits-Wahrscheinlichkeit $\hat{p}(\hat{c}|\boldsymbol{x_0})$.

Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit durch Nutzung der lokalen Nachbarschaft. Nicht alle Klassifikationstechniken machen Verteilungsannahmen. Einige von ihnen verwenden die lokale Nachbarschaft eines Moleküls um die Klassenzugehörigkeits-Wahrscheinlichkeit zu berechnen. Zu diesen Techniken zählen der KNN und der RF. Wenn angenommen wird, dass es sich bei $\mathcal{N}_0$ um die Indices der k-nächsten Nachbarn des Trainingsdatensatzes von $\boldsymbol{x_0}$ handelt, dann kann die Wahrscheinlichkeit, dass $\boldsymbol{x_0}$ zur Klasse j gehört, geschätzt werden als der Anteil, den die Moleküle der Klasse j an $\mathcal{N}_0$ insgesamt ausmachen:

$$\hat{p}(j|\boldsymbol{x_0}) = \frac{1}{k}\sum_{i\in\mathcal{N}_0} \mathbb{I}(c_i = j).$$

Bei $\mathbb{I}(g)$ handelt es sich um eine Indikatorfunktion, welche den Wert 1 annimmt, wenn g wahr ist und den Wert 0 wenn g falsch ist. Im Folgenden wird $\hat{p}(\hat{c}|\boldsymbol{x_0})$ vom KNN mit $\hat{p}_{kNN}$ bezeichnet.

Bei einem Entscheidungsbaum kann die Klassenzugehörigkeits-Wahrscheinlichkeit mit wenigen Modifikationen vergleichbar geschätzt werden. In diesem Fall wird angenommen, dass es sich bei $\mathcal{N}_0$ um die Indices der k Moleküle der Trainingsdatenpartition des terminalen Blattes handelt, welchem $\boldsymbol{x_0}$ zugeordnet wurde. Hierbei ist k im Vergleich zum KNN keine feste Zahl, sondern kann variieren. Allerdings wird bei Entscheidungs-

bäumen, genau wie beim KNN, der Anteil, den die Moleküle der Klasse j an $\mathcal{N}_0$ insgesamt ausmachen, als Vertrauensschätzer $\hat{p}(j|\boldsymbol{x_0})$ für die Klassenzugehörigkeits-Wahrscheinlichkeit eines neuen Moleküls $\boldsymbol{x_0}$ genutzt. Da der RF aus einem Ensemble an Entscheidungsbäumen aufgebaut ist, wird $\hat{p}(j|\boldsymbol{x_0})$ über alle Bäume des Ensembles gemittelt:

$$\bar{p}_j(\boldsymbol{x_0}) = \frac{1}{n_{Baum}} \sum_{i=1}^{n_{Baum}} \hat{p}(j|\boldsymbol{x_0}, Baum_i),$$

$Baum_i$ ist in diesem Fall der i-te Klassifikations-Baum des RF, welcher entscheidet, welchem terminalen Blatt $\boldsymbol{x_0}$ zugeordnet wird. Der Mittelwert über alle Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen wird im Englischen im Klassifikationsfall als Prediction Score bezeichnet. Die geschätzte Klasse ist wiederum diejenige, die die größte Klassenzugehörigkeits-Wahrscheinlichkeit aufweist. Dieser Schätzer wird im Folgenden mit $\bar{p}_{RFC}$ abgekürzt. Neben dem gerade beschriebenen Prediction Score gibt es noch einen weiteren, verwandten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer für den RF im Klassifikationsmodus. Wenn ein bisher ungesehenes Molekül $\boldsymbol{x_0}$ vorhergesagt wird, passiert es alle n Bäume des Ensembles und somit werden auch n Klassenvorhersagen erhalten. Ohne Berechnung von Wahrscheinlichkeiten basiert die Klassenzuweisung in diesem Fall auf einer Mehrheitsentscheidung der Mitglieder des Ensembles. Hier soll $v_j(\boldsymbol{x_0})$ die Fraktion der Stimmen sein, die für Klasse j abgegeben wurden:

$$v_j(\boldsymbol{x_0}) = \frac{1}{n_{Baum}} \sum_{i=1}^{n_{Baum}} \hat{p}(j = \hat{c}_i|\boldsymbol{x_0}, Baum_i),$$

es wird diejenige Klasse vorhergesagt $\hat{c}_{RFC}(\boldsymbol{x_0})$, welche die größte Stimmfraktion erhält. Diese Fraktion $v_j(\boldsymbol{x_0})$ für die Klasse $j = \hat{c}_{RFC}(\boldsymbol{x_0})$ kann direkt als Schätzer für die Klassenzugehörigkeits-Wahrscheinlichkeit des Moleküls $\boldsymbol{x_0}$ verwendet werden. Dieses Maß wird auch als engl.: Concordance bezeichnet [91]. Es kann als grobe Version des Prediction Scores verstanden werden, folglich sind keine großen Unterschiede in der Leistung der beiden Schätzer zu erwarten.

Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit im Regressionsfall. Regressionsmethoden minimieren, im Gegensatz zu Klassifikationsmethoden, in der Regel quadratische Verlustfunktionen. Quadratische Verlustfunktionen können allerdings

ebenfalls genutzt werden, um Zwei-Klassen-Klassifikationsprobleme zu lösen. In diesem Fall bekommt der Regressionsalgorithmus anstelle von kontinuierlichen Variablen zwei (dichotome) numerische Klasseninformationen (angenommen $y_i = 1$ steht für Klasse 1 und $y_i = 0$ steht für Klasse 2). Der Regressionsalgorithmus minimiert die quadratische Abweichung der angepassten Werte von diesen Klasseninformationen. Unter diesen Voraussetzungen schätzt diese Regressionsfunktion $\hat{f}(\boldsymbol{x_0})$ Klassenzugehörigkeits-Wahrscheinlichkeiten [15]:

$$\hat{y}(\boldsymbol{x_0}) = \hat{f}(\boldsymbol{x_0}) = E(1|\boldsymbol{x_0}) = p = (1|\boldsymbol{x_0}).$$

Die Klassenzuweisung erfolgt nun auf Basis folgender Regel:

$$\hat{c}(\boldsymbol{x_0}) = \begin{matrix}1\ if\ \hat{y}(\boldsymbol{x_0}) > 0.5\\ 2\ if\ \hat{y}(\boldsymbol{x_0}) \leq 0.5\end{matrix}$$

In der Praxis können allerdings Probleme auftreten, wenn $\hat{y}(\boldsymbol{x_0})$ nicht zwischen 0 und 1 begrenzt sein muss, wie beispielsweise bei der Multiplen Linearen Regression. Für die Anwendung auf reale Probleme ist es von großer Bedeutung, dass die Regressionsfunktion bedingte Erwartungswerte $E(1|\boldsymbol{x_0})$ gut approximieren kann. Je besser diese Funktion das kann, desto nützlicher sind deren Schätzwerte für die Klassenzugehörigkeits-Wahrscheinlichkeit. Bei vielen nichtparametrischen Regressionstechniken schätzt $\hat{y}(\boldsymbol{x_0})$ $p = (1|\boldsymbol{x_0})$ konsistent [110]. Wenn die Stichprobengröße gegen unendlich strebt, schätzen diese Regressionstechniken, wie zum Beispiel, die u.a. in dieser Arbeit verwendeten Techniken, KNN Regression [111], Neuronale Netze bei denen mit der Back-Propagation Regel die Summe quadrierter Fehler minimiert wurde (NN) oder Random Forests (RFR) im Regressionsmodus [110], Klassenzugehörigkeits-Wahrscheinlichkeiten asymptotisch korrekt. Allerdings muss an dieser Stelle bedacht werden, dass Konsistenz alleine nichts über die Eigenschaften eines bestimmten Schätzers aussagt, wenn die Stichproben klein sind [112]. Als Schätzer können folglich die Fehlerwahrscheinlichkeiten $1 - \hat{p}(1|\boldsymbol{x_0}) = 1 - \hat{y}(\boldsymbol{x_0})$ für Moleküle verwendet werden, welche der Klasse 1 zugeordnet werden und $1 - \hat{p}(2|\boldsymbol{x_0}) = \hat{p}(1|\boldsymbol{x_0}) = \hat{y}(\boldsymbol{x_0})$ für Moleküle, welche der Klasse 2 zugeordnet werden. Es werden also immer die kleineren Fehlerwahrscheinlichkeiten verwendet.

Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit bei SVM. SVM klassifizieren neue Moleküle, je nachdem auf welcher Seite der Entscheidungsebene sie sich befinden. Diese Information geht aus dem Vorzeichen des Entscheidungs-Wertes hervor. Die

Höhe des Entscheidungs-Wertes ist abhängig von der Distanz des Moleküls zur Entscheidungsebene und wird ausgedrückt als Vielfaches der Margin [113]. Es gilt, je kürzer die Distanz zur Entscheidungsebene ist, desto unsicherer ist die Vorhersage. Diese Distanz ist im engeren Sinn keine Klassenzugehörigkeits-Wahrscheinlichkeit, aber durch Kalibrierung können daraus Wahrscheinlichkeitsschätzer für die Klassenzugehörigkeit erhalten werden $\hat{p}_{SVC}$. Hierfür wird die sogenannte Platt Skalierung verwendet [96]

Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit bei Neuronalen Netzen im Klassifikationsmodus. NN im Klassifikationsmodus (NNC) haben hier zwei Output-Nodes. Die Aktivierungsfunktion und die Ausgabefunktion (softmax) stellen sicher, dass Klassenzugehörigkeits-Wahrscheinlichkeiten geschätzt werden, wobei der größte ausgegebene Wert die vorhergesagte Klasse bestimmt. Da in dieser Studie ein Ensemble mit fünf Mitgliedern verwendet wurde, wird der gemittelte größte Ausgabewert verwendet.

3.1.3 Auswertung von Zuverlässigkeits-Diagrammen

Wie bereits in der Einleitung beschrieben, gibt es für die Beurteilung der Genauigkeit von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern keine Standardmethode. Ein Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer ist gut bzw. genau, wenn die geschätzte Wahrscheinlichkeit nah an der wahren Wahrscheinlichkeit liegt. In der Regel sind die wahren Wahrscheinlichkeiten nicht bekannt, sondern lediglich die Klasseninformation. Durch die Darstellung der Wahrscheinlichkeitsschätzungen sowie der Klasseninformationen in einem Zuverlässigkeitsdiagramm, entspricht die wahre Wahrscheinlichkeit der Winkelhalbierenden. Bei Betrachtung des Diagramms kann somit bereits abgeschätzt werden, ob die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen gut kalibriert sind oder nicht.

3.1.4 Bewertung der Exaktheit von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern

Da es, wie im oberen Abschnitt bereits beschrieben, keine Standardmethode für die Beurteilung der Genauigkeit von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern gibt, werden an dieser Stelle nun die in der Literatur gängigen sowie die in dieser Arbeit verwendeten Methoden kurz vorgestellt.

Die erste Methode hat keinen Eigennamen, daher wird sie im Folgendem **Quadratischer Fehler (QF)** genannt. Dieser wird für ein beliebiges Objekt $\boldsymbol{x}$ wie folgt berechnet:

$$(t(j|\boldsymbol{x}) - p(j|\boldsymbol{x}))^2.$$

Bei $p(j|\boldsymbol{x})$ handelt es sich um die vorhergesagte Klassenzugehörigkeits-Wahrscheinlichkeit für eine bestimmte Klasse *j* und bei $(t(j|\boldsymbol{x})$ um die wahre Klassenzugehörigkeits-Wahrscheinlichkeit. Für die Testdatenpartition sind die wahren Klassenzugehörigkeits-Wahrscheinlichkeiten für eine bestimmte Klasse *j* unbekannt. Lediglich die Klassen sind bekannt. Aus diesem Grund wird $(t(j|\boldsymbol{x})$ als 1 definiert, wenn die Klasse des Objektes gleich *j* ist und 0, wenn dies nicht der Fall ist [97].

Im Zweiklassenfall wird der QF als **Brier-Score** [114] bezeichnet. Nachdem der *QF* für jedes Testobjekt berechnet wurde, wird der **MSE** für die gesamten Datensatzpartitionen berechnet. In dieser Arbeit wird nur mit Zweiklassenfällen gearbeitet, daher wird im Folgenden nur der Begriff Brier-Score verwendet. [97, 98, 114]. Ursprünglich war der Brier-Score keine Methode um die Effizienz einer Kalibrierung zu beurteilen, sondern lediglich die Nicht-Richtigkeit. Allerdings konnte gezeigt werden, dass sich diese Nicht-Richtigkeit zerlegen lässt in einen Teil, welcher misst, wie nah die vorhergesagten Wahrscheinlichkeiten an den wahren Wahrscheinlichkeiten sind (engl.: Reliability Component) und einen Teil, welcher die Fähigkeit des Klassifikators misst den Objekten Klassen zuzuordnen, sodass der Anteil an TPs bezogen auf eine Klasse maximal divers ist (engl.: Resolution Component). Hieraus geht hervor, dass der Brier-Score minimal wird, wenn die wahren Wahrscheinlichkeiten als Vorhersagen verwendet werden [115]. Allerdings muss an dieser Stelle beachtet werden, dass auf diese Weise nur Vergleiche der Leistungsfähigkeit der Kalibrierung möglich sind, wenn die Leistung der Klassifikationstechnik konstant ist. Darüber hinaus gibt es noch einige weitere Zerlegungen des Brier-Score (siehe [116] für weiterführende Informationen).

Sowohl Fawcett und Niculescu-Mizil, als auch Flach und Takashi Matsubara erkannten 2007, unabhängig voneinander, die Beziehung zwischen den AUC-basierten Maßen der ROC-Analyse und kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern. Sie entdeckten, dass ein perfekt kalibrierter Klassifikator immer eine konvexe ROC Kurve ergibt. Flach und Takashi Matsubara zerlegten in ihrer Publikation den Brier-Score in einen sogenannten engl.: Calibration Loss und einen engl.: Refinement Loss. Der Calibra-

tion Loss, welcher den Kalibriererfolg messen soll, ist definiert als die mittlere quadratische Abweichung der empirischen Wahrscheinlichkeiten von der Steigung der Segmente der ROC Kurve (Anzahl Segmente entspricht Anzahl unterschiedlicher geschätzter Klassenzugehörigkeits-Wahrscheinlichkeiten) [94].

Des Weiteren gibt es noch eine Methode, welche hauptsächlich verwendet wird, wenn Kalibrierung wichtig ist. Sie wird als engl.: **Cross Entropy** [117] oder auch als engl.: LogLoss bezeichnet [94, 118, 119].

Nach Zadrozny und Elkan sind die Methoden Cross Entropy sowie die QF-Methode und der Brier-Score alle geeignet um Klassenzugehörigkeits-Wahrscheinlichkeiten bezüglich ihrer Nähe zur wahren Wahrscheinlichkeit zu evaluieren und somit gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeiten von schlecht kalibrierten zu unterscheiden.

Verfahren auf Basis von sogenannten **Lift Graphiken** [120] und **ROC Kurven** [121], sollten nicht zur Bewertung der Exaktheit von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen benutzt werden. Der Grund dafür ist, dass sie lediglich die Exaktheit der Rangreihenfolge der Testobjekte beurteilen können [122, 97].

Sehr einfach zu berechnende Methoden basieren auf den bereits erläuterten Zuverlässigkeits-Diagrammen. Die Abweichung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer-Werte zur wahren Wahrscheinlichkeit kann berechnet werden, indem der Mittlere Absolute Fehler **MAE** (engl. Mean Absolute Error) oder auch der Mittlere Quadratische Fehler **MSE** (engl. Mean squared Error) der zehn Punkte der Kurve zur Winkelhalbierenden berechnet wird. Der MSE bestraft größere Abweichungen stärker als der MAE. Je kleiner der MAE oder MSE, desto näher sind die Wahrscheinlichkeitsschätzungen an der wahren Wahrscheinlichkeit. Da weder in den MAE noch in den MSE die Besetzungsstatistik der Segmente eingeht, kann zusätzlich der **MAE_w** bzw. **MSE_w** berechnet werden, bei welchem die jeweiligen Segmentmittelwerte noch einmal durch die Anzahl der Objekte geteilt werden. Neben diesen Varianten wurde von Caruana und Niculescu-Mizil noch eine weitere Variante vorgestellt, welche anstelle der zehn Mittelwerte der Segmente, immer 100 Objekte in ein Segment packt und dann von (S_x) {S_1=1-100, S_2=2-101,S_3=3-102 etc. } die Fehler berechnet [123, 102]. Diese Variante wird im Folgenden als **MAE_Cu** bzw. **MSE_Cu** bezeichnet.

In dieser Arbeit wurden zur Evaluation der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der MAE, MSE, MAE_w, MSE_w, MAE_Cu und MSE_Cu verwendet. Diese berechnen ausschließlich den Fehler zur wahren Wahrscheinlichkeit und ergänzen die visuelle Inspektion der Zuverlässigkeitsdiagramme. Zusätzlich wird auch der Brier-Score berechnet, da dieser in der Literatur die gängigste Methode ist [117].

3.1.5 Modellvalidierung

Das Ziel dieser Arbeit ist es zu beurteilen, wie die unterschiedlichen Techniken für bisher ungesehene Moleküle die Klassenzugehörigkeits-Wahrscheinlichkeiten schätzen können, von welchen Faktoren dies abhängig ist und ob die Kalibrierung mit der logistischen Regression zu einer Verbesserung führt. Da in dieser Studie keine zukünftigen Moleküle für die verwendeten Datensätze vorhanden sind, werden die Klassifikations- und Regressionsmodelle mit Hilfe von zufällig ausgelassenen Datensatzpartitionen überprüft. Hier wird, um die Daten möglichst effizient zu nutzen, eine 50*50 % „Lass-mehrere-Objekte-heraus-Kreuzvalidierung“ (LMO-CV) durchgeführt, welche zukünftige Daten simulieren soll. Hier soll darauf hingewiesen werden, dass keine Optimierung von Hyperparametern durchgeführt wurde. Folglich wurde auch keine Modellselektion durchgeführt, wodurch ein systematischer Fehler durch Modellselektion (engl.: Model Selection Bias) entstehen kann [124, 39]. Aufgrund der Tatsache, dass bei der 50% LMO-CV die Daten jeweils in eine Trainingspartition (50%) und eine unabhängige Testpartition (50%) unterteilt werden, werden die Vorhersagefehler für die jeweilige Trainingspartition unverzerrt geschätzt. Die Hyperparameter zeigen im Durchschnitt eine gute Leistung. Dadurch, dass diese nicht optimiert wurden, kann dies bei einigen Datensätzen zu suboptimalen Modellen führen, allerdings sind keine besonders großen Unterschiede zu erwarten. Außerdem soll der Fokus der Studie auf der Charakterisierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer liegen und es ist nicht zu erwarten, dass diese Unterschiede die Charakterisierung beeinflussen. Für die Bewertung der Leistung der Klassifikationstechniken wurde die Korrektklassifizierungsrate (Acc) berechnet.

3.1.6 Datensätze und molekulare Deskriptoren

Alle in dieser Arbeit verwendeten Datensätze sind frei verfügbar. Tabelle 3 fasst die Charakteristiken dieser Datensätze zusammen. Es kann beobachtet werden, dass sich die verschiedenen Datensätze in ihrer Größe und in ihren Klassenverhältnissen unter-

scheiden. Detailliertere Informationen sind zusätzlich in den dazu gehörigen Referenzen zu finden. Für die folgenden vier Datensätze wurden die publizierten Deskriptoren verwendet: MUSK2 [125, 126], QSAR [127, 128], BBB [129],PGP [65]. Für die übrigen 6 Datensätze wurden die zur Verfügung gestellten SMILES verwendet um zwei unterschiedliche molekulare Deskriptoren zu berechnen. Auf der einen Seite wurden die engl.: MACCS Keys (166 bit) [130] berechnet, welche die Besetzungsstatistik von Substrukturen aufzeichnen und auf der anderen Seite die MOE Deskriptoren. Hierbei wurden die rotations- und translationsinvarianten Deskriptoren benutzt. Zur Berechnung der Deskriptoren wurde die Chemical Computing Group's Molecular Operating Environment (MOE) software (Release 2013.08) verwendet [131]. Eine Liste der MOE Deskriptoren ist im Anhang (Kapitel 9.1.7) zu finden. Alle Deskriptoren mit Ausnahme der MACCS Keys und der engl.: Binarized Atom Pairs wurden autoskaliert.

Tabelle 3: Eigenschaften der in der Arbeit verwendeten Datensätze.

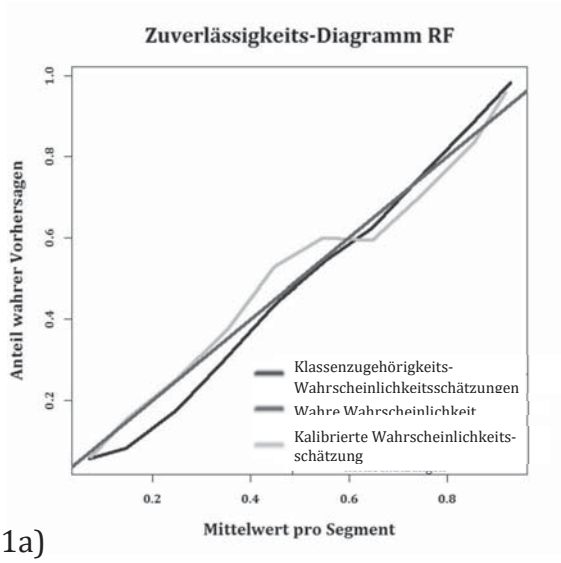
Datensatz	**Deskriptortyp**	**Anzahl Moleküle**	**Anzahl Deskriptoren**	**Klassen-verhältnis**
MUSK2	Shape/Conformation	6598	166	85/15
QSAR	DRAGON	1055	41	34/66
BBB	Chemical/Physical Properties	325	9	45/55
CYP1A2	MACCS/MOE/ E-State	7485	166/181/192	46/54
PGP	Binarized Atom Pairs	186	1522	42/58
FactorXa	MACCS/MOE	435	166/181	36/64
Liver	MACCS	951	166/181	32/68
hERG	MACCS	561	166/181	62/38
Cancer	MACCS	7747	166/181	41/59
Ames	MACCS	6512	166	46/54

3.1.7 Simulationsaufbau

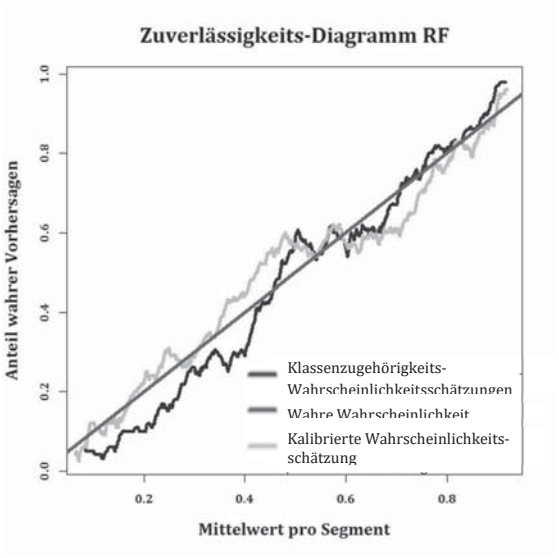
Für die unterschiedlichen Simulationsstudien wird der jeweilige Versuchsaufbau vor der Beschreibung der Ergebnisse kurz erläutert.

3.1.8 Festlegung einer Zuverlässigkeitsgrenze

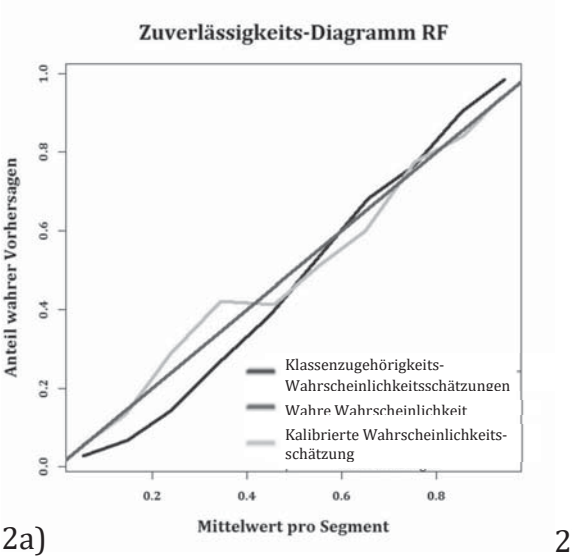
Wenn der MAE/ MSE bzw. die verschieden berechneten MAE/ MSE-Werte unterschiedlicher Klassifikations- und Regressionstechniken (kalibriert und unakalibriert etc) miteinander verglichen werden sollen, konnte bisher lediglich die Aussage getroffen werden, dass die Technik/ Methode mit dem geringeren MAE/ MSE exaktere Klassenzugehörigkeits-Wahrscheinlichkeiten schätzen kann. Darüber hinaus ist es allerdings ebenso interessant zu wissen, welche Methoden insgesamt gut kalibrierte Wahrscheinlichkeitsschätzer hervorbringen. Die MAE/ MSE-Werte, sowie die gewichtete Variante (MAE_w /MSE_w) und die Variante mit dem gleitenden Mittelwert (MAE_Cu/ MSE_Cu) werden, wenn es sich um die Fehler nach der Kalibrierung mittels logistischer Regression handelt (Platt-Scaling), mit dem Kürzel _Log versehen. Zu diesem Zweck sollte ein MAE/ MSE Grenzwert festgelegt werden, bis zu welchem davon ausgegangen werden kann, dass die betrachtete Technik gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorbringt. Zur Festlegung dieses Grenzwertes wurden Zuverlässigkeitsdiagramme visuell inspiziert und der entsprechende MAE/ MSE wurde berechnet. Sobald die Wahrscheinlichkeitsschätzkurve nicht mehr nah der Winkelhalbierenden verläuft, wurde angenommen, dass die Wahrscheinlichkeitsschätzung zu ungenau ist (siehe Abbildung 11). Neben dem klassischen Zuverlässigkeits-Diagramm (Typ a) wurde noch die Variante abgebildet, welche auf der Fehlerberechnung von Caruana (auf Basis des gleitenden Mittelwertes) basiert (Typ b). Die dazugehörigen Werte der Fehler sind in Abbildung 12 und Abbildung 13 zu finden.



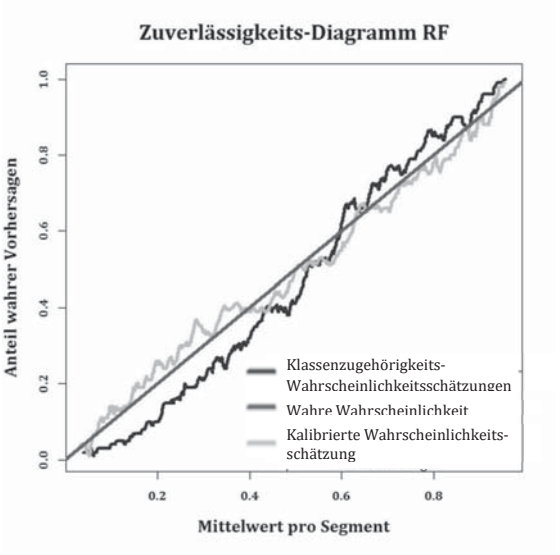
1a)



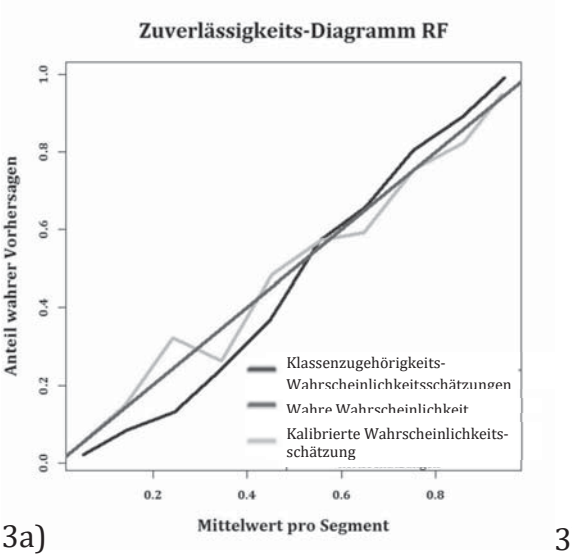
1b)



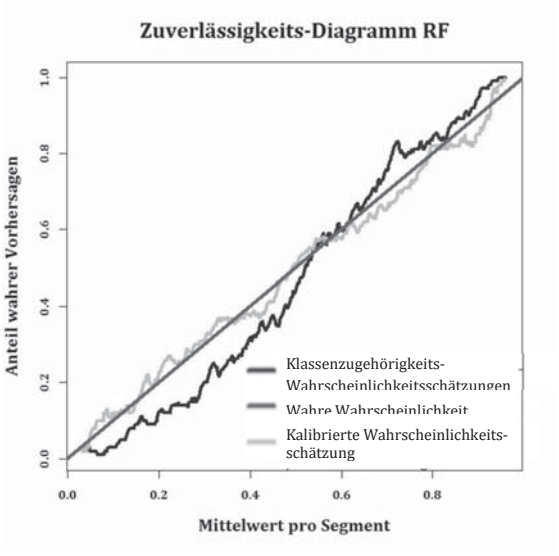
2a)



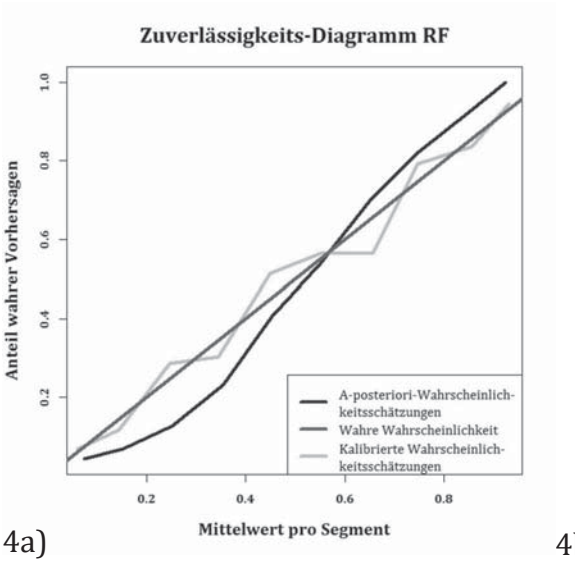
2b)



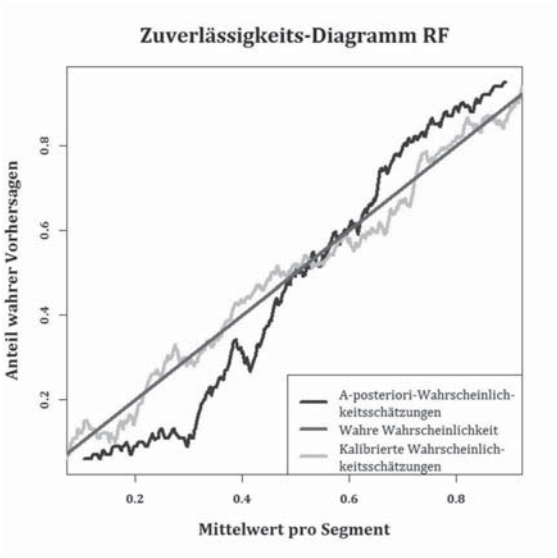
3a)



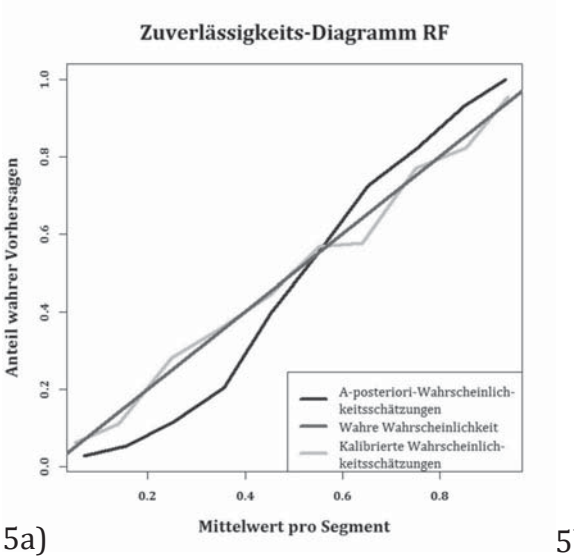
3b)



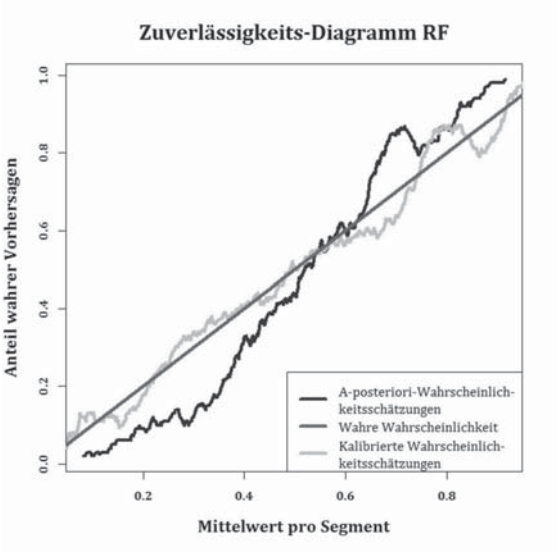
4a)



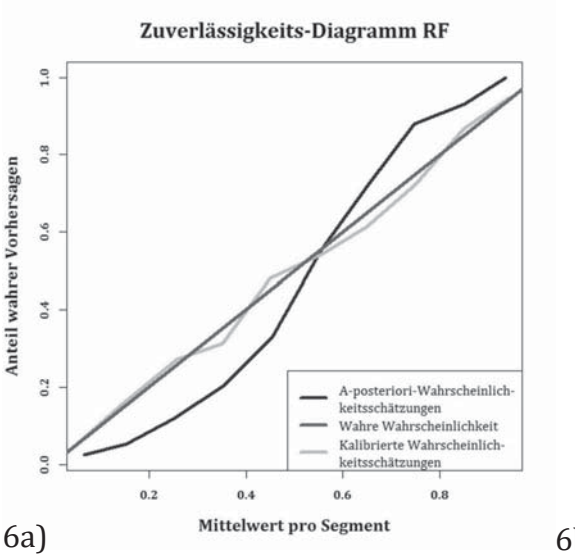
4b)



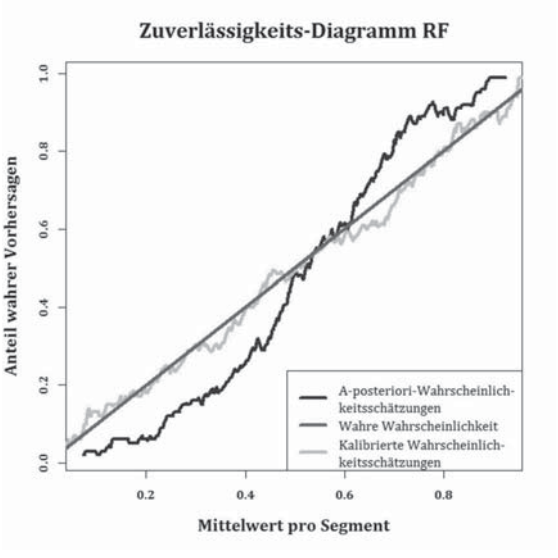
5a)



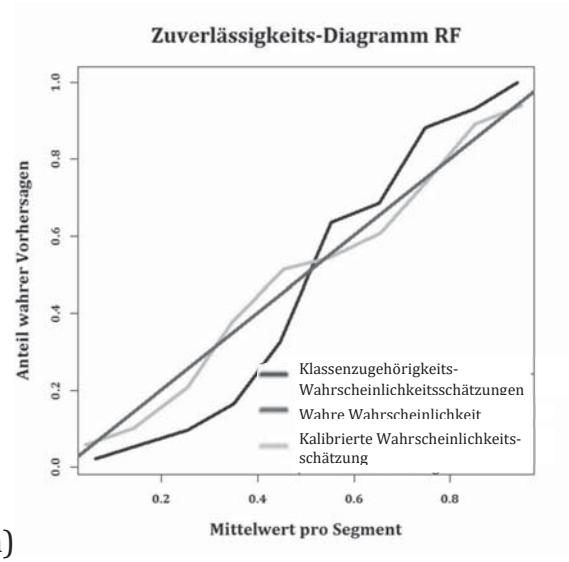
5b)



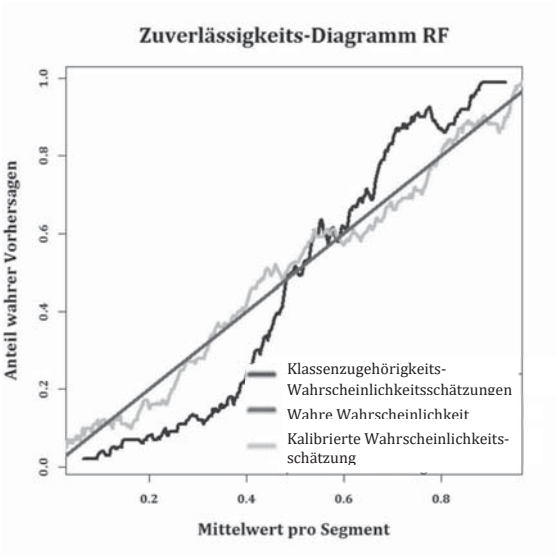
6a)



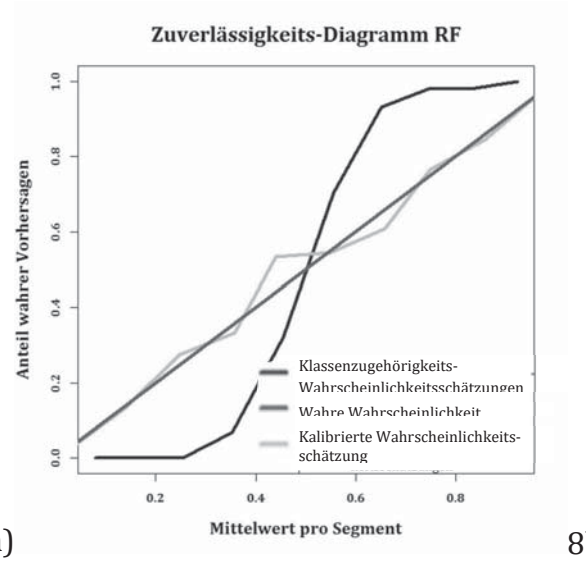
6b)



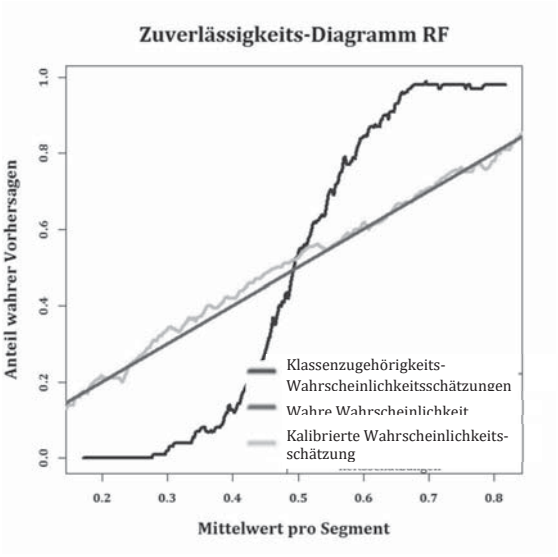
7a)



7b)



8a)



8b)

Abbildung 11: Zuverlässigkeits-Diagramme mit aufsteigendem MAE/ MSE (basierend auf blauer Kurve berechnet). Die cyane Kurve zeigt die kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer (nach logistischer Regression), auf dessen Basis der kalibrierte Fehler berechnet wurde. Es handelt sich um simulierte Daten. Ein Datensatz wurde mit 2000 Objekten und 40 Variablen generiert, welcher sich aus zwei Klassen zusammensetzt. Beide Klassen wurden jeweils aus multivariaten Normalverteilungen ($\boldsymbol{N}$ ($\boldsymbol{\mu}$, $\boldsymbol{I}$)) generiert, das bedeutet, Normalverteilungen mit dem Mittelwert $\boldsymbol{\mu}$ und Covarianzmatrix $\boldsymbol{I}$ (unkorreliert), wobei sich die Mittelwerte unterscheiden. Um eine gewisse Klassifikationsleistung einzustellen, wurden die Mittelwerte der beiden Klassen zusammen oder auseinander geschoben. Je weiter diese Mittelwerte auseinander liegen, desto leichter ist das Klassifikationsproblem. Auf diese Weise konnten unterschiedliche MAE/ MSE-Werte simuliert werden. Bei den mit a) gekennzeichneten Diagrammen handelt es sich um die „klassische" Variante und bei den mit b) gekennzeichneten Diagrammen um die Variante nach Caruana (auf Basis des gleitenden Mittelwertes).

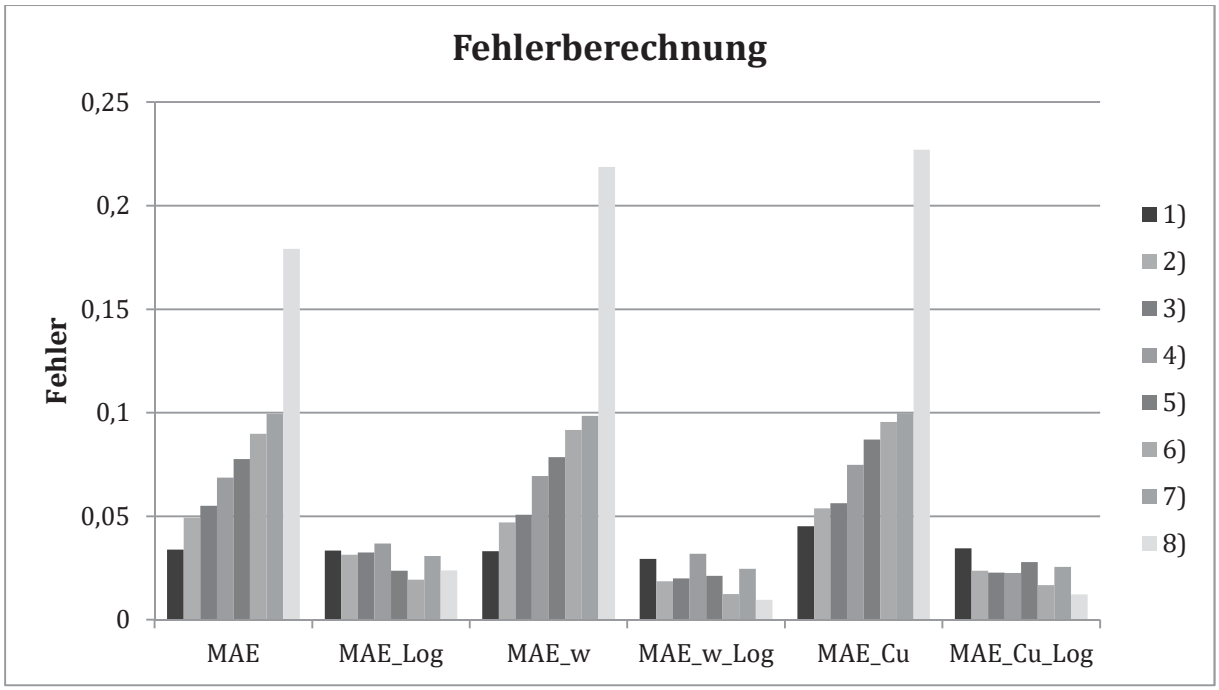


Abbildung 12: Zur Abbildung 11 gehörigen absoluten Fehlervarianten der Graphiken 1-8. Das Kürzel _Cu kennzeichnet die Fehlervarianten des Abbildungtypes b), die übrigen gehören zum Abbildungstyp a). Das Kürzel _Log kennzeichnet den Fehler nach logistischer Regression (Kalibrierung).

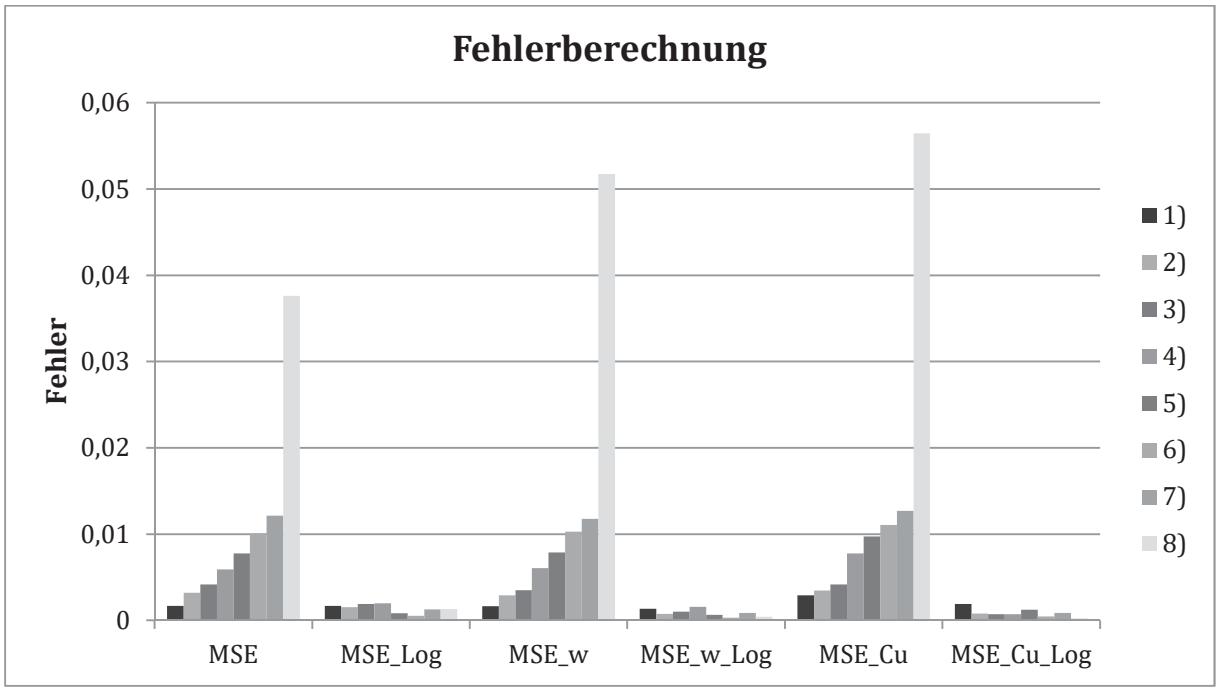


Abbildung 13: Zur Abbildung 11 gehörigen quadrierten Fehlervarianten der Graphiken 1-8. Das Kürzel _Cu kennzeichnet die Fehlervarianten des Abbildungtypes b), die übrigen gehören zum Abbildungstyp a). Das Kürzel _Log kennzeichnet den Fehler nach logistischer Regression (Kalibrierung).

Bei visueller Inspektion der Zuverlässigkeits-Diagramme ist zu erkennen, dass die unkalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen der Diagramme 1-3 sehr nah an der wahren Wahrscheinlichkeit liegen und größtenteils sogar auf der entsprechenden Geraden liegen. In den Diagrammen 4 bis 7 entfernen sich die unkalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen an den Rändern kontinuierlich von der wahren Wahrscheinlichkeit. Schließlich ist der Abstand in Diagramm 8 an den Rändern enorm groß. Die kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen liegen alle sehr nah an der wahren Wahrscheinlichkeit. Dies gilt sowohl für den Diagramm Typ a), als auch für den Diagramm Typ b). Durch Verwendung des gleitenden Mittelwertes nach Caruana ist der Diagramm Typ b) nicht so stark geglättet und erscheint somit etwas „unruhiger“. Der Diagramm Typ b) hat gegenüber dem Diagramm Typ a) einen entscheidenden Vorteil, er kann keine leeren Segmente haben. Dies kann allerdings bei Typ a) durch eine extreme Verteilung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der Fall sein und dann ist die Berechnung des Fehlers praktisch nicht möglich.

Basierend auf der visuellen Inspektion kann sowohl für die unkalibrierten (MAE), als auch für die kalibrierten Wahrscheinlichkeitsschätzungen (MAE_Log), eine grobe Fehlergrenze bei 0.05 gesetzt werden, bis zu dieser angenommen wird, dass die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen noch relativ nah an der wahren Wahrscheinlichkeit liegen und somit folglich noch als gut kalibriert angesehen werden können. Für die gewichteten MAE Werte (MAE_w, MAE-Log_w), welche zusätzlich noch die Besetzungsstärke des Segments berücksichtigen, kann dieselbe Fehlergrenze festgelegt werden. Dasselbe gilt somit auch für das Fehlermaß von Caruana (MAE_Cu und MAE_Cu_Log), da durch die gleitende Mittelwertbildung ebenfalls die Besetzungsstärke des Segments berücksichtigt wird. Die resultierenden Fehlermaße, welche jeweils den quadratischen Fehler berechnen, liefern Werte in einer kleineren Größenordnung. Bei diesen Werten könnte ein grober Grenzwert bei 0.005 festgelegt werden. Im Folgenden werden diese unterschiedlichen Arten der Fehlerberechnung noch einmal näher betrachtet. Allgemein lässt sich sagen: je niedriger der MAE/ MSE, desto besser sind die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen kalibriert.

3.2 Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeits-schätzern versus CP

3.2.1 Übersicht über die verwendeten Klassifikationstechniken sowie deren Hyperparametereinstellungen

Die nachfolgende Tabelle 4 listet die verwendeten R-Pakete sowie Hyperparameter auf, welche abweichend von den Standardparametern eingestellt wurden. Sofern Abweichungen von dieser Auflistung vorgenommen wurden, wurde dies an gegebener Stelle kenntlich gemacht.

Tabelle 4: Auflistung der verwendeten Pakete sowie relevanter Hyperparameter, welche abweichend von den Standardeinstellungen eingestellt wurden.

Methode	Parameter 1	Parameter 2	Paketname	Version
CP	nb_trees=500	/	conformal	0.1

3.2.2 Modellvalidierung

Das Ziel dieser Studie ist es zu beurteilen, wie effizient der Conformal Predictor bei gegebenem Signifikanzlevel einen AB definieren kann. An dieser Stelle meint effizient wie viele bisher ungesehene Moleküle als uninformativ vorhergesagt werden. Außerdem soll getestet werden, ob Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer ebenfalls verwendet werden können um einen AB zu definieren, indem das vorgegebene Signifikanzlevel als Grenzwert dient. Moleküle außerhalb dieses Grenzwertes werden zurückgewiesen und werden somit ebenfalls uninformativ vorhergesagt. Bereits in früheren Publikationen wurden Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer dazu verwendet, Objekte nahe der Entscheidungsebene zurück zu weisen und somit die Fehlerrate zu senken, dies wurde als „Reject Option“ bezeichnet [132]. Anschließend soll verglichen werden, welche der beiden Methoden weniger Moleküle uninformativ vorhersagt und somit effizienter ist. Genauer betrachtet wird, um die Daten möglichst effizient zu nutzen, eine 5-fache-Kreuzvalidierung (CV) durchgeführt, welche zukünftige Daten simulieren soll. An dieser Stelle soll noch einmal darauf hingewiesen werden, dass keine Optimierung von Hyperparametern durchgeführt wurde. Für die Bewertung der Leistung der Klassifikationstechniken wurde die Korrektklassifizierungsrate (Acc) berechnet.

3.2.3 Datensätze und molekulare Deskriptoren

Es wurden exemplarisch drei frei verfügbare Datensätze, welche bereits in der letzten Studie (siehe Kapitel 3.1.3) Anwendung gefunden haben, verwendet. Mit der einzigen Ausnahme, dass weniger Deskriptorsätze verwendet wurden (siehe Tabelle 5). Die Datenvorbehandlung war ebenfalls identisch. Es kann beobachtet werden, dass sich die verschiedenen Datensätze in ihrer Größe und in ihren Klassenverhältnissen unterscheiden. Detailliertere Informationen sind zusätzlich in den dazu gehörigen Referenzen zu finden. Für den folgenden Datensatz wurden die publizierten Deskriptoren verwendet: MUSK2 [125, 126] Für die übrigen 2 Datensätze wurden die zur Verfügung gestellten SMILES verwendet um zwei unterschiedliche molekulare Deskriptoren zu berechnen. Auf der einen Seite wurden die engl.: MACCS Keys (166 bit) [130] berechnet, welche die Besetzungsstatistik von Substrukturen aufzeichnen und auf der anderen Seite die MOE Deskriptoren, hierbei wurden die rotations- und translationsinvarianten Deskriptoren benutzt. Zur Berechnung der Deskriptoren wurde die Chemical Computing Group's Molecular Operating Environment (MOE) software (Release 2013.08) verwendet [131]. Eine Liste der MOE Deskriptoren ist im Anhang (Kapitel 9.1.7) zu finden. Alle Deskriptoren mit Ausnahme der MACCS Keys wurden autoskaliert.

Tabelle 5: Eigenschaften der in der Arbeit verwendeten Datensätze.

Datensatz	Deskriptortyp	Anzahl Moleküle	Anzahl Deskriptoren	Klassen-verhältnis
MUSK2	Shape/Conformation	6598	166	85/15
FactorXa	MACCS/MOE	435	166/181	36/64
Ames	MACCS	6512	166	46/54

3.2.4 Einhaltung des Signifikanzlevels mit dem R package „conformal"

Zuerst wird ein RF auf einer Trainingsdatenpartition unter Verwendung einer 5-fachen CV trainiert. Die Nichtkonformitäts-Scores werden klassenweise berechnet. Dieses Verfahren wird auch als klassenweise Mondrian off-line inductive conformal prediction

(MICP) bezeichnet [133, 134]. Im Detail wird die Anzahl an Bäumen, welche für eine bestimmte Klasse votiert hat, durch die gesamte Anzahl an Bäumen geteilt. Dieses Nichtkonformitäts-Measure wird auch als Fraction (engl.: Anteil) bezeichnet. Wenn beispielsweise ein binäres Klassifikationsproblem vorliegt und 87 von 100 Bäumen für ein bestimmtes Objekt für Klasse 0 votieren, dann würde die Nichtkonformitäts-Bewertungszahl 0.87 für Klasse 0 und 0.13 für Klasse 1 betragen. Auf diese Weise entsteht eine Nichtkonformitäts-Bewertungszahl-Matrix, dessen Reihen zu den Objekten der Trainingsdatenpartition gehören und dessen Spalten zu den Klassen (0 und 1). Anschließend wird jede Spalte aufsteigend sortiert. Die Spalten werden nun als Mondrian Klassenlisten bezeichnet. An dieser Stelle darf nicht vergessen werden ein Signifikanzlevel $(1 - \delta)$ festzulegen [134].

Das Modell, welches auf der Trainingsdatenpartition trainiert wurde, wird anschließend verwendet, um die Objekte in der Testdatenpartition vorherzusagen. Im Detail wird für ein bisher ungesehenes Objekt $\boldsymbol{x_0}$ zunächst die Nichtkonformitäts-Bewertungszahl (p) berechnet:

$$p(\boldsymbol{x_0}; 0) = \frac{N_{Bäume\ für\ Klasse\ 0}}{N_{Bäume}}$$

$$p(\boldsymbol{x_0}; 1) = \frac{N_{Bäume\ für\ Klasse\ 1}}{N_{Bäume}}$$

Danach werden die p-Werte für jede Klasse berechnet. Hierbei wird die Anzahl der Objekte in der entsprechenden Mondrianliste, welche einen kleineren p-Wert als das betrachtete Objekt aufweisen, durch die Gesamtanzahl der Objekte der Trainingsdatenpartition, geteilt.

$$p - Wert\ (\boldsymbol{x_0}; 0) = \frac{|\{MCL(0) < p(\boldsymbol{x_0}; 0)\}|}{N_{Trainingsdaten}}$$

$$p - Wert\ (\boldsymbol{x_0}; 1) = \frac{|\{MCL(1) < p(\boldsymbol{x_0}; 1)\}|}{N_{Trainingsdaten}}$$

Schließlich werden diese p-Werte mit dem zuvor definierten Signifikanzlevel verglichen. Damit ein betrachtetes Objekt einer bestimmten Klasse zugeordnet wird, muss der p-Wert größer sein als das Signifikanzlevel. Wenn die p-Werte für beide Klassen das Signifikanzlevel überschreiten, dann würde das Objekt beiden Klassen zugeordnet werden

und im umgekehrten Fall folglich keiner Klasse [134]. In beiden Fällen wäre die Vorhersage uninformativ.

3.2.5 Einhaltung des Signifikanzlevels mit Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern

Zuerst werden mit dem RF auf Basis der Trainingsdatenpartition die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der Testdatenpartition bestimmt. In diesem Fall wurde, aus Gründen der Vergleichbarkeit, ebenfalls das Nichtkonformitäts-Measure Fraction berechnet. Danach werden die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer in aufsteigender Reihenfolge aus Sicht einer Klasse sortiert. Aufgrund der Sortierung der Wahrscheinlichkeitsschätzer aus Sicht von nur einer Klasse, befinden sich somit alle unsicher vorhergesagten Objekte in der Mitte und alle sicher vorhergesagten Objekte an den Rändern. Anschließend wird das Signifikanzlevel festgelegt. Das Signifikanzlevel nimmt hier einen Wert zwischen 0.1 und 0.5 an (die fünf möglichen Werte sind: 0.1; 0.2; 0.3; 0.4; 0.5). Ein Signifikanzlevel von 0.1 beinhaltet alle Objekte, die mit höchstens 10%iger Wahrscheinlichkeit falsch zugeordnet wurden bzw. mindestens mit 90%iger Wahrscheinlichkeit richtig zugeordnet wurden (siehe Abbildung 14). An dieser Stelle wird davon ausgegangen, dass die zuvor berechneten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer nah an der wahren Wahrscheinlichkeit liegen. Somit können sichere Objekte von unsicheren, basierend auf ihrem Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwert, unterschieden werden.

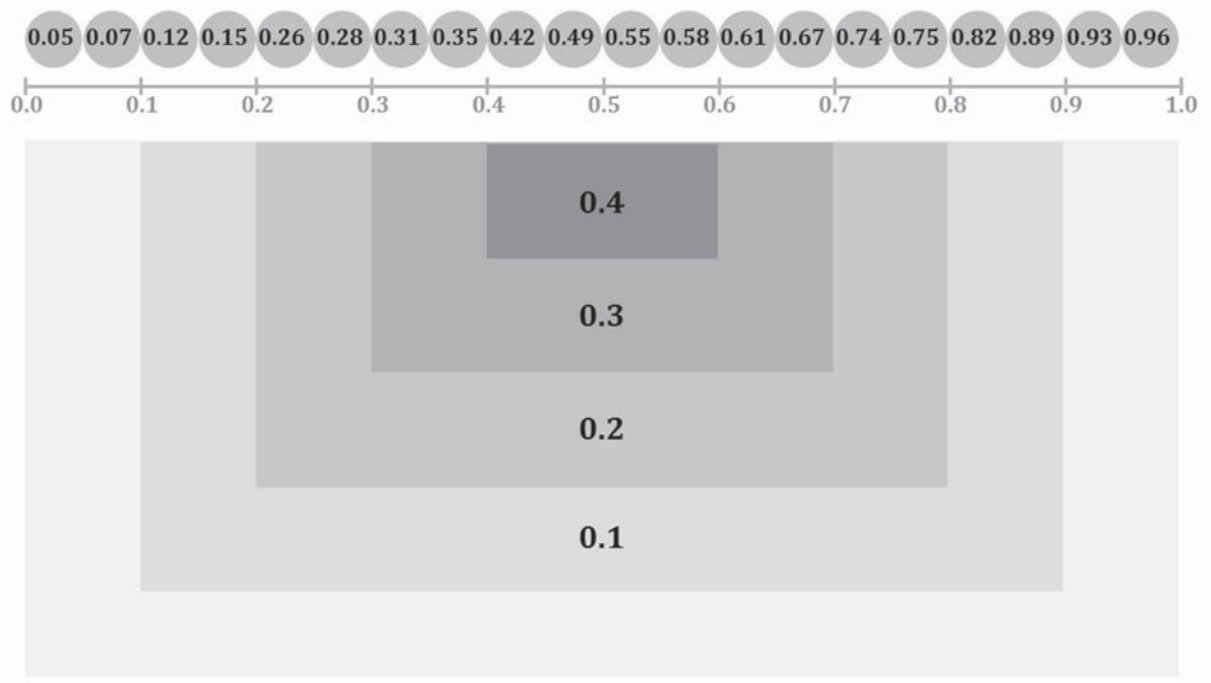


Abbildung 14: Sortierung der Objekte nach ihren Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern. Danach werden, um ein bestimmtes Signifikanzlevel einzuhalten, bestimmte Segmente ausgeschnitten. In dieser Graphik stellen die grauen Kästen den Bereich dar, welcher ausgeschnitten wird, um das Signifikanzlevel, welches in den bestimmten Kästen steht, einzuhalten. Beispielsweise würde, um ein Signifikanzlevel von 0.3 einhalten zu wollen, der Bereich von 0.3 bis 0.7 ausgeschnitten werden und diese Objekte würden zurückgewiesen werden. (Diese Objekte wären, verglichen mit dem CP, uninformativ). Dieser Ansatz beruht auf der Annahme, dass Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer vorliegen, welche nah an der wahren Wahrscheinlichkeit liegen.

4 Ergebnisse

4.1 Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern

4.1.1 Visuelle Analyse der Zuverlässigkeits-Diagramme und Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken vor und nach Kalibrierung

In dieser Studie wurden zunächst die Zuverlässigkeits-Diagramme unterschiedlicher Klassifikations- und Regressionstechniken analysiert. Um die Vergleichbarkeit zu garantieren wurden die Diagramme unter den gleichen Rahmenbedingungen generiert. Im Detail wurde ein Datensatz mit 2000 Objekten und 40 Variablen generiert, welcher sich aus zwei Klassen zusammensetzt. Beide Klassen wurden jeweils aus multivariaten Normalverteilungen ($N(\mu, I)$) generiert, das bedeutet Normalverteilungen mit dem Mittelwert μ und Covarianzmatrix I (unkorreliert), wobei sich die Mittelwerte unterschieden. Um eine gewisse Klassifikationsleistung einzustellen, wurden die Mittelwerte der beiden Klassen zusammen oder auseinander geschoben. Je weiter diese Mittelwerte auseinander liegen, desto leichter ist das Klassifikationsproblem. Die Korrektklassifizierungsrate (Acc) wurde auf 0.8 eingestellt. Die angewendeten Kalibriermethoden für die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sowie eine Beschreibung der Zuverlässigkeits-Diagramme sind in Kapitel 1.7.2 und 1.7.3 zu finden.

Die folgenden Abbildungen 15 und 16 zeigen die resultierenden Kurven der Wahrscheinlichkeitsschätzungen in den Zuverlässigkeits-Diagrammen sowie die Histogramme der Wahrscheinlichkeitsschätzungen für die Klassifikationsmethoden RF und KNN. Es ist zu erkennen, dass unter den genannten Bedingungen die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer-Kurven sowohl von RF als auch KNN einen sigmoiden Verlauf zeigen, welcher durch Kalibrierung mittels logistischer Regression den wahren Wahrscheinlichkeiten gut angenähert werden kann. Die Krümmung ist beim RF jedoch deutlich stärker als beim KNN. Der Effekt ist ebenfalls anhand der Histogramme erkennbar, nach der logistischen Regression sind die Wahrscheinlichkeitsschätzungen gleichmäßig verteilt, im Vergleich zu dem vorherigen Histogramm, bei dem es zu Häufungen in der Mitte kam. Rein visuell ist bereits zu erkennen, dass Kalibrierung in beiden Fällen

den Fehler reduziert und, dass der Fehler nach der Kalibrierung sehr niedrig sein muss. Folglich sind die Wahrscheinlichkeitsschätzer sehr nah an der wahren Wahrscheinlichkeit. Im Fall des KNN sind bereits die unkalibrierten Wahrscheinlichkeitsschätzungen näher an der wahren Wahrscheinlichkeit.

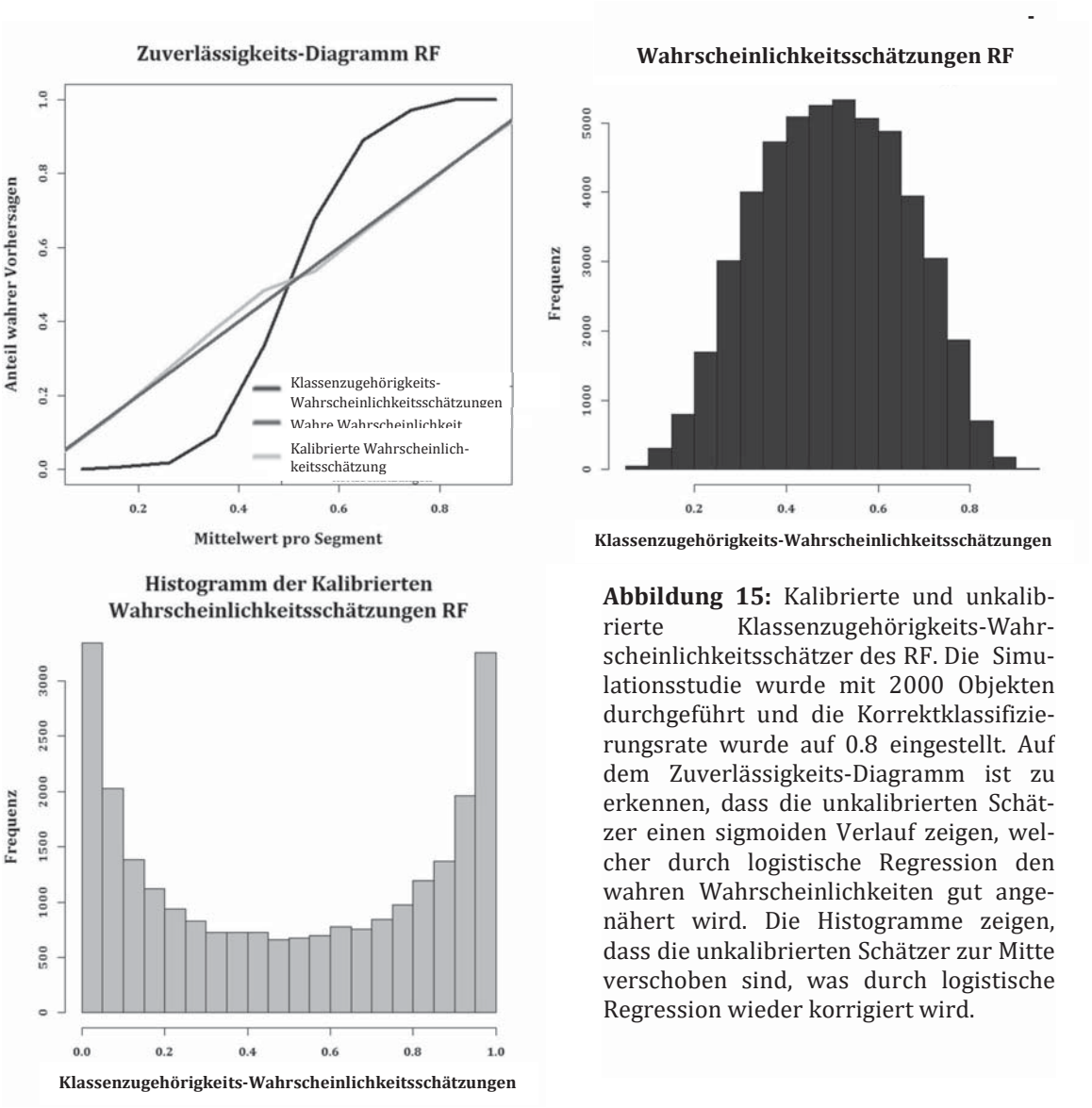
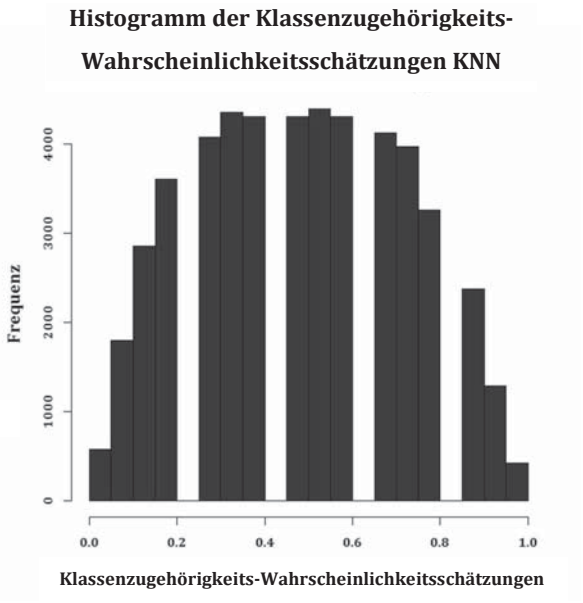


Abbildung 15: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression den wahren Wahrscheinlichkeiten gut angenähert wird. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.



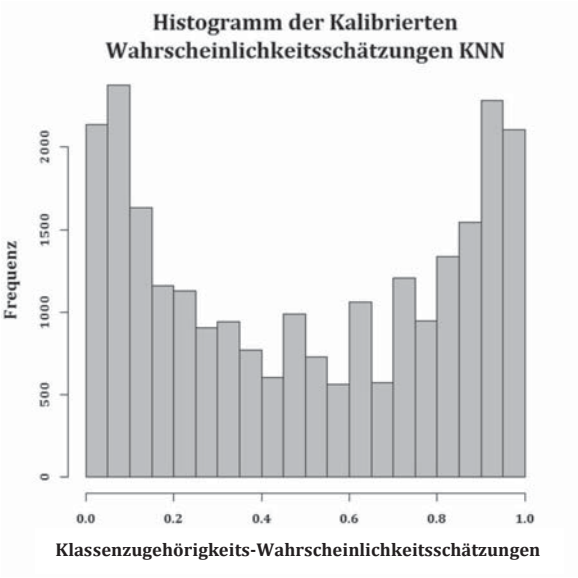
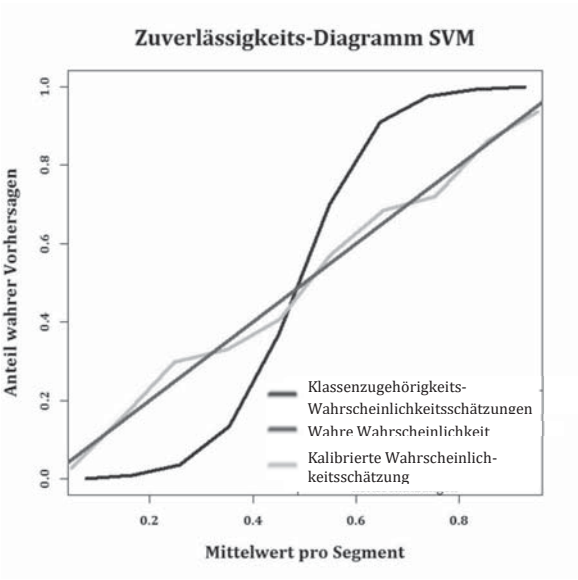
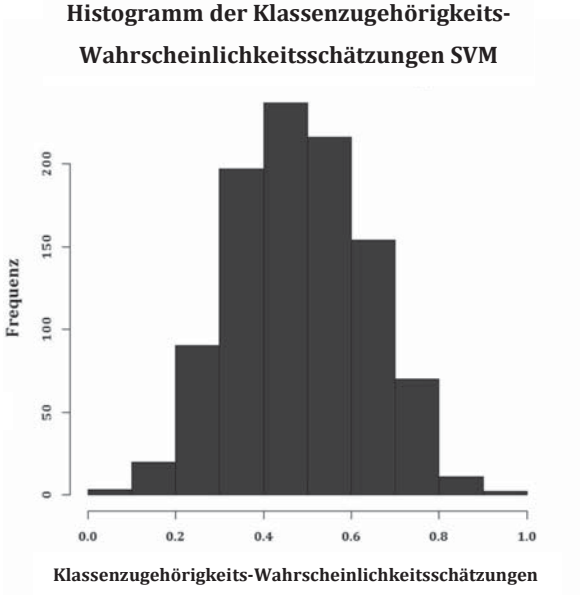


Abbildung 16: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des KNN. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert werden kann. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.

Auffällig sind beim KNN die „Lücken“ im Histogramm der unkalibrierten Wahrscheinlichkeiten, dies rührt her von der begrenzten Anzahl an Nachbarn (k=15), welche auch nur eine begrenzte Anzahl an Klassenzugehörigkeits-Wahrscheinlichkeiten ermöglichen. Die nachfolgende Abbildung 17 zeigt die Zuverlässigkeits-Diagramme und Histogramme der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der Klassifikationstechnik SVM. Ähnlich wie beim RF und beim KNN zeigen die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer im Zuverlässigkeits-Diagramm der SVM ebenfalls einen sigmoiden Verlauf, welcher mit Hilfe der logistischen Regression wieder behoben wird. Auch in den Histogrammen kann beobachtet werden, dass die Vorhersagen zunächst zur

Mitte verschoben sind und dann wieder in Richtung Gleichverteilung zurückgeschoben werden.





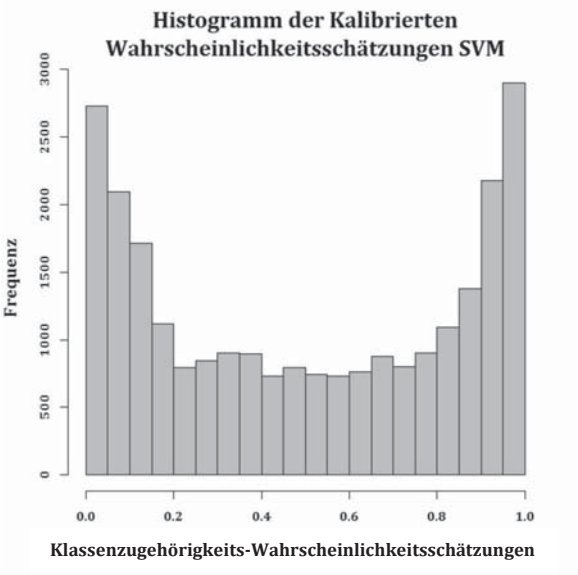


Abbildung 17: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der SVM. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert werden kann. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.

Die LDA (Abbildung 18), die NN und der NBC bringen von vornherein gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervor, welche sich nah an der wahren Wahrscheinlichkeit befinden. Dies ist sowohl anhand des Zuverlässigkeits-Diagramms als auch anhand der Histogramme zu erkennen. Die Kalibrierung hat lediglich einen geringen Einfluss oder führt zu einer Verschlechterung. An dieser Stelle werden die Ergebnisse der LDA exemplarisch gezeigt. Die Abbildungen 60 und 61 für die NN und den NBC sind im Anhang (Kapitel 9.1.1) zu finden.

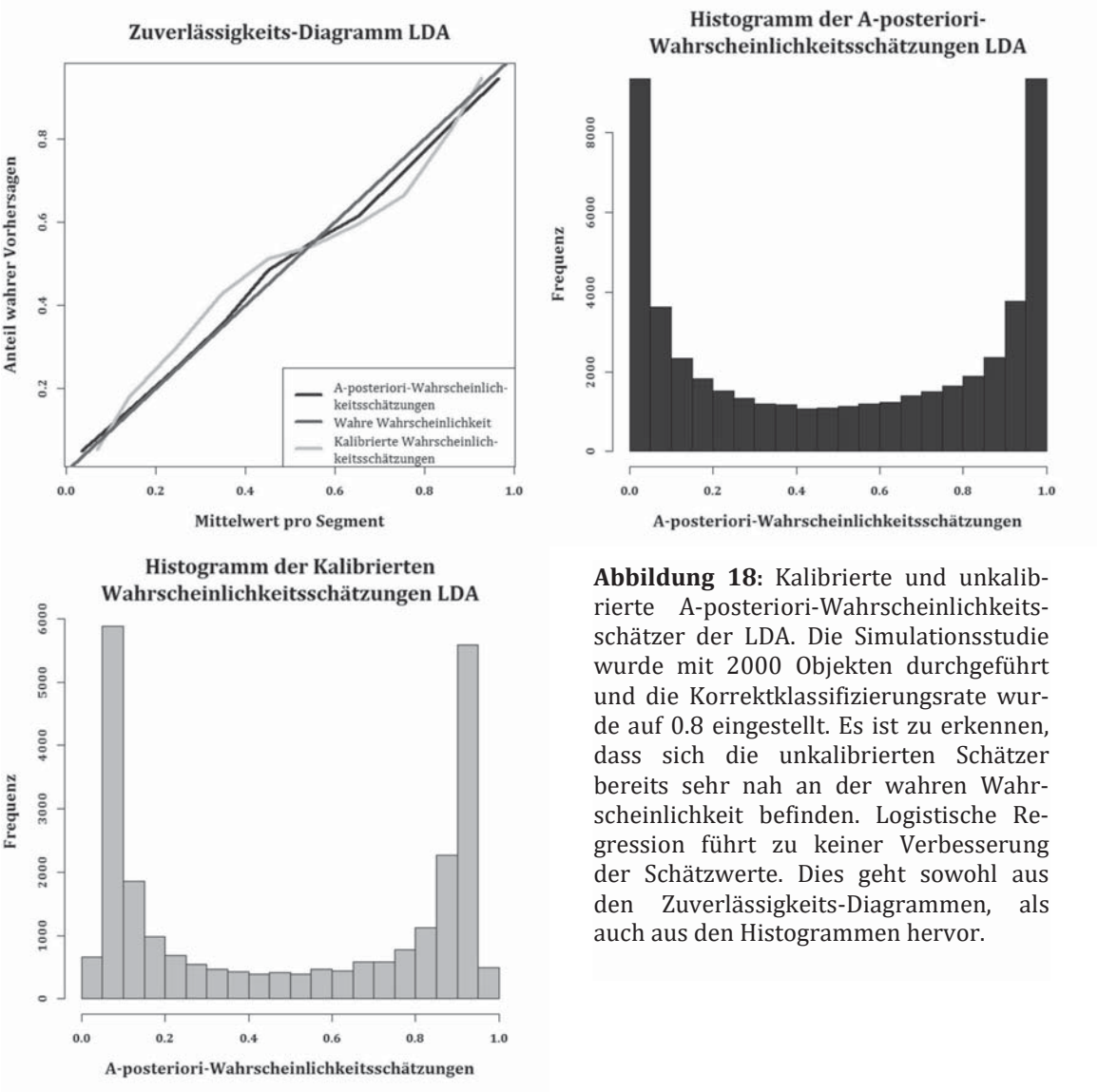


Abbildung 18: Kalibrierte und unkalibrierte A-posteriori-Wahrscheinlichkeitsschätzer der LDA. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Es ist zu erkennen, dass sich die unkalibrierten Schätzer bereits sehr nah an der wahren Wahrscheinlichkeit befinden. Logistische Regression führt zu keiner Verbesserung der Schätzwerte. Dies geht sowohl aus den Zuverlässigkeits-Diagrammen, als auch aus den Histogrammen hervor.

Bei der PLSDA (Abbildung 19) kommt es ebenfalls zu einer leichten Verschiebung der Vorhersagen zur Mitte hin. Allerdings ist dieses Phänomen deutlich schwächer ausgeprägt als beim RF, KNN oder den SVM. Durch die Kalibrierung mit Hilfe der logistischen Regression kann diese Verschiebung wieder ausgeglichen werden.

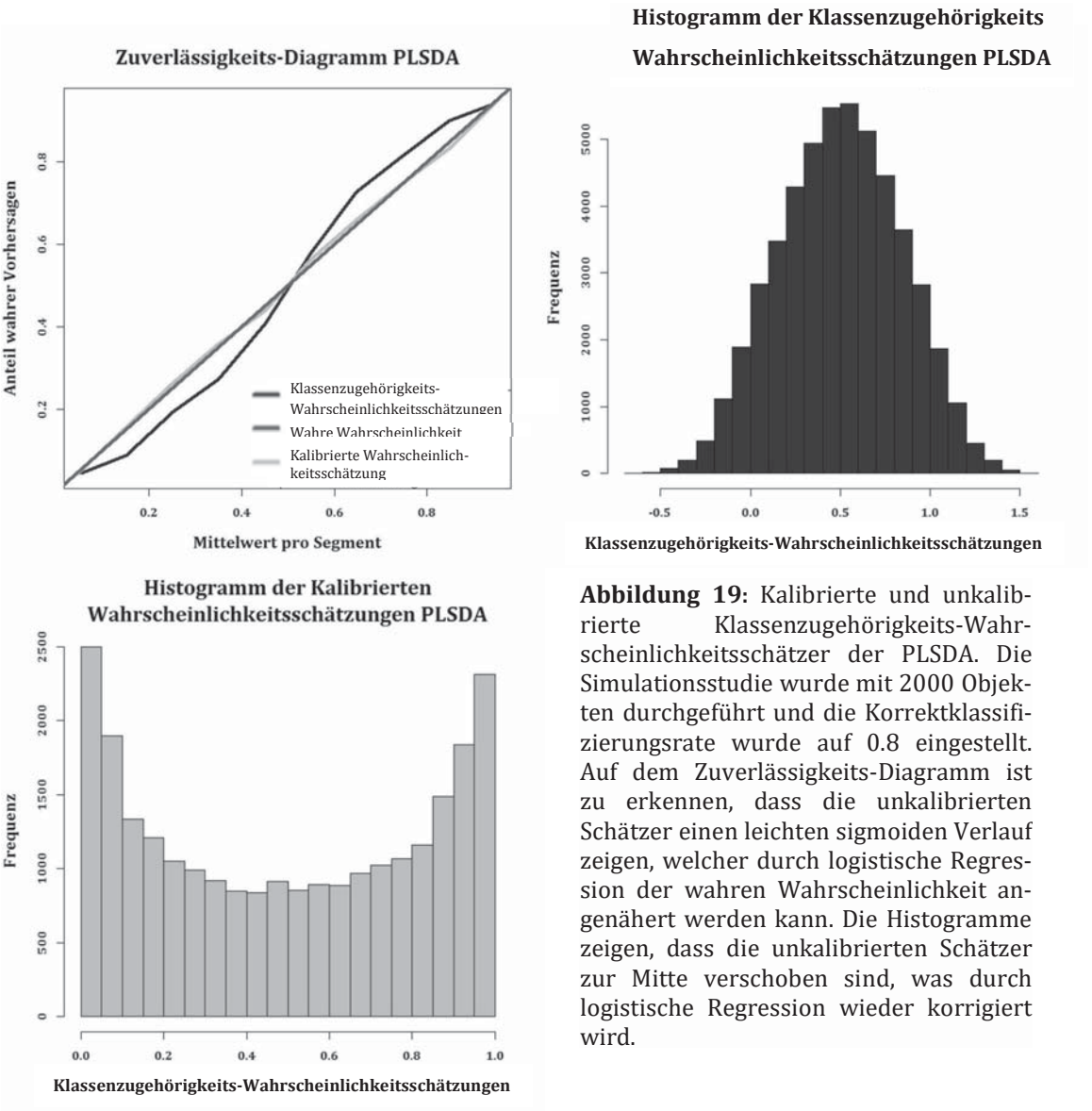


Abbildung 19: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der PLSDA. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen leichten sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert werden kann. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.

Im Regressionsfall werden die vorhergesagten kontinuierlichen Werte zur Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit genutzt (siehe Kapitel 3.1.2). Im Fall des RFR und der SVR liegen diese bereits zwischen 0 und 1. Die übrigen Regressionstechniken sind allerdings in der Lage über 0 und 1 hinaus zu extrapolieren. Aus diesem Grund werden für diese Techniken die kontinuierlichen Vorhersagen skaliert, sodass diese wieder zwischen 0 und 1 liegen. (In Kapitel 4.1.3 wird zur Lösung dieses Problems noch ein weiterer Ansatz vorgestellt und mit der Skalierung verglichen). Die Nutzung der kontinuierlichen Vorhersagen der Regressionstechniken folgt dem Ansatz, dass nichtparametrische Regressionstechniken, die den Zusammenhang zwischen Prädiktor und Klassenzugehörigkeiten gut approximieren können, ebenfalls gut zur Wahrscheinlich-

keitsschätzung geeignet sein können [110]. In den Abbildungen 20 und 21 werden die Regressionsmethoden SVR und Lasso, exemplarisch für die übrigen Regressionstechniken, näher betrachtet. Bei allen Regressionsmethoden kommt es ebenfalls zu einer Verschiebung der unkalibrierten Werte zur Mitte hin, welche durch Kalibrierung wieder korrigiert werden können. Die Abbildungen 62 bis 65 der übrigen Regressionstechniken sind im Anhang (Kapitel 9.1.1) aufgeführt.

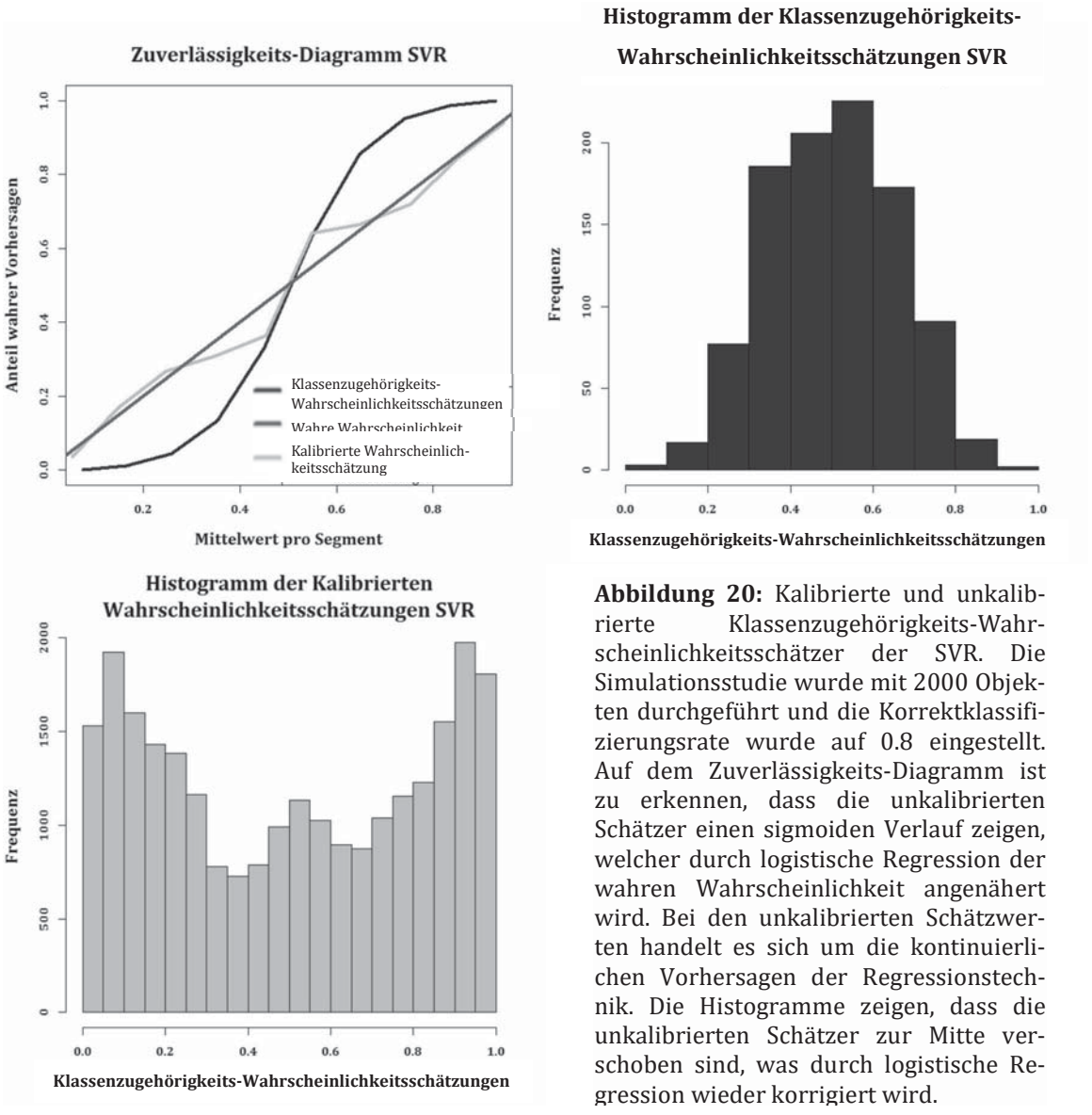


Abbildung 20: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der SVR. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert wird. Bei den unkalibrierten Schätzwerten handelt es sich um die kontinuierlichen Vorhersagen der Regressionstechnik. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.

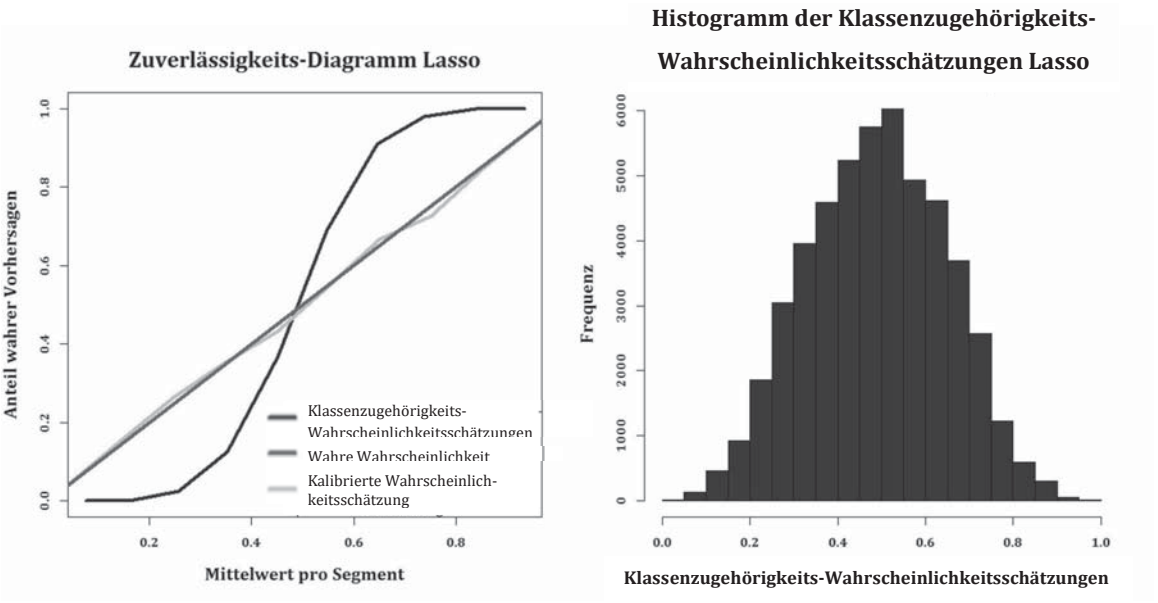


Abbildung 21: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des Lasso. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert wird. Bei den unkalibrierten Schätzwerten handelt es sich um die kontinuierlichen Vorhersagen der Regressionstechnik, welche skaliert wurden, sodass diese zwischen 0 und 1 liegen. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.

Zusammenfassend lässt sich sagen, dass unter gegebenen Voraussetzungen die Kalibrierung in allen Fällen mit Ausnahme von LDA, den NN und dem NBC, die Qualität der Wahrscheinlichkeitsschätzungen verbessert hat. Diese Techniken bringen bereits vor der Kalibrierung gute Wahrscheinlichkeitsschätzer hervor. Im Fall der LDA und des NBC ist es sogar nachteilig die Wahrscheinlichkeitsschätzungen zu kalibrieren. Die Methoden werden in weiterführenden Simulationsstudien genauer untersucht. Im Detail wird analysiert von welchen Faktoren die Qualität der Wahrscheinlichkeitsschätzung abhängt.

4.1.2 Vorversuch: Einfluss der Variablenanzahl des Datensatzes auf den Fehler sowie Beurteilung der Fehlermaße

Um den Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer zu untersuchen wurde eine Simulationsstudie durchgeführt, welche wie folgt aufgebaut wurde. Es wurde ein Datensatz bestehend aus 2000 Objekten und 2 Klassen erstellt. Die beiden Klassen wurden aus zwei multivariaten Normalverteilungen mit unterschiedlichem Erwartungswert gezogen. Die Variablen wurden zunächst nicht korreliert. Die Korrektklassifizierungsrate (Acc) wurde auf 0.9 eingestellt. Die Anzahl der Objekte und die Korrektklassifizierungsrate (Acc) wurden konstant gehalten um eine Beeinflussung durch diese Faktoren ausschließen zu können. Auf diese Weise wurden insgesamt fünf Datensätze mit 20, 40, 60, 80 und 100 Variablen erzeugt. Zur Evaluierung wurde wie bereits im Methodenteil beschrieben (Kapitel 3.1.5) eine 50*50% LMO-CV verwendet. Außerdem wurde jeder Versuch mit zufällig gewählten Startpartitionen zehn Mal wiederholt, d. h. insgesamt zehn Mal wurden die Daten neu generiert und der Mittelwert sowie die Standardabweichung der Einzelwerte wurden berechnet. Die Ergebnisse für ausgewählte Techniken (RF, SVM, SVR und Ridge) sind in den Abbildungen 22 bis 25 dargestellt.

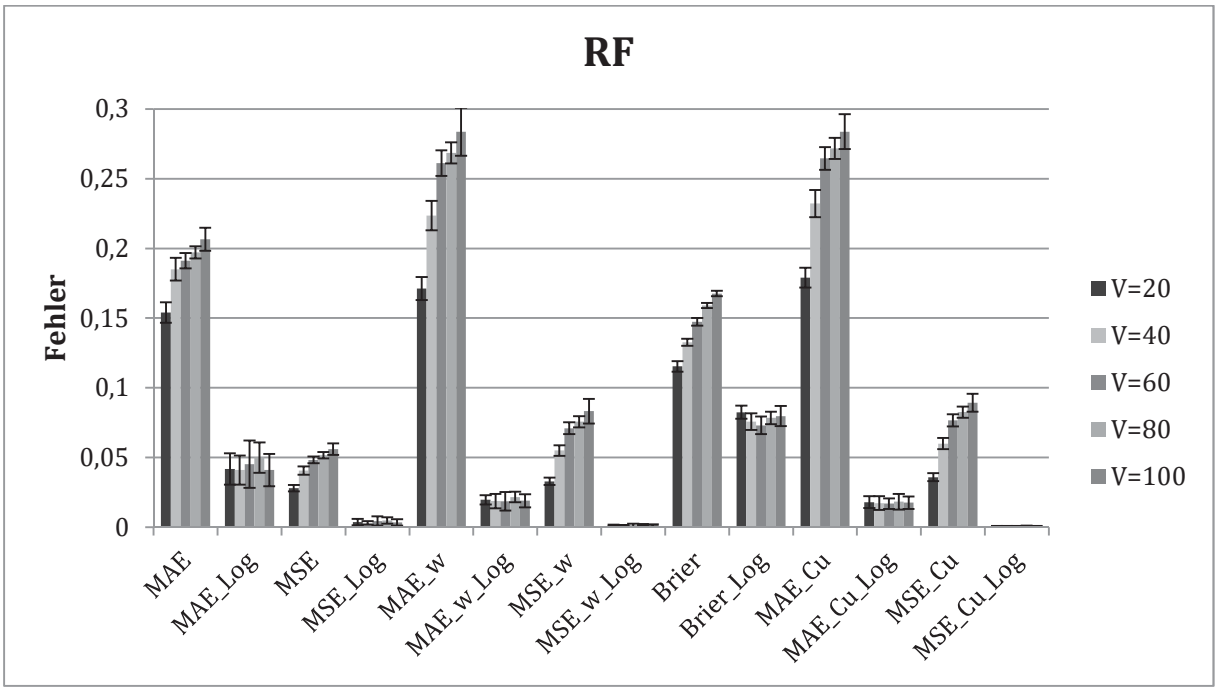


Abbildung 22: Ergebnisse der Simulationsstudie zum Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF. Bei unkorrelierten Daten steigt der Fehler vor logistischer Regression mit Zunahme der Variablenanzahl an. Aufgetragen sind die Mittelwerte der Fehler

aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte. (Eine Übersicht über alle Abkürzungen ist am Anfang dieser Arbeit im Abkürzungsverzeichnis zu finden).

Für den RF ist erkennbar, dass alle unkalibrierten Fehlermaße eine Abhängigkeit der Variablenanzahl zeigen, genauer gesagt steigt der unkalibrierte Fehler mit zunehmender Anzahl an Variablen an. Bei den Fehlermaßen nach Verwendung der logistischen Regression ist solch eine Abhängigkeit nicht erkennbar (siehe Abbildung 22). Ein vergleichbares Bild wie beim RF zeigt sich beim KNN (siehe Anhang Kapitel 9.1.2). Dies ist soweit aufgrund der Verwandtschaft der beiden Techniken zu erwarten gewesen.

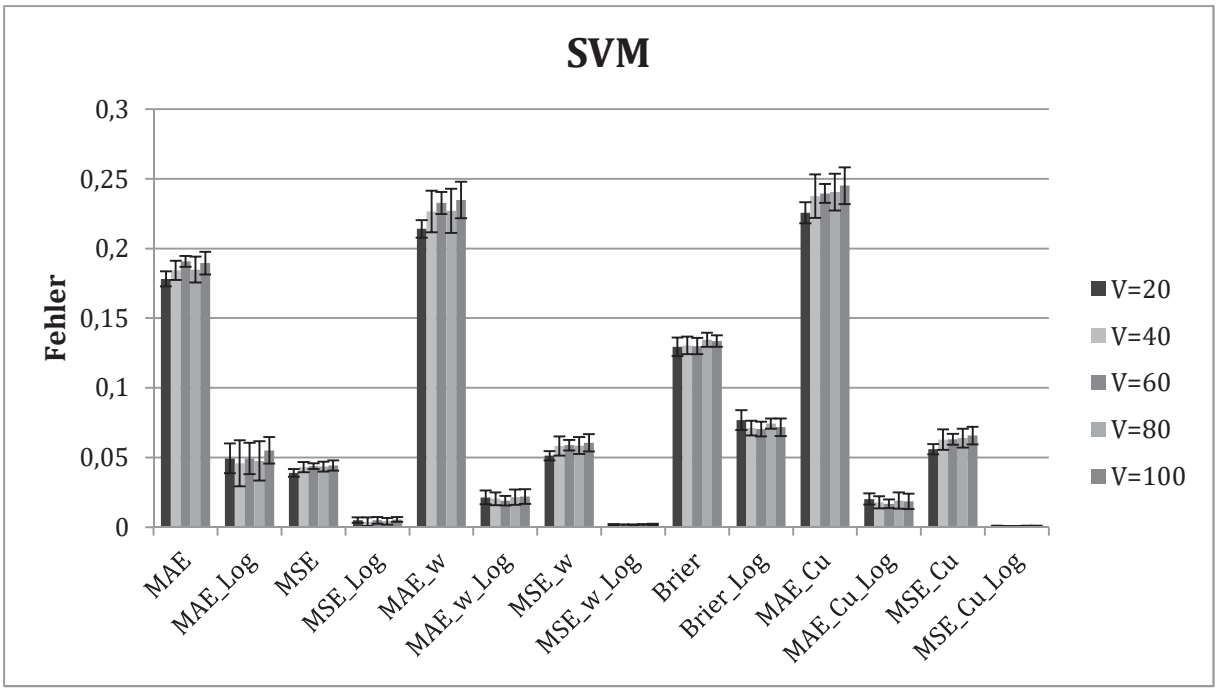


Abbildung 23: Ergebnisse der Simulationsstudie zum Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der SVM. Es ist keine Zunahme des Fehlers mit steigender Variablenanzahl zu erkennen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

Im Gegensatz zum RF und zum KNN sind bei der SVM (siehe Abbildung 23) keine Einflüsse der Variablenanzahl erkennbar, weder bei den unkalibrierten noch bei den kalibrierten Fehlermaßen. Genauso verhält es sich bei den NN (siehe Anhang Kapitel 9.1.2). Bei der SVR (siehe Abbildung 24) hingegen kann wie beim RF und KNN eine Abhängigkeit zwischen den unkalibrierten Fehlern und der Anzahl Variablen beobachtet werden. Um zu überprüfen, ob dies bei allen Regressionstechniken der Fall ist, wurde dieses Ex-

periment zusätzlich mit der Ridge Regression (siehe Abbildung 25) durchgeführt, welche darüber hinaus repräsentativ für das Elastic Net und das Lasso sein soll.

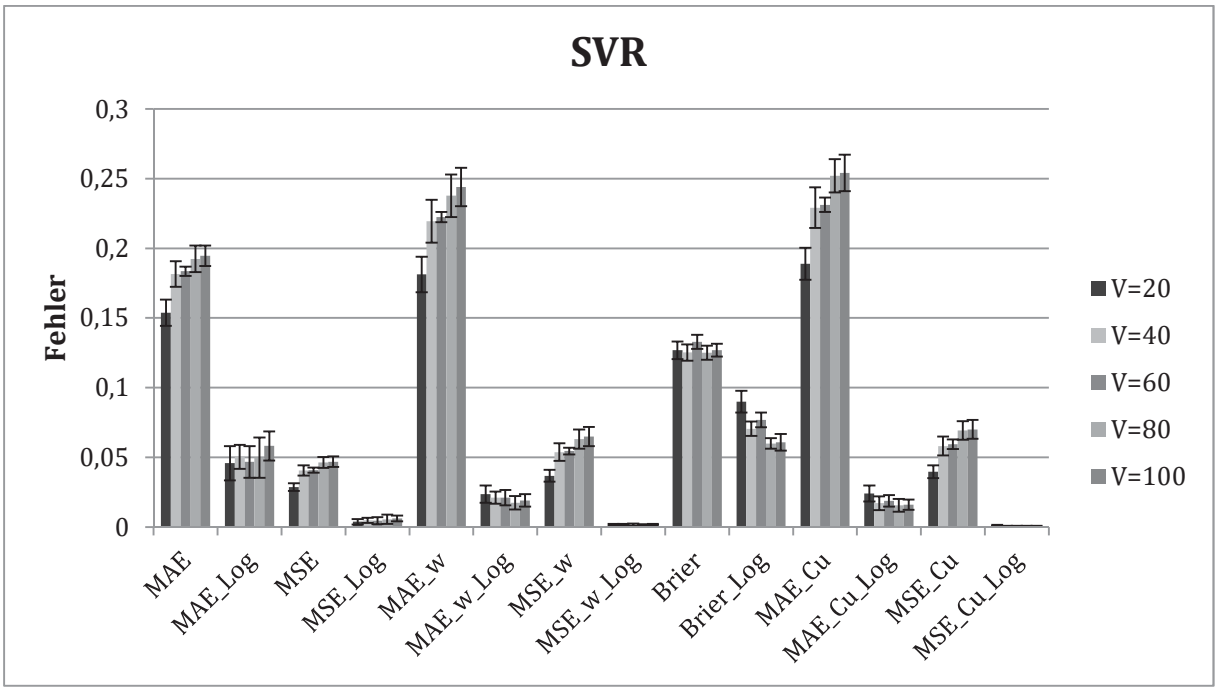


Abbildung 24: Ergebnisse der Simulationsstudie zum Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der SVR. Bei unkorrelierten Daten steigt der Fehler vor logistischer Regression mit Zunahme der Variablenanzahl an. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

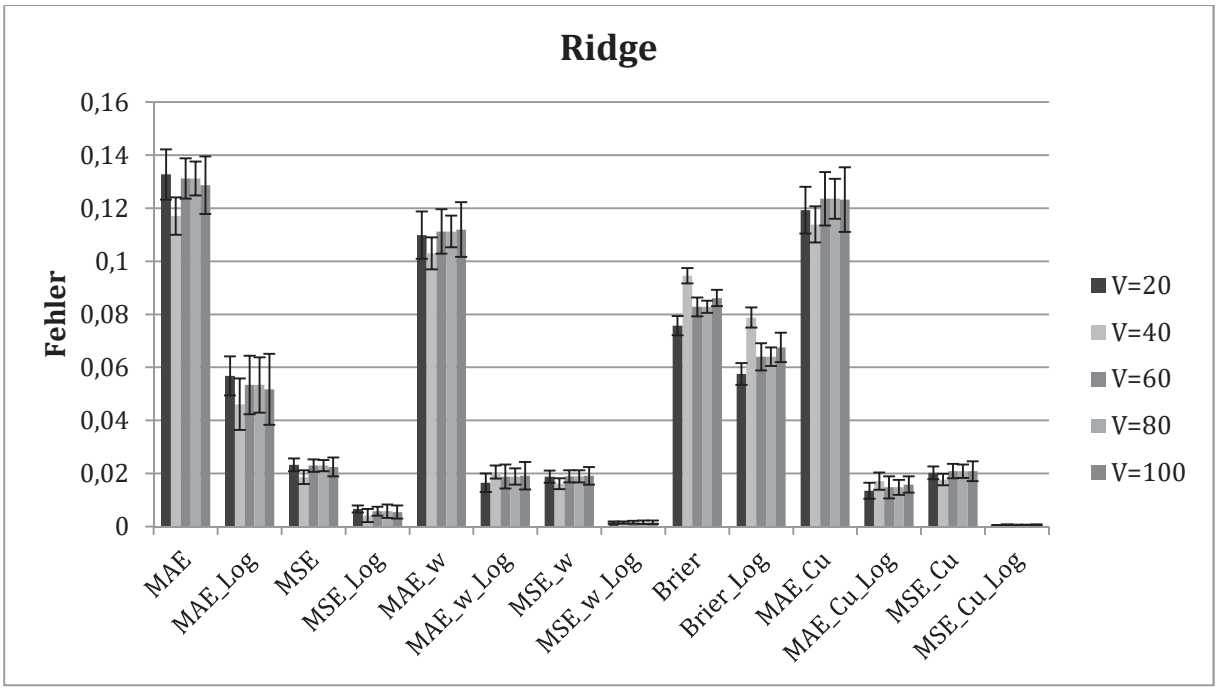


Abbildung 25: Ergebnisse der Simulationsstudie zum Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der Ridge. Es ist keine Zunahme des Fehlers mit steigender Variablenanzahl zu erkennen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

Bei dieser Technik kann wiederum kein Einfluss der Anzahl an Variablen erkannt werden. Aus kleineren vorherigen Testreihen ging hervor, dass die Korrelation der Daten möglicherweise einen stärkeren Einfluss hat. Deshalb wurden diejenigen Versuche mit korrelierten Daten wiederholt, bei denen eine Abhängigkeit erkannt wurde (RF, KNN und SVM). Die Daten wurden mit Hilfe einer Kovarianz Matrix mit Equikorrelation $\sum_E$ korreliert (rho (r) =0.2). Das bedeutet die Korrelation rho ist an jeder Stelle identisch.

$$\sum\nolimits_E = \begin{pmatrix} 1 & \dots & rho \\ \dots & 1 & rho \\ rho & rho & 1 \end{pmatrix}$$

An dieser Stelle werden exemplarisch ausschließlich die Ergebnisse des RF gezeigt, die übrigen Ergebnisse sind im Anhang zu finden (siehe Kapitel 9.1.2). Es wird beobachtet, dass die zuvor erkannte Abhängigkeit der Daten bei korrelierten Variablen nicht beobachtet werden kann. Diese Beobachtung wird auch beim KNN und bei der SVM gemacht.

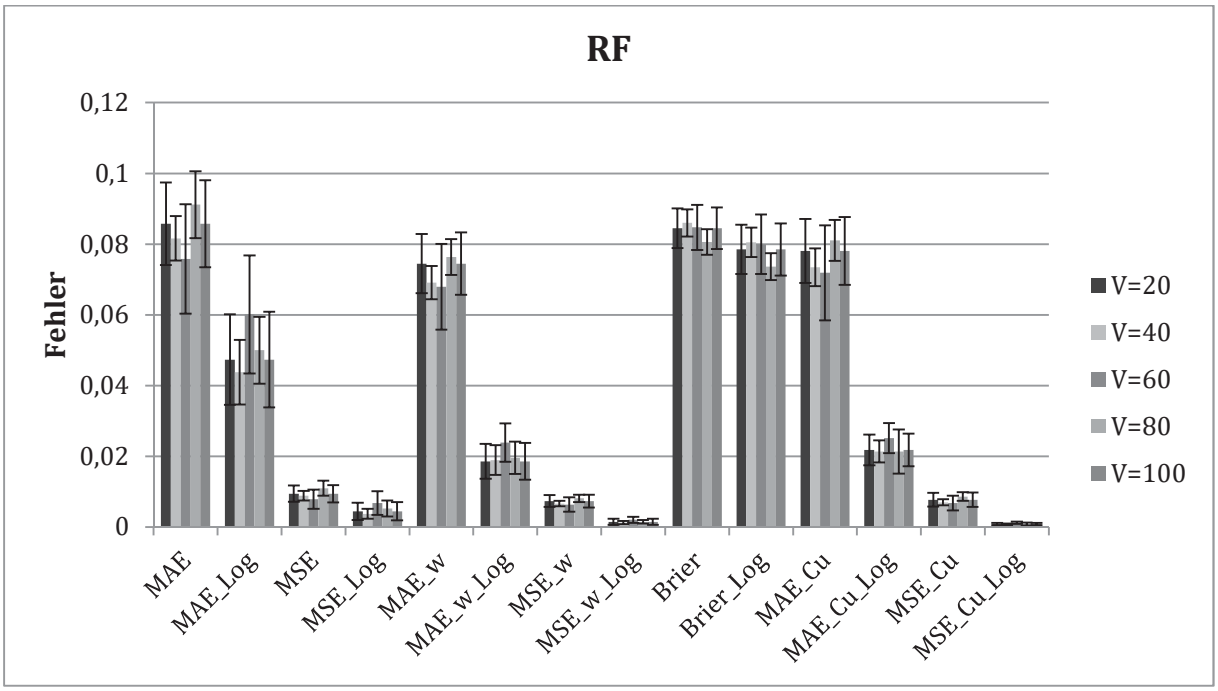


Abbildung 26: Ergebnisse der Simulationsstudie zum Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF mit korrelierten Daten. Es ist keine Zunahme des Fehlers mit steigender Variablenanzahl zu erkennen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

Auf den oben gezeigten Abbildungen ist darüber hinaus erkennbar, dass sich die jeweiligen MAE- und MSE-Werte und auch die gewichteten MAE- und MSE-Werte jeweils vor und nach logistischer Regression nicht in ihrer Verlaufsform unterscheiden. Aufgrund der Berechnungsweise war eine Ähnlichkeit der Ergebnisse zu erwarten. Allerdings bestand die Vermutung, dass sich der quadratische Fehler aufgrund der stärkeren Gewichtung von Ausreißern anders verhalten könnte. Die einzige Ausnahme ist der Brier-Score, weil dieser nicht nur den Fehler zur wahren Wahrscheinlichkeit, sondern auch die Leistung der Klassifikationstechnik bewertet. Um diese Beobachtungen numerisch zu belegen wurde deshalb am Beispiel der Simulationsstudie des RF mit zehn Variablen eine Korrelationsmatrix erstellt (Tabelle 6). Diese Tabelle untermauert die zunächst beobachteten Ergebnisse. Auf Grundlage dieser Ergebnisse werden im weiteren Verlauf dieser Arbeit nicht mehr alle Maße im Text beschrieben, es werden allerdings alle Ergebnisse (zu allen Maßen) im Anhang (Kapitel 9.1.2) tabellarisch aufgeführt.

Tabelle 6: Korrelationsmatrix der oben abgebildeten Simulationsstudie des RF mit zehn Variablen. Es ist zu erkennen, dass die jeweiligen MAE- und MSE-Werte (blau unterlegt) und auch die gewichteten MAE- und MSE-Werte jeweils vor und nach logistischer Regression (gelb unterlegt) stark miteinander korrelieren. Die einzige Ausnahme stellt der Brier-Score dar, da dieser nicht nur den Fehler zur wahren Wahrscheinlichkeit sondern auch die Leistung der Klassifikationstechnik bewertet.

	MAE	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MAE	1	-0.13	0.97	-0.08	0.97	-0.29	0.94	-0.23	-0.54	-0.85	0.93	-0.55	0.92	-0.44
MAE_Log		1	0.02	0.93	-0.22	0.91	-0.02	0.96	-0.31	-0.03	-0.20	0.73	-0.06	0.77
MSE			1	0.07	0.96	-0.18	0.98	-0.10	-0.64	-0.90	0.94	-0.43	0.95	-0.28
MSE_Log				1	-0.16	0.82	0.03	0.94	-0.36	-0.11	-0.17	0.66	0.01	0.73
MAE_w					1	-0.36	0.97	-0.32	-0.56	-0.86	0.98	-0.56	0.96	-0.41
MAE_w_Log						1	-0.21	0.94	-0.14	0.22	-0.35	0.84	-0.25	0.81
MSE_w							1	-0.14	-0.68	-0.92	0.97	-0.41	0.98	-0.24
MSE_w_Log								1	-0.23	0.08	-0.32	0.77	-0.17	0.80
Brier									1	0.85	-0.57	0.00	-0.67	-0.15
Brier_Log										1	-0.87	0.42	-0.93	0.23
MAE_Cu											1	-0.58	0.97	-0.43
MAE_Cu_Log												1	-0.49	0.96
MSE_Cu													1	-0.30
MSE_Cu_Log														1

4.1.3 Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels Simulationsstudien

Für jede untersuchte Klassifikationstechnik (RF, KNN, SVM, NN, LDA, PLSDA, NBC) und jede Regressionstechnik (SPLS, RFR, SVR, Ridge, Elastic Net, Lasso) wurden Simulationsstudien durchgeführt. Bei diesen wurde die Anzahl an Objekten (500, 1000, 2000, 4000), die Korrektklassifizierungsrate (Acc) (0.7, 0.75, 0.8, 0.85, 0.9) und die Korrelation r (r=0, r=0.1, r=0.2) variiert, um den Einfluss dieser Faktoren auf die Höhe des Fehlers und damit auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer zu analy-

sieren. Die Anzahl an Variablen wurde konstant auf 40 gesetzt, da der Einfluss dieser bereits im letzten Kapitel untersucht wurde. Zur Evaluierung wurde wie bereits im Methodenteil beschrieben (Kapitel 3.1.5) eine 50*50% LMO-CV verwendet. Außerdem wurde jeder Versuch mit zufällig gewählten Startpartitionen zehn Mal wiederholt, d. h. insgesamt zehn Mal wurden die Daten neu generiert und der Mittelwert sowie die Standardabweichung der Einzelwerte der MAE_Cu- und MAE_Cu_Log-Werte wurden berechnet. Begonnen wurden mit den Klassifikationstechniken. In Abbildung 27 sind die Ergebnisse für den RF mit 4000 Objekten dargestellt. Es ist zu erkennen, dass der MAE_Cu mit steigender Korrektklassifizierungsrate (Acc) zunimmt und mit zunehmender Korrelation abnimmt. Mit r=0.2 und einer Korrektklassifizierungsrate (Acc) von 0.7-0.8 werden gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorgebracht (unter der roten Linie). Nach logistischer Regression sind unabhängig von der Korrektklassifizierungsrate (Acc) und der Korrelation der Daten alle Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer gut kalibriert.

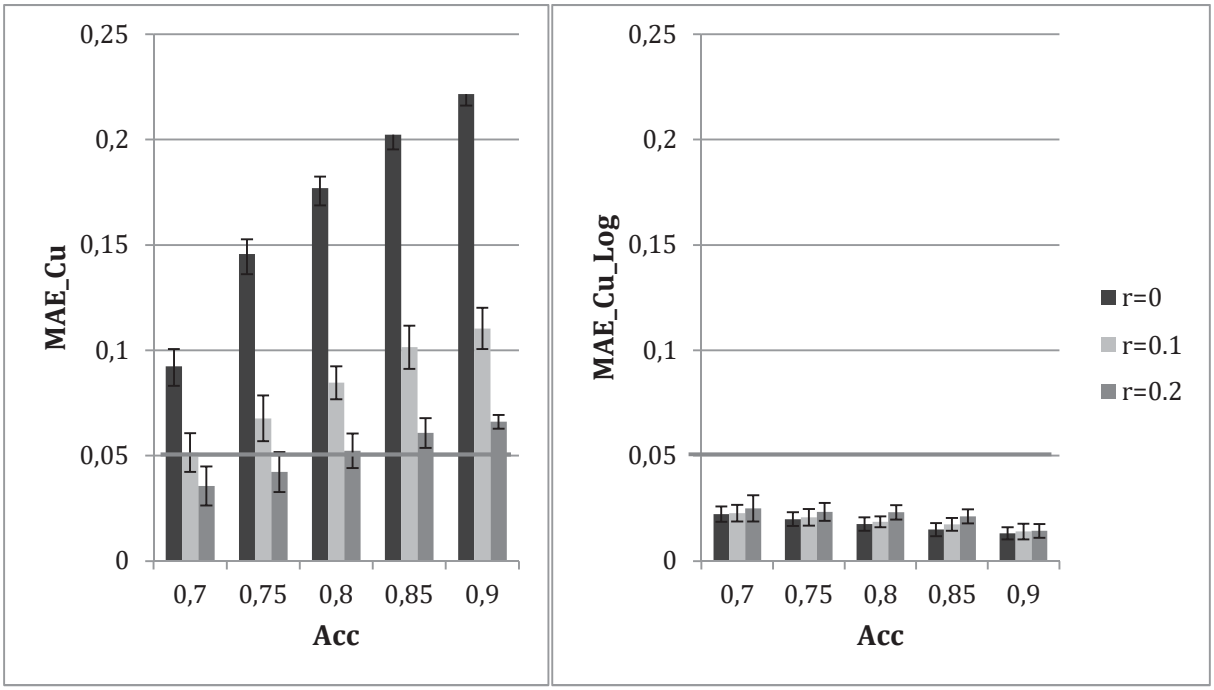


Abbildung 27: Ergebnisse für den **RF** mit 4000 Objekten, 40 Variablen und r=0. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) und mit abnehmender Korrelation steigt der MAE_Cu an. Nur bei korrelierten Daten und mittlerer Korrektklassifizierungsrate (Acc) gibt der RF gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

Die übrigen Fehlermaße werden im Anhang tabellarisch aufgeführt (Kapitel 9.1.3). Abbildung 28 zeigt exemplarisch die Abhängigkeit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF, bei einer Korrektklassifizierungsrate (Acc) von 0.8 und r=0, von der Anzahl der Objekte im Datensatz.

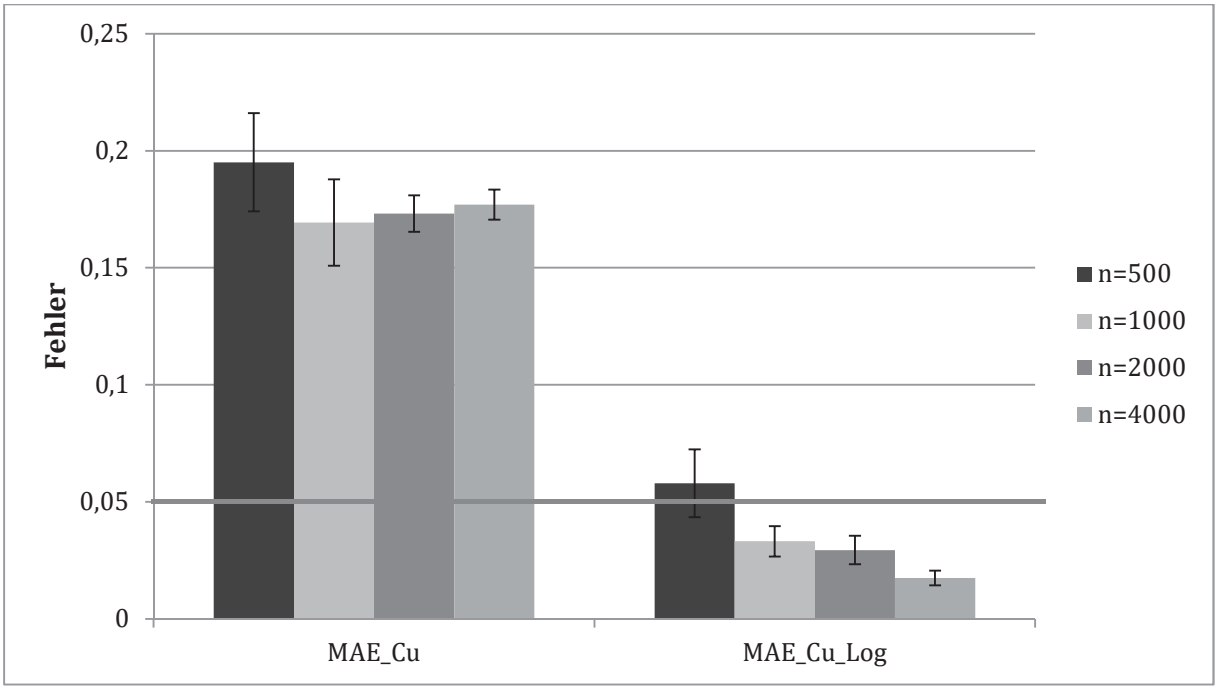


Abbildung 28: Ergebnisse für den **RF** mit einer Korrektklassifizierungsrate (Acc)=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab, genauso wie die Standardabweichung. Nur nach Kalibrierung werden ab 1000 Molekülen gut kalibrierte Wahrscheinlichkeitsschätzer hervorgebracht (Werte befinden sich unterhalb der roten Linie).

Aus der Abbildung geht hervor, dass der MAE_Cu unabhängig von der Datensatzgröße relativ konstant bleibt. Lediglich bei 500 Objekten ist der Fehler tendenziell etwas höher, genauer gesagt auf der Höhe der roten Linie (Fehler: 0.05). Wie bereits erwähnt, sind die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF bei unkorrelierten Daten nicht sehr nah an der wahren Wahrscheinlichkeit (rote Linie). Der MAE_Cu_Log hingegen zeigt eine Abhängigkeit von der Datensatzgröße, der Fehler nimmt mit steigender Größe ab. Ab einer Datensatzgröße von 1000 Molekülen werden nach Kalibrierung nah an der wahren Wahrscheinlichkeit liegende Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorgebracht. Es kann zusätzlich beobachtet werden, dass mit zunehmender Anzahl an Objekten bei beiden Fehlermaßen erwartungsgemäß

die Standardabweichung abnimmt. Diese Ergebnisse sind auch repräsentativ für die übrigen Korrektklassifizierungsraten (Acc) und Korrelationsvarianten (siehe im Anhang Kapitel 9.1.3). Beim KNN (Abbildung 29) kann ein ähnlicher Verlauf beobachtet werden wie beim RF. Allerdings sind die Absolutwerte der unkalibrierten Fehler beim RF höher als beim KNN. Mit steigender Korrektklassifizierungsrate (Acc) steigt auch der MAE_Cu. Jedoch gilt dies ausschließlich für r=0. Für r=0.1 und r=0.2 ist der Fehler unabhängig von der Korrektklassifizierungsrate (Acc) und, im Gegensatz zum Fehler bei r=0, nah an der wahren Wahrscheinlichkeit. Der MAE_Cu_Log ist unabhängig von der Korrektklassifizierungsrate (Acc) und der Korrelation und befindet sich ebenfalls unterhalb der roten Linie.

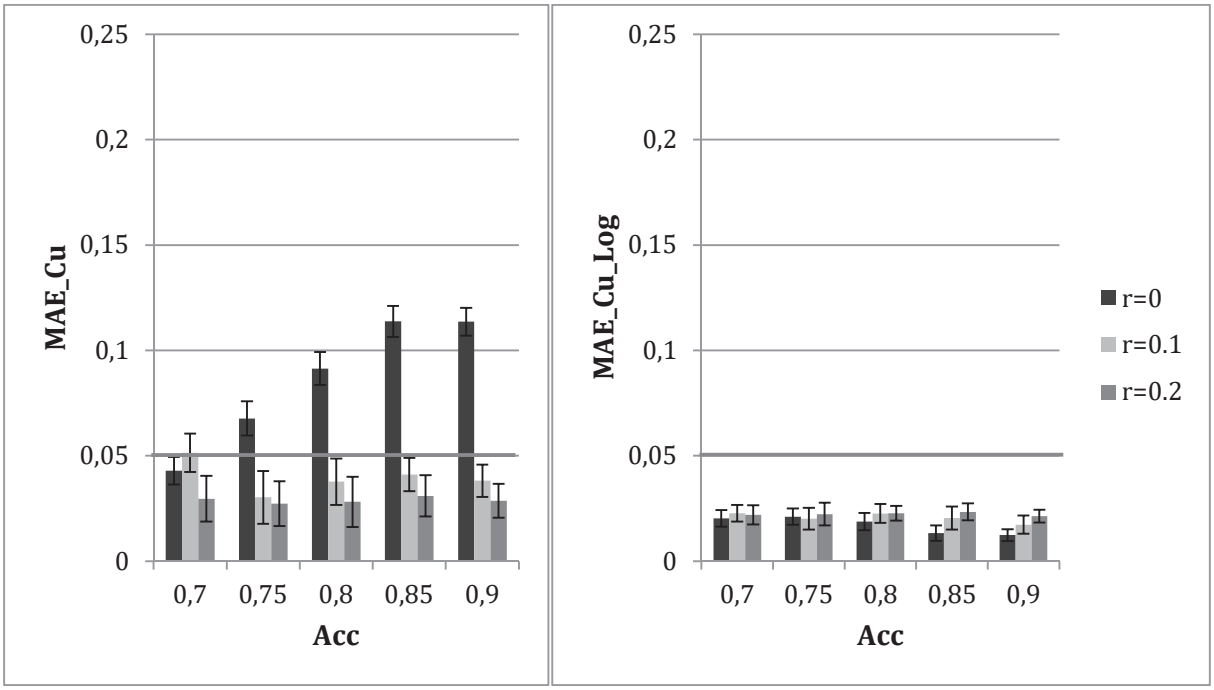


Abbildung 29: Ergebnisse für den **KNN** mit 4000 Objekten, 40 Variablen und r=0. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) und mit abnehmender Korrelation steigt der MAE_Cu an. Nur bei korrelierten Daten gibt der KNN gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

Bei der Betrachtung der Abhängigkeit der Fehlergröße beim KNN von der Datensatzgröße (Abbildung 30), fällt auf, dass sowohl der MAE_Cu, als auch der MAE_Cu_Log mit steigender Anzahl an Objekten tendenziell abnimmt. Allerdings befinden sich lediglich

die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer nach Kalibrierung mit logistischer Regression nah an der wahren Wahrscheinlichkeit.

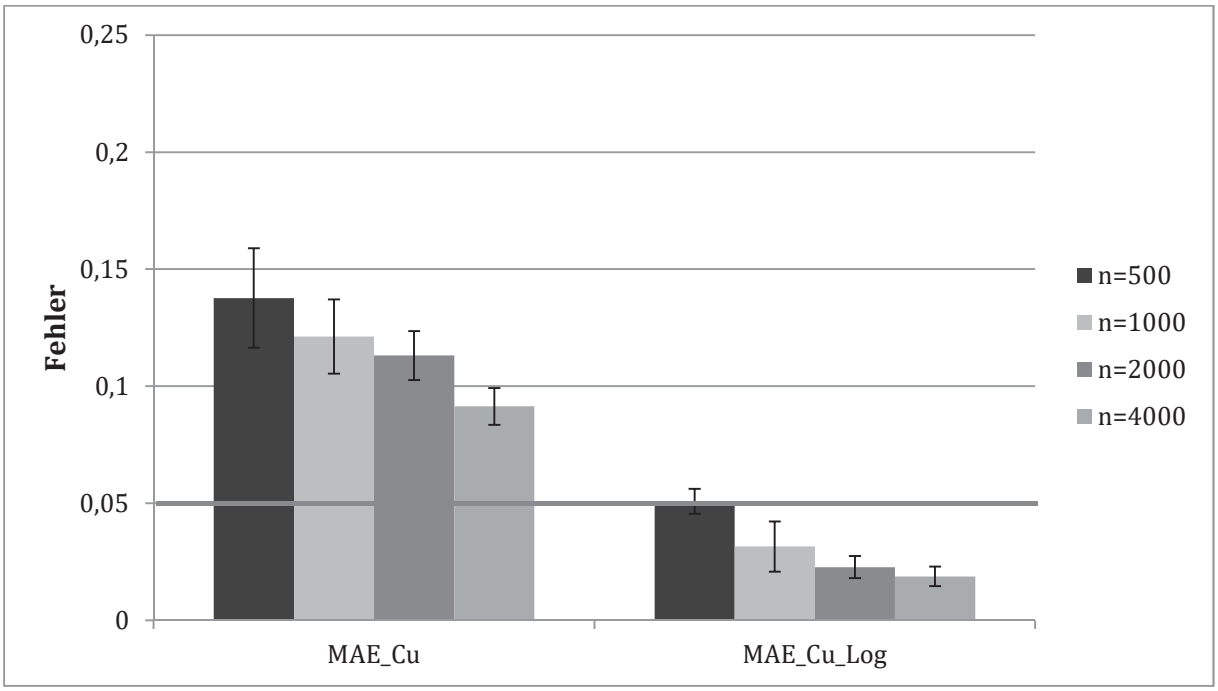


Abbildung 30: Ergebnisse für den **KNN** mit einer Korrektklassifizierungsrate (Acc)=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Sowohl für den MAE_Cu, als auch für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab. Nur nach Kalibrierung werden Wahrscheinlichkeitsschätzer nah der wahren Wahrscheinlichkeit hervorgebracht (Werte befinden sich unterhalb der roten Linie).

In Abbildung 31 sind die Ergebnisse für die SVM dargestellt. Beobachtet wird, dass der Fehler für die unkalibrierten Schätzwerte wie beim RF mit steigender Korrektklassifizierungsrate (Acc) zunimmt und mit stärkerer Korrelation abnimmt. Dennoch liegt mit Ausnahme des Wertes bei Korrektklassifizierungsrate (Acc) 0.7 und r=0.2, keiner der Werte nah an der wahren Wahrscheinlichkeit. Im Gegenteil, diese sind wie beim RF sehr hoch. Anders verhält es sich bei den kalibrierten Werten, alle MAE_Cu_Log-Werte verlaufen unterhalb der roten Linie und somit nah an der wahren Wahrscheinlichkeit. In Abbildung 32 wird für die SVM die Abhängigkeit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von der Datensatzgröße exemplarisch gezeigt. Für die SVM werden dieselben Beobachtungen wie beim RF gemacht. Der MAE_Cu ist unabhängig von der Datensatzgröße stabil und der MAE_Cu_Log nimmt mit zunehmender Anzahl an

Objekten ab. Erst nach logistischer Regression werden ab circa 500 Molekülen gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorgebracht.

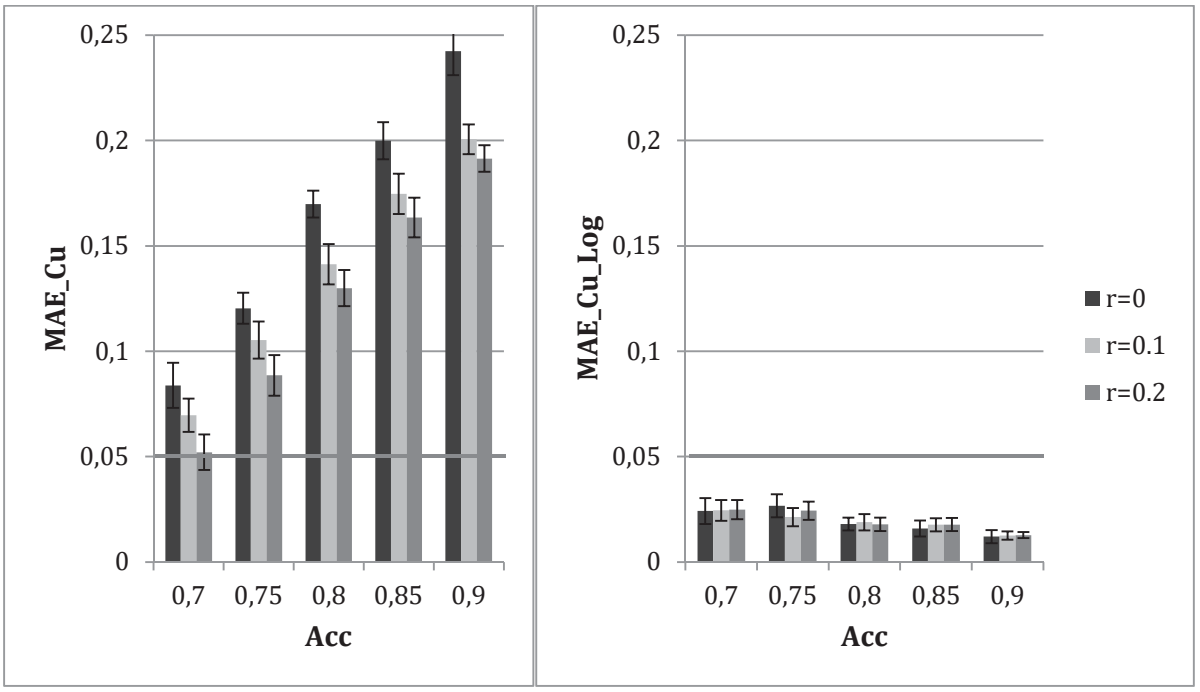


Abbildung 31: Ergebnisse für die **SVM** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) und mit abnehmender Korrelation steigt der MAE_Cu an. Nur bei einer Korrektklassifizierungsrate (Acc)=0.7 und r=0.2 gibt die SVM gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

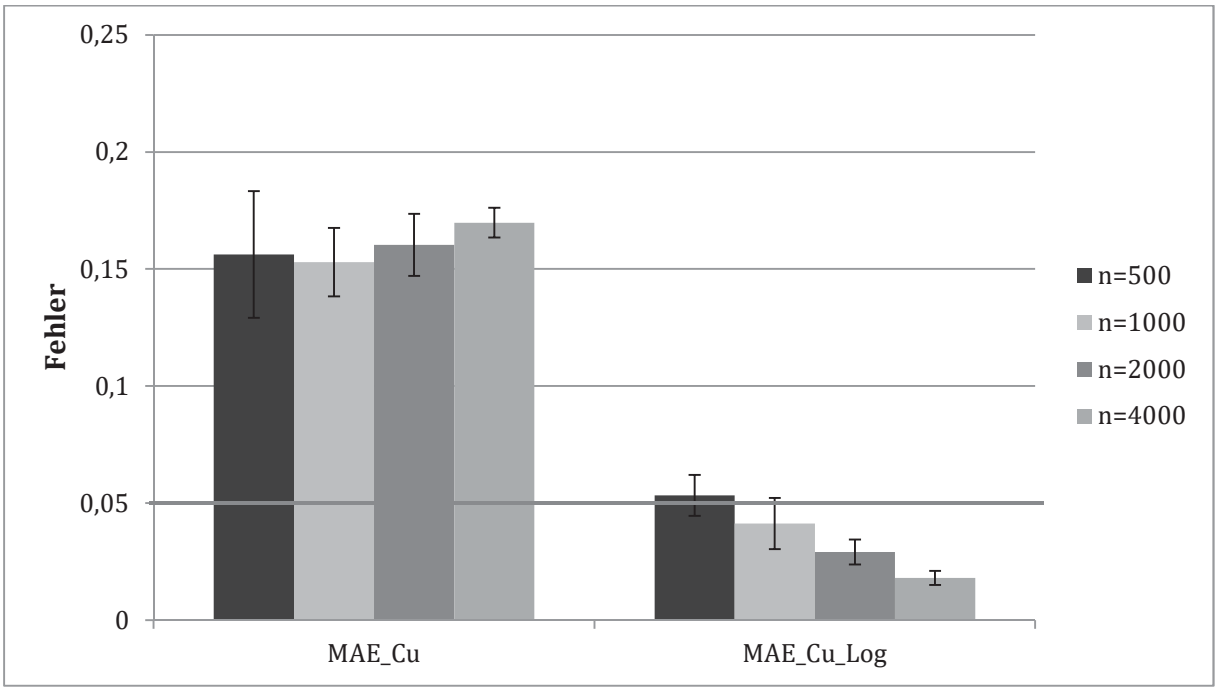


Abbildung 32: Ergebnisse für die **SVM** mit einer Korrektklassifizierungsrate (Acc)=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab, genauso wie die Standardabweichung. Nur nach Kalibrierung werden ab 500 Molekülen gut kalibrierte Wahrscheinlichkeitsschätzer hervorgebracht (Werte befinden sich unterhalb der roten Linie).

Als nächstes werden die Ergebnisse der NN betrachtet, welche in Abbildung 33 zu finden sind. Aus der Abbildung geht hervor, dass der MAE_Cu mit steigender Korrektklassifizierungsrate (Acc), unabhängig von der Korrelation leicht abnimmt. Der MAE_Cu_Log weist keinen Trend auf. Er ist unabhängig von der Korrektklassifizierungsrate (Acc) und der Korrelation gleich niedrig. Sowohl der unkalibrierte als auch der kalibrierte Fehler sind sehr niedrig. Der unkalibrierte Fehler befindet sich knapp überhalb der roten Linie und der kalibrierte Fehler knapp unterhalb. Folglich sind beide recht nah an der wahren Wahrscheinlichkeit. In Abbildung 34 wird, wie bei den letzten Techniken, beispielhaft die Abhängigkeit des Fehlers von der Datensatzgröße gezeigt. Es wird keine Abhängigkeit der Fehler von der Datensatzgröße beobachtet, mit Außnahme des Wertes bei 500 Objekten, welcher im Vergleich zu allen anderen Werten, deutlich höher liegt. Ab einer Datensatzgröße von 1000 Molekülen sind sowohl die kalibrierten als auch die unkalibrierten Wahrscheinlichkeitsschätzer nah an der wahren Wahrscheinlichkeit.

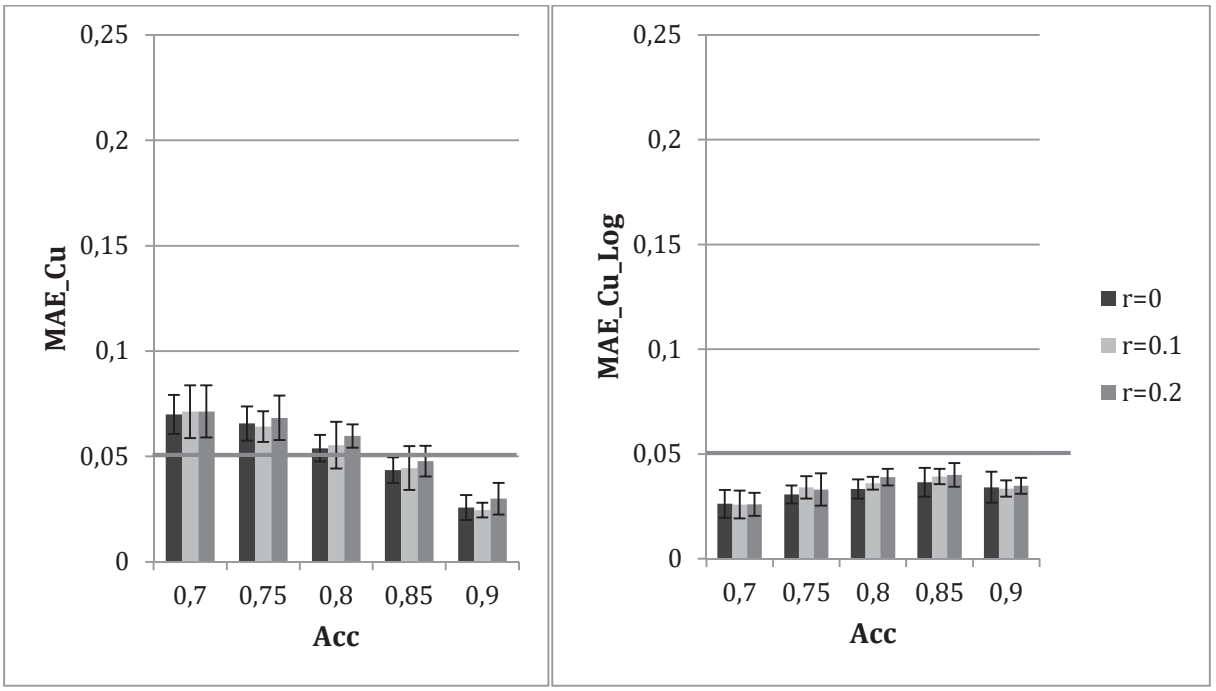


Abbildung 33: Ergebnisse für die **NN** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) nimmt der MAE_Cu leicht ab. Es ist keine Abhängigkeit von der Korrelation zu erkennen. Bereits vor der Kalibrierung geben die NN gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (alle Werte sind nicht weit von der roten Linie entfernt). Der MAE_Cu_Log hingegen bleibt ebenfalls immer niedrig. Alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

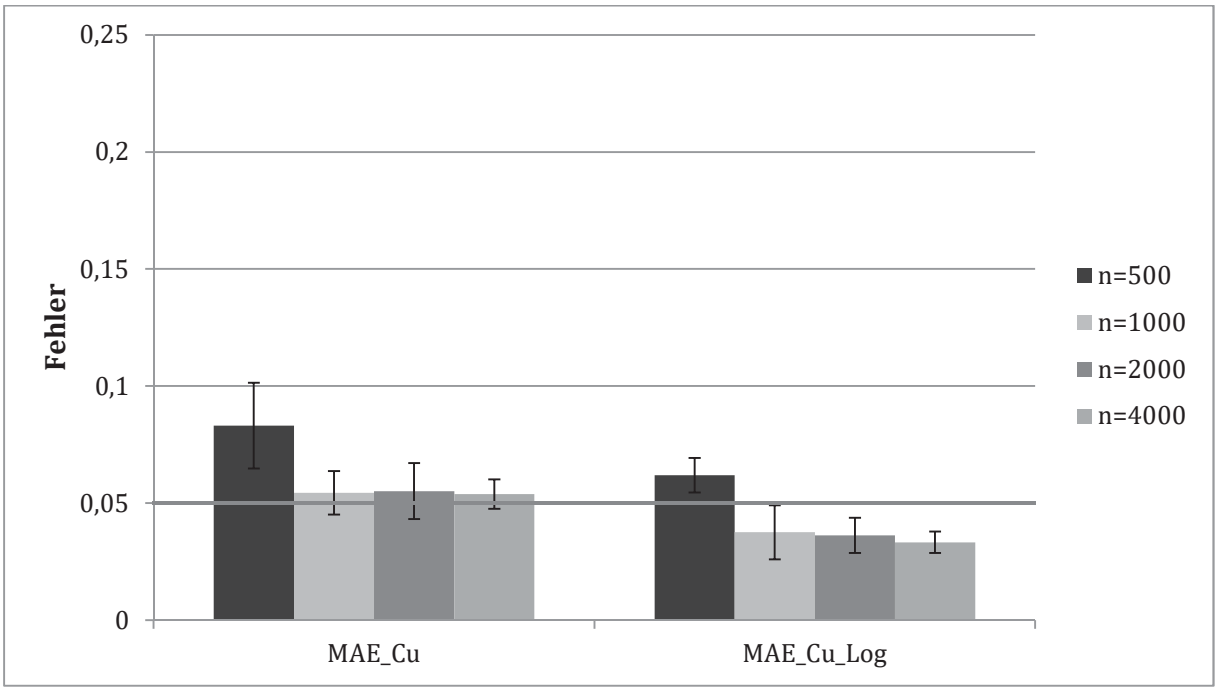


Abbildung 34: Ergebnisse für die **NN** mit einer Korrektklassifizierungsrate (Acc)=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Weder für den MAE_Cu noch für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Die Standardabweichung nimmt erwartungsgemäß mit zunehmender Datensatzgröße ab. Ab einer Datensatzgröße von 1000 Molekülen werden sowohl kalibriert als auch unkalibriert gute Wahrscheinlichkeitsschätzer erhalten (Werte befinden sich unterhalb oder auf Höhe der roten Linie).

Nachfolgend werden die Ergebnisse der LDA beschrieben, welche auf Abbildung 35 zu sehen sind. Es ist zu sehen, dass sowohl der MAE_Cu als auch der MAE_Cu_Log sehr niedrig sind und unterhalb der roten Linie verlaufen. Mit steigender Korrektklassifizierungsrate (Acc) sinkt der MAE_Cu minimal, allerdings ist keine Abhängigkeit von der Korrelation zu erkennen. Beim MAE_Cu_Log ist weder von der Korrektklassifizierungsrate (Acc) noch von der Korrelation eine Abhängigkeit zu erkennen. Die LDA bringt sowohl unkalibriert als auch kalibriert sich sehr nah an der wahren Wahrscheinlichkeit befindene Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervor.

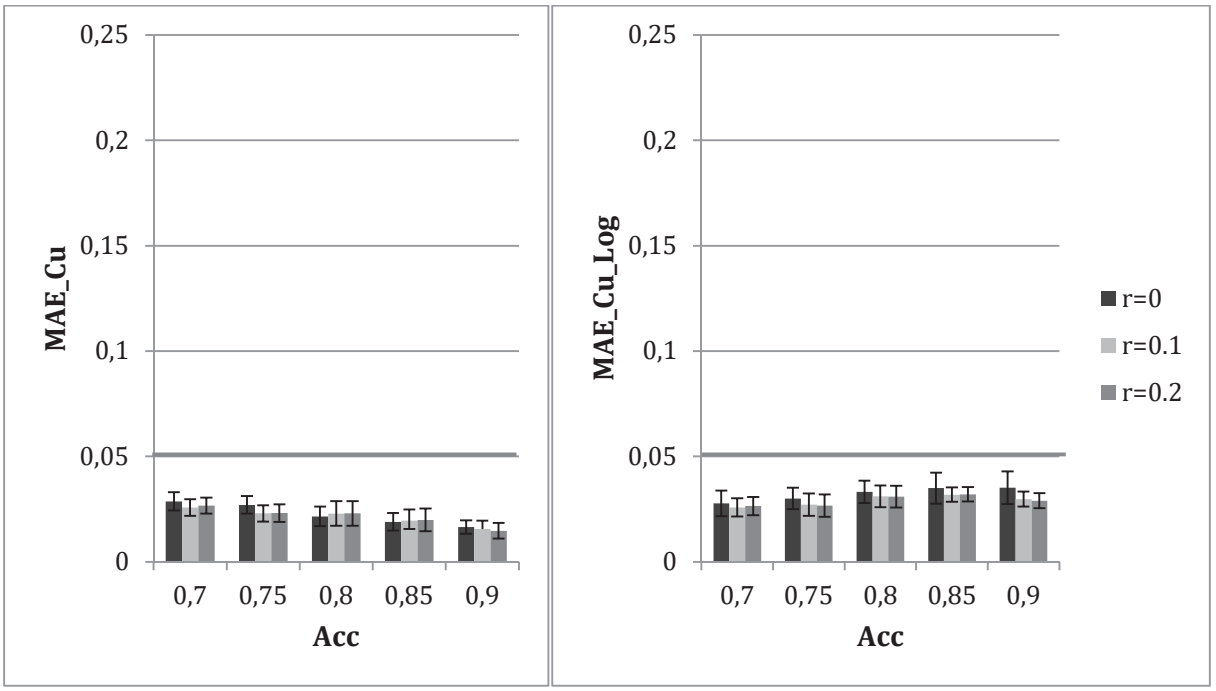


Abbildung 35: Ergebnisse für die **LDA** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) nimmt der MAE_Cu minimal ab. Es ist keine Abhängigkeit von der Korrelation zu erkennen. Bereits vor der Kalibrierung gibt die LDA sehr gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus. (Alle Werte befinden sich unterhalb der roten Linie). Der MAE_Cu_Log bleibt ebenfalls, unabhängig von der Korrelation und der Korrektklassifizierungsrate (Acc) immer niedrig (unterhalb der roten Linie). Folglich sind alle unkalibrierten und kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer nah an der wahren Wahrscheinlichkeit.

Bei der Betrachtung der Abhängigkeit der Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von der Datensatzgröße (Abbildung 36) fällt auf, dass sowohl der MAE_Cu als auch der MAE_Cu_Log mit steigender Anzahl an Objekten sinkt. Ab 1000 Molekülen sind sowohl die kalibrierten als auch die unkalibrierten Schätzwerte nah an der wahren Wahrscheinlichkeit (unterhalb der roten Linie).

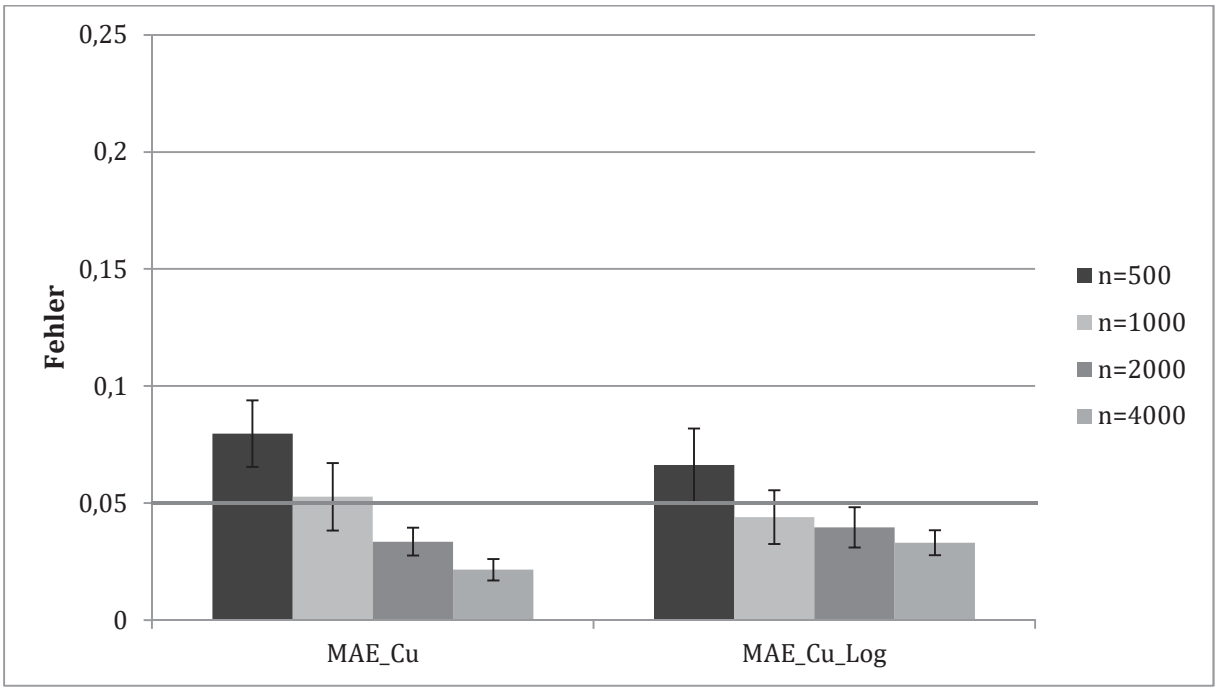


Abbildung 36: Ergebnisse für die **LDA** mit einer Korrektklassifizierungsrate (Acc)=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Sowohl für den MAE_Cu als auch für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt ab. Die Standardabweichung nimmt erwartungsgemäß mit zunehmender Datensatzgröße ebenfalls ab. Ab einer Datensatzgröße von 1000 Molekülen werden sowohl kalibriert als auch unkalibriert gute Wahrscheinlichkeitsschätzer erhalten (Werte befinden sich unterhalb oder auf Höhe der roten Linie).

4.1.3.1 Detailliertere Betrachtung der Kalibrierung des NBC

Aufgrund der Beobachtung, dass die logistische Regression keine gute Technik zur Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des NBC ist, sollte eine Alternative, welche bereits in der Literatur beschrieben ist, untersucht werden. Bei der Alternative handelt es sich um die isotonische Regression [99, 97, 98]. In Abbildung 37 sind die Ergebnisse einer Simulationsstudie mit dem NBC mit 2000 Objekten und 40 Variablen dargestellt. Die Korrektklassifizierungsrate (Acc) wurde auf 0.8 eingestellt. Berechnet wurden der MAE_Cu, der MAE_Cu_Log und der MAE_Cu_Iso. In Abbildung 37 wurden die Mittelwerte aus zehn wiederholten Versuchen aufgetragen sowie deren Standardabweichung. Zur Validierung wurde eine 50*50% LMO-CV verwendet. Die übrigen Fehlermaße sind im Anhang (Kapitel 8.2.4) zu finden. Es wurden sowohl unkorrelierte als auch korrelierte Variablen verwendet.

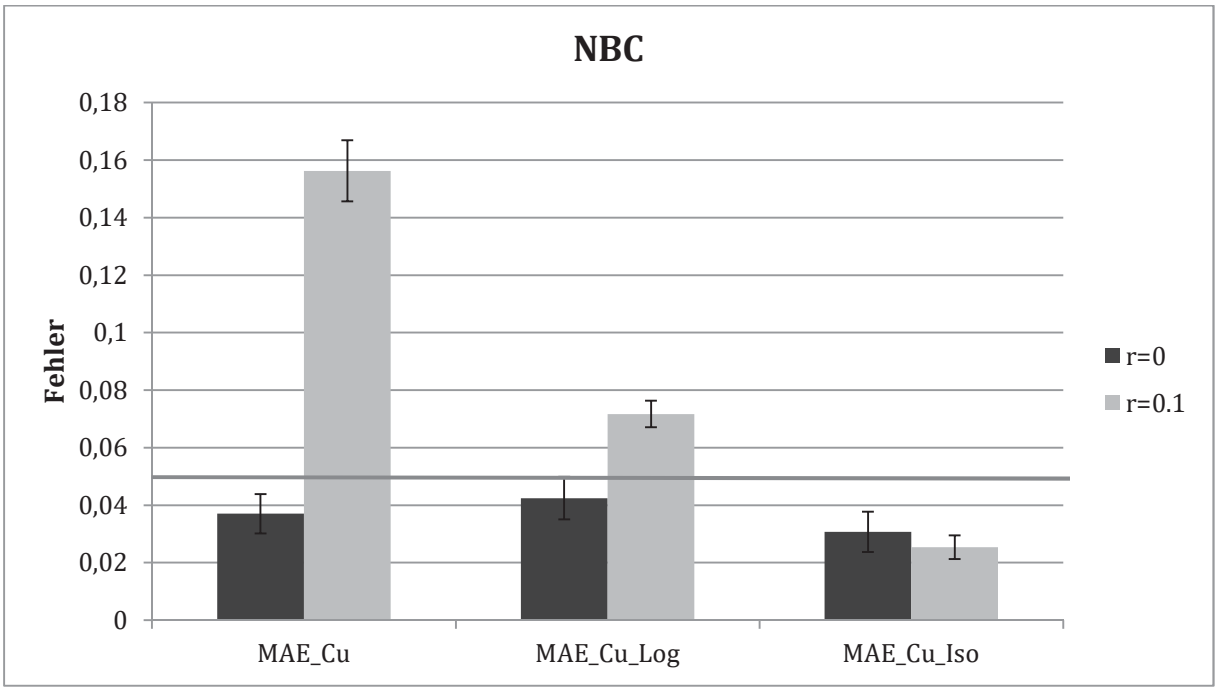


Abbildung 37: MAE_Cu, MAE_Cu_Log und MAE_Cu_Iso Mittelwerte der Klassifikationstechnik NBC aus zehn Versuchen sowie deren Standardabweichung. Die rote Linie kennzeichnet die „Nähe" zur wahren Wahrscheinlichkeit. Gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sollten kleiner als diese Linie sein. Die unkalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des NBC bei unkorrelierten Daten sind bereits sehr gut, Kalibrierung führt zu keiner Verbesserung. Bei korrelierten Variablen hingegen verschlechtern sich die Schätzwerte und nur die isotonische Regression kann diese gut rekalibrieren.

Es ist zu erkennen, dass der NBC bereits ohne anschließende Kalibrierung gute Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorbringt, sofern die Daten unkorreliert vorliegen. In diesem Fall führt weder die logistische noch die isotonische Regression zu einer Verbesserung. Wenn die Variablen allerdings korreliert sind, verschlechtert sich die Leistung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer stark. Die logistische Regression führt zu einer Verbesserung, allerdings sind die Schätzer immer noch nicht gut. Lediglich die Rekalibrierung mit der isotonischen Regression bringt Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervor, welche sich nah an der wahren Wahrscheinlichkeit befinden (unterhalb der roten Linie).

Mit dem Siegeszug der SVM und des RF um das Jahr 2000 herum hat der NBC an Bedeutung verloren und wird im chemieinformatischen Bereich nur noch wenig angewendet.

(Ende des Unterkapitels 4.1.3.1)

Die Ergebnisse für die PLSDA sind im Anhang (Kapitel 9.1.3) graphisch dargestellt, da sie vergleichbar sind mit den Ergebnissen der SPLS, welche im nächsten Abschnitt exemplarisch vorgestellt werden.

An dieser Stelle folgen die Ergebnisse der Regressionstechniken. Die Techniken RFR und SVR weisen nahezu dieselben Ergebnisse auf wie die bereits beschriebenen Klassifikationsmethoden (siehe Anhang Kapitel 9.1.3). Aus diesem Grund werden diese hier nicht näher beschrieben. Für die Auswertung der übrigen Regressionstechniken gibt es zwei verschiedene Varianten, welche anhand des Lassos exemplarisch gezeigt werden sollen. Für die anderen Techniken sind die Ergebnisse im Anhang aufgelistet (Kapitel 9.1.3). Die Vorhersagen der abhängigen Variablen der Techniken Lasso, Ridge, Elastic Net und SPLS können, im Gegensatz zu den Techniken RFR und SVR, auch über die Klassengrenzen 0 und 1 hinaus extrapolieren. Um diese Werte wieder auf eine Skala zwischen 0 und 1 zu bekommen, können diese einerseits skaliert werden oder Werte größer 1 oder kleiner 0 werden lediglich abgeschnitten. In der folgenden Abbildungen (Abbildung 38 und Abbildung 39) werden zunächst die skalierten Ergebnisse gezeigt.

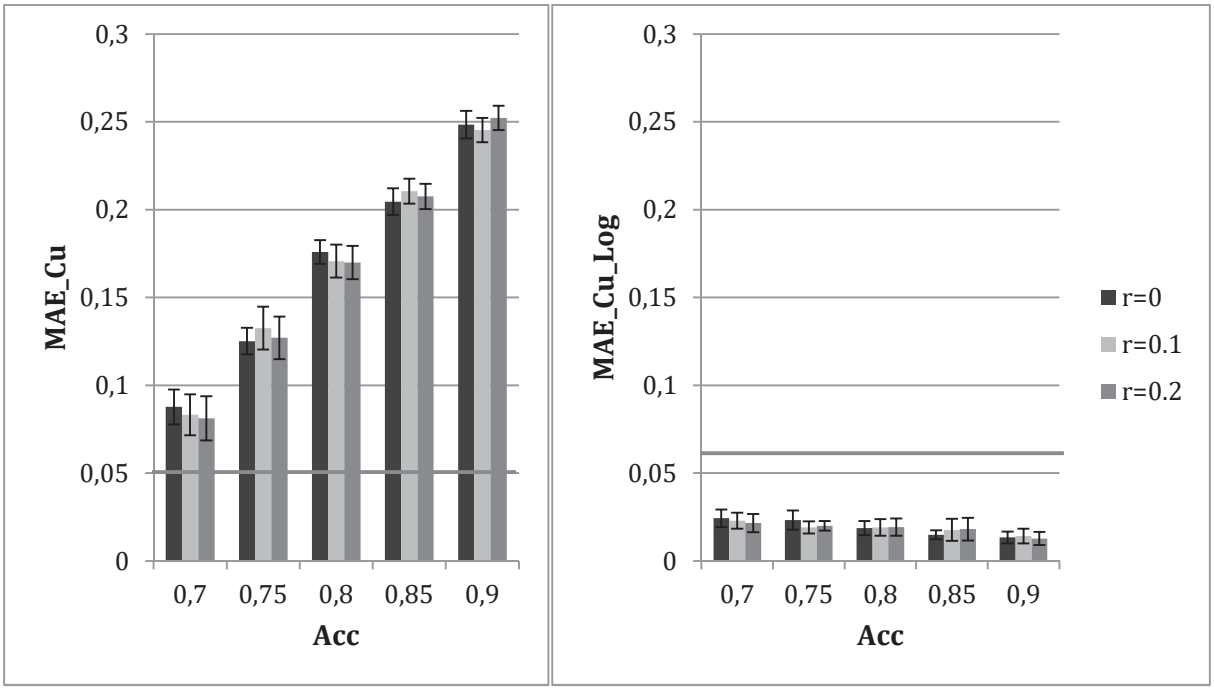


Abbildung 38: Ergebnisse für das **Lasso** (skaliert) mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc), unabhängig von der Korrelation, steigt der MAE_Cu an. Alle unkalibrierten Wahrscheinlichkeitsschätzer liegen weiter entfernt von der wahren Wahrscheinlichkeit. Im Gegensatz zum MAE_Cu bleibt der MAE_Cu_Log immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

Bei den unkalibrierten Wahrscheinlichkeitsschätzern ist zu erkennen, dass der MAE_Cu mit steigender Korrektklassifizierungsrate (Acc), unabhängig von der Korrelation, ansteigt. Darüber hinaus liegen alle unkalibrierten Wahrscheinlichkeitsschätzer weiter entfernt von der wahren Wahrscheinlichkeit. Alle kalibrierten Wahrscheinlichkeitsschätzer hingegen befinden sich relativ nah an der wahren Wahrscheinlichkeit. Der MAE_Cu_Log ist, unabhängig von der Korrektklassifizierungsrate (Acc) und der Korrelation, sehr niedrig. Abbildung 39 zeigt, dass die Datensatzgröße ausschließlich den MAE_Cu_Log beeinflusst, dieser sinkt mit steigender Anzahl an Objekten. Lediglich die Standardabweichung nimmt erwartungsgemäß bei beiden Fehlern ab.

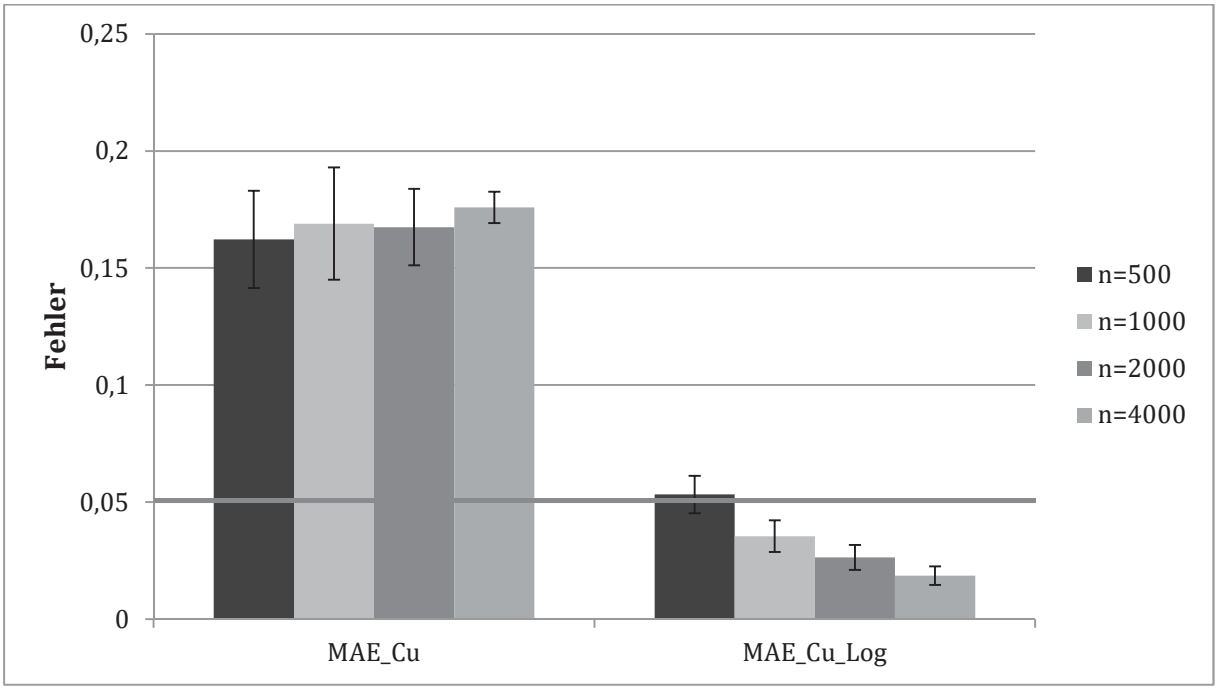


Abbildung 39: Ergebnisse für das **Lasso** (skaliert) mit einer Korrektklassifizierungsrate (Acc)=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab. Nur nach Kalibrierung werden gut kalibrierte Wahrscheinlichkeitsschätzer hervorgebracht (Werte befinden sich unterhalb der roten Linie).

Im Anschluss werden im direkten Vergleich mit den „skalierten" Ergebnissen, diejenigen der „abgeschnittenen" Variante betrachtet. Die Ergebnisse unterscheiden sich lediglich in der absoluten Höhe des MAE_Cu. Der MAE_Cu ist, im Verhältnis zu dem Wert der skalierten Ergebnisse, deutlich geringer. Wenn die Korrektklassifizierungsrate (Acc) 0.7 oder 0.75 beträgt, liegt der MAE_Cu unterhalb der roten Linie. Selbiges gilt für den Ein-

fluss der Datensatzgröße, außer der absoluten Höhe des MAE_Cu, sind keine Unterschiede erkennbar (siehe Anhang Kapitel 9.1.3).

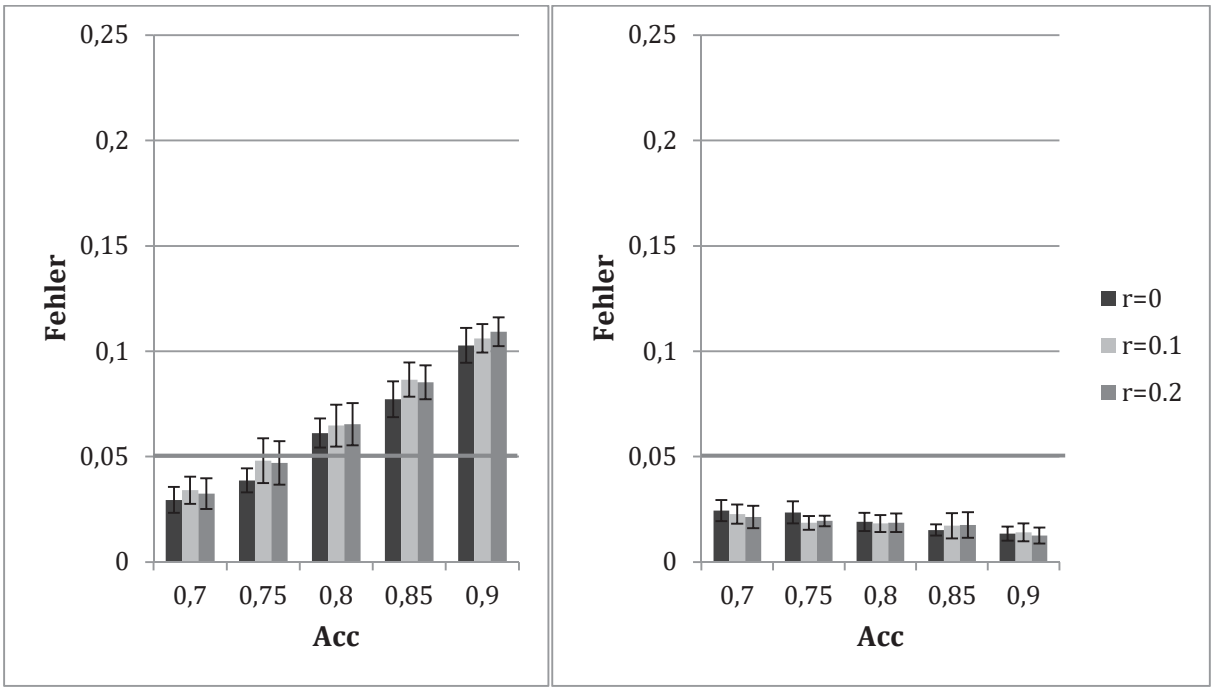


Abbildung 40: Ergebnisse für das **Lasso** (abgeschnitten) mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Im Vergleich zu den skalierten Ergebnissen unterscheiden sich die abgeschnittenen Ergebnisse lediglich durch die absolute Höhe des MAE_Cu. Dieser ist deutlich geringer. Bei einer Korrektklassifizierungsrate Acc=0.7 und 0.75 gibt das Lasso (abgeschnitten) nah an der wahren Wahrscheinlichkeit liegende Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie).

Nahezu dieselben Ergebnisse wie für das Lasso (abgeschnitten) werden auch für die SPLS (abgeschnitten) und für die PLSDA erhalten. Aus diesem Grund sind die Graphiken zu diesen Techniken im Anhang zu finden (Kapitel 9.1.3).

Zusammenfassend lässt sich sagen, dass der MAE_Cu der Techniken RF, RFR, SVM, SVR und KNN stark von der Korrektklassifizierungsrate (Acc) und der Korrelationsmatrix der Daten abhängt, der MAE_Cu_Log hingegen nicht. Aber dieser hängt dafür, im Gegensatz zum MAE_Cu, stärker von der Größe des Datensatzes ab. Sowohl die Größe des MAE_Cu, als auch die des MAE_Cu_Log der NN und der LDA ist leicht abhängig von der Datensatzgröße und unabhängig von der Korrektklassifizierungsrate (Acc) und der Korrelation. Der MAE_Cu der übrigen Regressionstechniken und der PLSDA ist lediglich abhängig von der Korrektklassifizierungsrate (Acc) und der MAE_Cu_Log von der Größe

des Datensatzes. Insgesamt profitieren alle Techniken, mit Ausnahme der LDA, den NN und dem NBC, von einer Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer mittels logistischer Regression. Der NBC profitiert von einer isotonischen Regression. Die LDA und die NN bringen bereits vorher gute Schätzer hervor, wie auch der RF und der KNN, wenn korrelierte Daten vorliegen. Bei den Regressionstechniken führt die abgeschnittene Variante zu einem niedrigeren MAE_Cu.

4.1.4 Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels realer Datensätze

In dieser Studie wurde analysiert, ob in den realen Datensatzbeispielen dieselben Beobachtungen gemacht werden können, wie in den Simulationsstudien. Zu diesem Zweck wurden die im Methodenteil bereits beschriebenen Datensätze mit variierender Korrektklassifizierungsrate (Acc) und Datensatzgröße verwendet. Zur Evaluierung wurde wie bereits im Methodenteil beschrieben (Kapitel 3.1.5) eine 50*50% LMO-CV verwendet. Jeder Versuch wurde einmal (bei allen Techniken gleicher random seed) durchgeführt. Zusätzlich wurden unterschiedliche Deskriptoren verwendet, um zu analysieren, ob die Auswahl eines unterschiedlichen Deskriptors die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen beeinflusst. Die verwendeten Fehlermaße waren erneut der MAE_Cu und der MAE_Cu_Log.

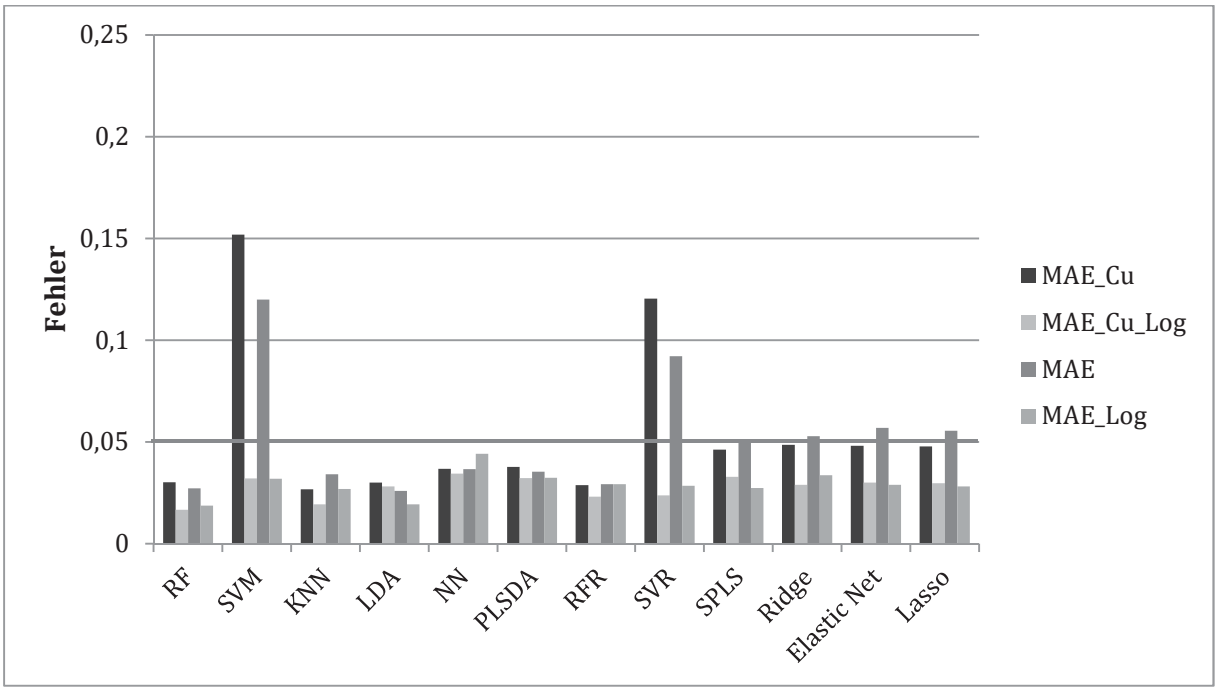


Abbildung 41: MAE_Cu und MAE_Cu_Log, sowie MAE und MAE_Cu, der auf der x-Achse aufgeführten Techniken basierend auf dem Ames Datensatz mit 6512 Molekülen und einer mittleren Korrektklassifizierungsrate (Acc) über alle Techniken von 0.75. Alle Techniken, mit Ausnahme der SVM und der SVR, bringen bereits vor der Kalibrierung gute Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervor (unter der roten Linie). Kalibrierung führt bei allen Techniken zumindest zu einer geringfügigen Verbesserung. Allerdings ist sie bei KNN, LDA, NN, PLSDA und RFR nicht notwendig. Ausgenommen der unkalibrierten Wahrscheinlichkeitsschätzer der SVM und SVR verlaufen die beiden Fehlermaße (MAE und MAE_Cu) sehr ähnlich.

Der zuerst betrachtete Datensatz Ames (MACCS) (Abbildung 41) hat eine Größe von 6512 Molekülen und die über alle Techniken gemittelte Korrektklassifizierungsrate (Acc) beträgt 0.75. Bei allen verwendeten Datensätzen wird von einer komplexeren Korrelationsstruktur in den Daten ausgegangen. Am Beispiel des Ames Datensatzes soll zu Beginn noch einmal exemplarisch die Ähnlichkeit der Fehlermaße gezeigt werden und die Schwierigkeit bei Verwendung der Brier-Score zur Auswertung.

Für alle Techniken wird beobachtet, dass die Ergebnisse mit denen aus den Simulationen übereinstimmen. Kalibrierung führt in allen Fällen zu besseren Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern. Allerdings sind diese Verbesserungen bei KNN, LDA, NN und PLSDA nur marginal. Die SVM und die SVR bringen unkalibriert sehr schlechte Schätzer hervor, die übrigen Techniken ergeben bereits gute unkalibrierte Schätzer, welche allerdings durch Kalibrierung noch sichtbar verbessert wurden. Bei Betrachtung des Brier-Score (Abbildung 42) ist ausschließlich bei den Techniken SVM

und SVR, welche eine hohe Differenz zwischen kalibrierten und unkalibrierten Fehler aufweisen, ein größerer Unterschied erkennbar. Die übrigen Differenzen zwischen Brier und Brier_Log sind nur sehr gering. Der Grund hierfür ist, dass die Korrektklassifizierungsrate zu einem größeren Teil in den Brier-Score mit einfließt, welche sich zumeist kaum ändert. Aus diesem Grund wird in den nächsten Datensätzen wiederum der MAE_Cu und der MAE_Cu_Log verwendet.

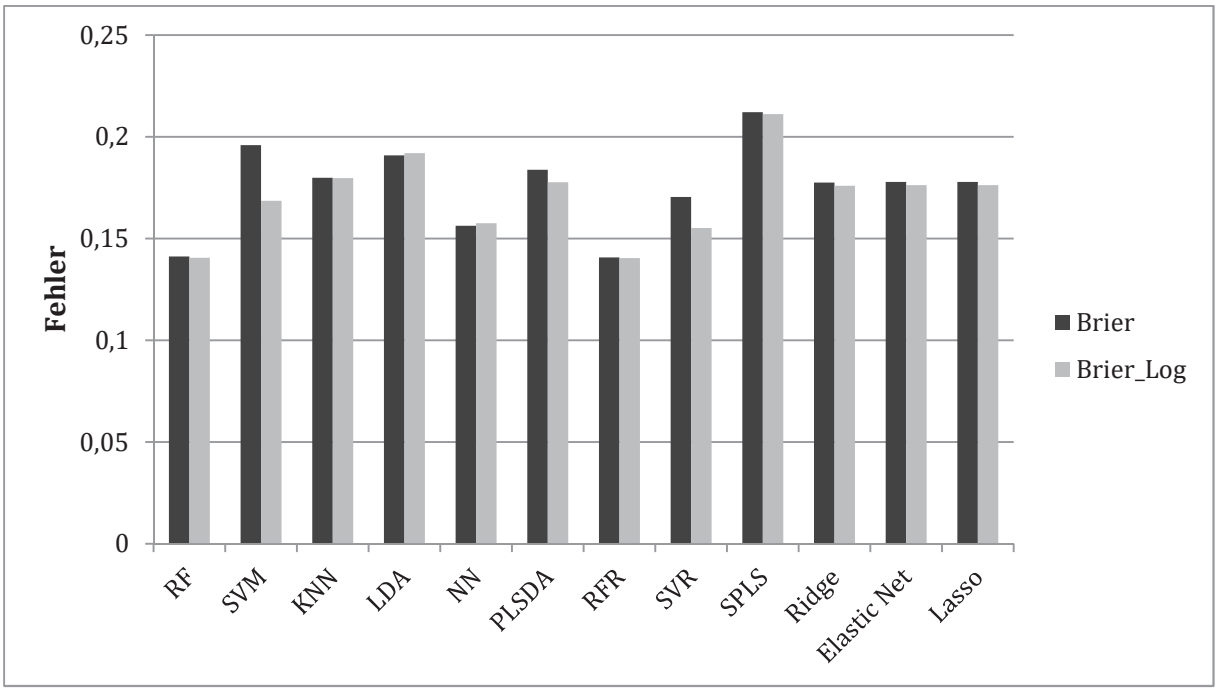


Abbildung 42: Brier und der Brier_Log der auf der x-Achse aufgeführten Techniken basierend auf dem Ames Datensatz mit 6512 Molekülen und einer mittleren Korrektklassifizierungsrate (Acc) über alle Techniken von 0.75. Nur bei den Techniken SVM und SVR, bei welchen die Differenz zwischen kalibrierten und unkalibrierten Fehler sehr hoch ist, ist auch ein etwas größerer Unterschied in der Brier-Score erkennbar.

Als nächstes wird zum einen, um die Ergebnisse des Ames Datensatzes noch einmal zu bestätigen, und zum anderen, um den Einfluss des ausgewählten Deskriptors auf den Fehler zu analysieren, der CYP1A2 Datensatz mit 7485 Molekülen und einer gesamt Korrektklassifizierungsrate (Acc) von 0.79 ausgewählt. Es wurden MACCS, MOE und E-State Deskriptoren verwendet. Wie bereits beim Ames Datensatz beobachtet, werden beim CYP1A2 Datensatz (Abbildung 43) ebenfalls die Ergebnisse der Simulationsstudien bestätigt. Nur bei Betrachtung des MAE_Cu ist ein Einfluss des Deskriptors bei den Techniken PLSDA, SVR, Ridge, Elastic Net und Lasso erkennbar. Zwischen den beiden

verglichenen Fehlermaßen sind, ausgenommen der unkalibrierten Fehlerschätzer der SVM und SVR, nur geringfügige Unterschiede erkennbar.

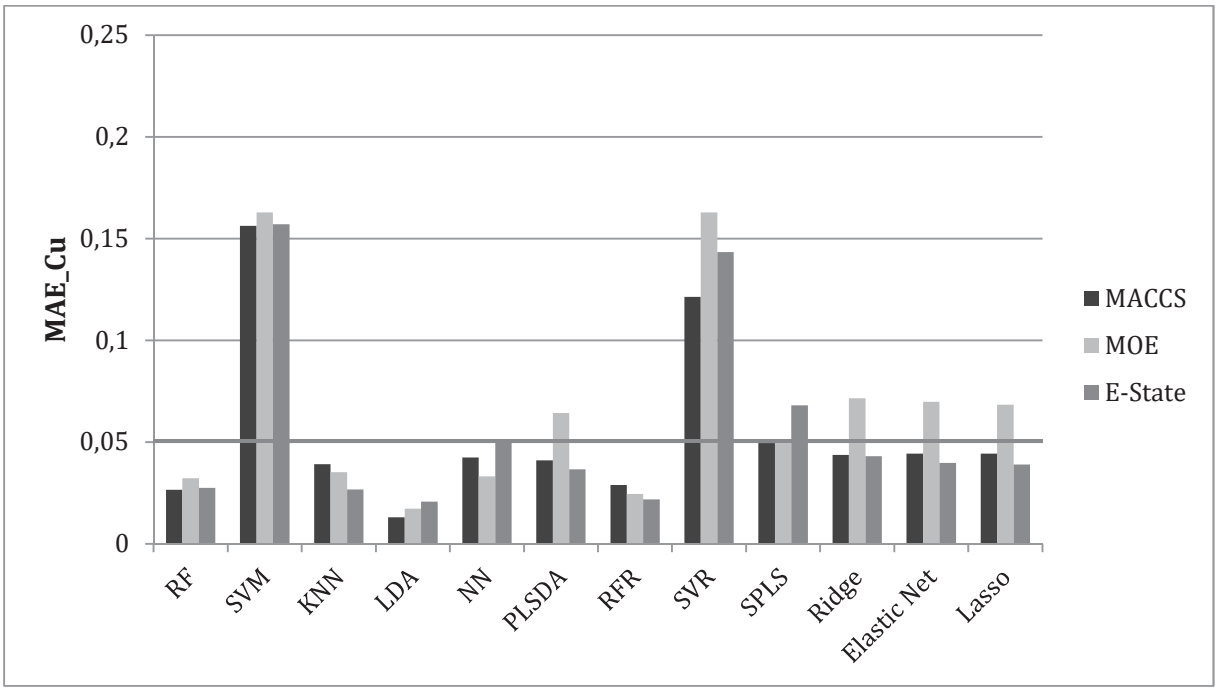


Abbildung 43: MAE_Cu der auf der x-Achse aufgeführten Techniken basierend auf dem CYP1A2 Datensatz mit 7485 Molekülen und einer mittleren Korrektklassifizierungsrate (Acc) über alle Techniken von 0.79. Alle Techniken, mit Ausnahme der SVM und der SVR, bringen bereits vor der Kalibrierung gute Wahrscheinlichkeitsschätzer hervor (unter der roten Linie). Die besten Schätzer stammen von den Techniken RF, KNN, LDA, NN und RFR. Bei den Techniken PLSDA, SVR, Ridge, Elastic Net und Lasso ist ein Einfluss des Deskriptors erkennbar. Bei Verwendung des MOE Deskriptors vergrößert sich der Fehler. Bei den übrigen Techniken sind die Unterschiede bei Verwendung verschiedener Deskriptoren sehr gering.

Dieser vergrößert sich bei Verwendung des MOE Deskriptors. Der MAE_Cu_Log hingegen weist nur sehr geringe Unterschiede bei Verwendung verschiedener Deskriptoren auf. Die schlechteren Ergebnisse der MOE Deskriptoren auf Basis des CYP Datensatzes lassen sich an dieser Stelle nicht erklären, da sich die Korrektklassifizierungsrate nicht ändert.

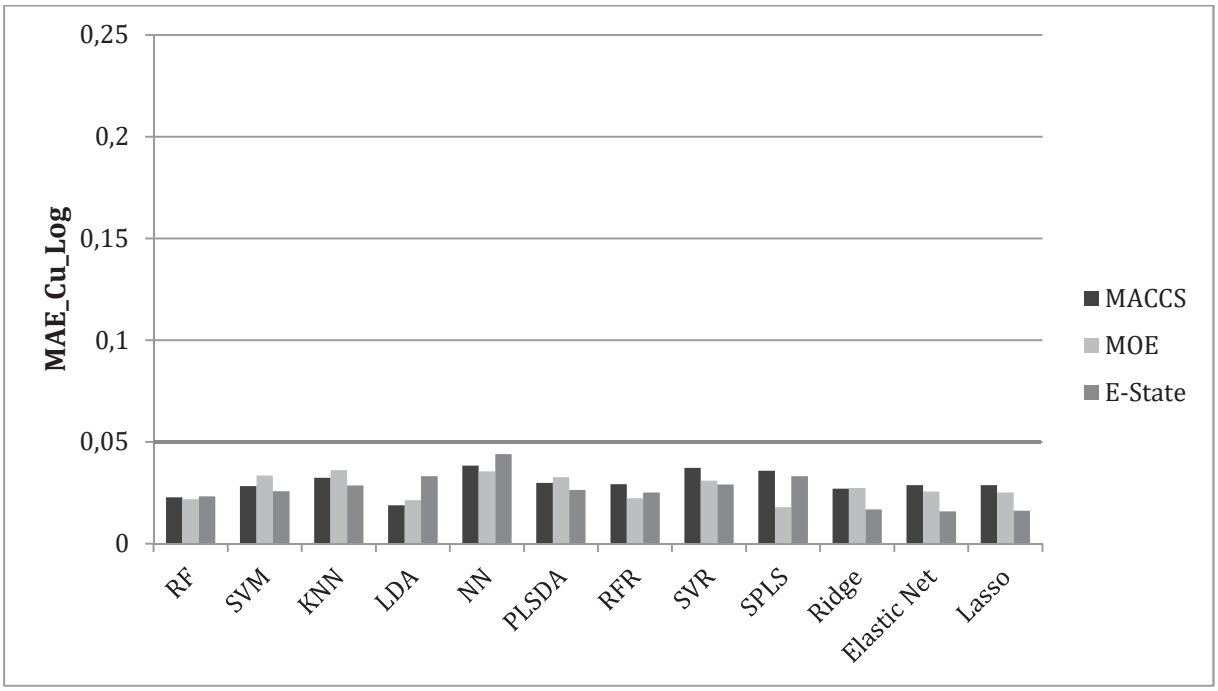


Abbildung 44: MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem CYP1A2 Datensatz mit 7485 Molekülen und einer mittleren Korrektklassifizierungsrate (Acc) über alle Techniken von 0.79. Alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit (unterhalb der roten Linie). Kalibrierung ist allerdings bei RF, KNN, LDA, NN, PLSDA und RFR nicht notwendig. Die Unterschiede zwischen den verschiedenen Deskriptoren sind nur sehr gering.

Um den Einfluss der Korrektklassifizierungsrate (Acc) auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzung zu analysieren, wurde der Factor Xa (435 Moleküle) Datensatz mit einer mittleren Korrektklassifizierungsrate (Acc) von 0.90 verwendet. Außerdem wurden zwei verschiedene Deskriptoren (MACCS und MOE) berechnet, um erneut den Einfluss dieser zu betrachten. Auf Abbildung 44 ist zu sehen, dass die Deskriptorauswahl sich zwar auf die Höhe des Fehlers auswirkt, allerdings dominiert kein Deskriptor den anderen. Des Weiteren ist zu erkennen, dass diejenigen Techniken, welche in den Simulationsstudien eine Abhängigkeit zwischen Korrektklassifizierungsrate (Acc) und MAE_Cu zeigten, diese auch auf dem betrachteten Datensatz zeigen. Der MAE_Cu der Techniken RF, KNN, SVM, PLSDA, SPLS, Ridge, Elastic Net und Lasso ist, verglichen mit dem Ames und dem CYP1A2 Datensatz, leicht angestiegen. Alle Techniken mit Ausnahme der LDA profitieren von der Kalibrierung. Für die Techniken KNN, LDA und NN wäre diese jedoch nicht notwendig gewesen, denn sie brachten bereits davor gute Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervor (unterhalb der roten

Linie). In diesem Kapitel wurden drei Datensätze exemplarisch gezeigt, die übrigen Realdatensatzbeispiele sind im Anhang aufgeführt und beschrieben (Kapitel 9.1.4).

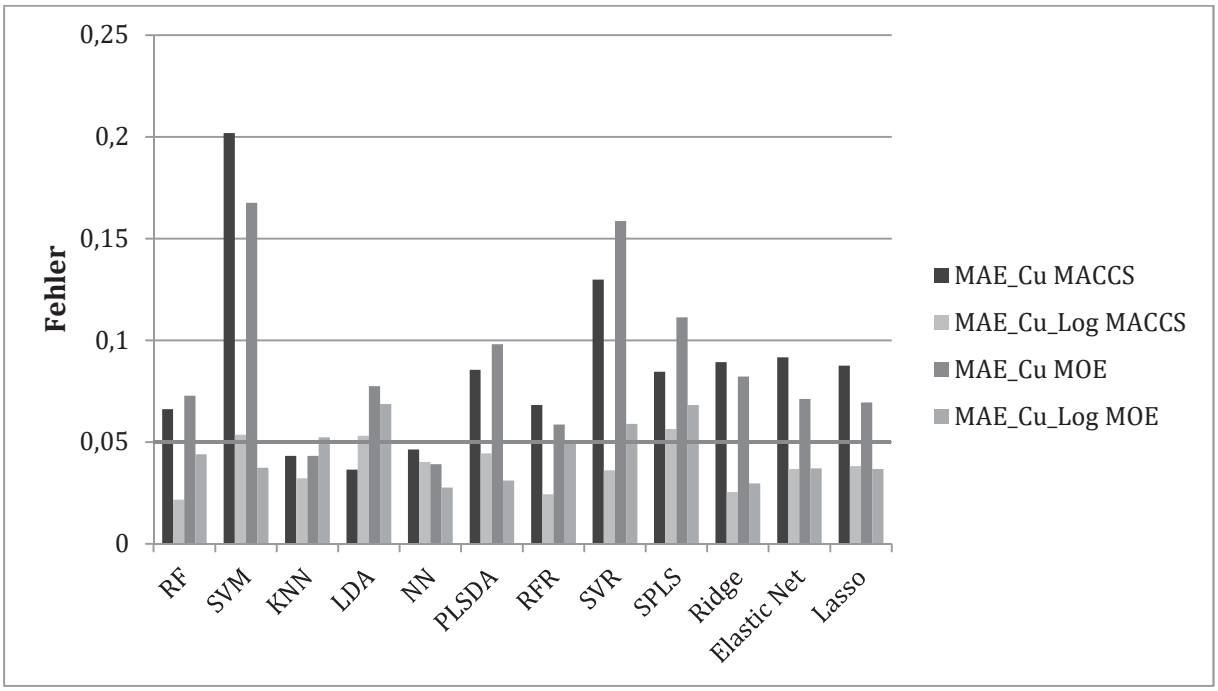


Abbildung 45: MAE_Cu und MAE_Cu Log der auf der x-Achse aufgeführten Techniken, berechnet mit MACCS und MOE Deskriptoren, basierend auf dem Factor Xa Datensatz mit 435 Molekülen. Die mittlere Korrektklassifizierungsrate (Acc) über alle Techniken beträgt 0.90. Zwischen den beiden Deskriptoren sind Unterschiede erkennbar. Jedoch dominiert kein Deskriptor den anderen. Die Ergebnisse der Simulationsstudien bestätigen sich erneut, da in diesen bereits für die Methoden RF, KNN, SVM, PLSDA, SPLS, Ridge, Elastic Net und Lasso eine Abhängigkeit des unkalibrierten Fehlers von der Korrektklassifizierungsrate (Acc) zu erkennen war. Die Fehler liegen im Vergleich zum Ames und CYP1A2 Datensatz etwas höher. Für alle Techniken, ausgenommen KNN, LDA und NN, ist eine Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sinnvoll.

4.1.5 Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels Simulationsstudien

Der Aufbau dieser Simulationsstudien gleicht denen der Kapitel 4.1.2 und 4.1.3. Zur Evaluierung wurde wie bereits im Methodenteil beschrieben (Kapitel 3.1.5) eine 50*50% LMO-CV verwendet. Jeder Versuch wurde einmal (bei allen Techniken gleicher random seed) durchgeführt. Es wurden immer 4000 Objekte verwendet. Wie auch in Studie 4.1.2 wurde sowohl die Korrektklassifizierungsrate (Acc) (0.7, 0.8, 0.9), als auch die Korrelation (r=0, r=0.1, r=0.2) variiert. Für jede Technik wurde der MAE_Cu und der MAE_Cu_Log zunächst separat berechnet. Danach wurden alle Techniken zu einem Hetero-Ensemble kombiniert. Anschließend wurden vier verschiedene Arten von Fehlern be-

rechnet. Für die ersten beiden Fehlerarten wurden zunächst alle unkalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aller verwendeten Techniken gemittelt. Danach wurden sowohl der MAE_Cu, als auch der MAE_Cu_Log berechnet. In dieser Studie wird diese Art der Klassifikatorfusion als Ensemble A bezeichnet. Beim sogenannten Ensemble B wurde zunächst jede Technik einzeln rekalibriert und der Mittelwert gebildet. Anschließend wurde der MAE_Cu_Log berechnet. Aus den bisherigen Untersuchungen ging bereits hervor, dass die LDA und die NNs in der Regel bereits gute unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorbringen. Aus diesem Grund wurden bei dem Ensemble C alle Techniken mit Ausnahme der LDA und den NN einzeln rekalibriert und mit den unkalibrierten Schätzern dieser ausgenommenen Techniken gemittelt. Danach wurde der MAE_Cu_Log berechnet. Das Ziel dieser Studie ist es zu analysieren, in wieweit sich die Bildung eines Hetero-Ensembles auf die Wahrscheinlichkeitsschätzung auswirkt und ob eines der drei Hetero-Ensemble „Typen“ (A, B, C) möglicherweise den anderen überlegen ist.

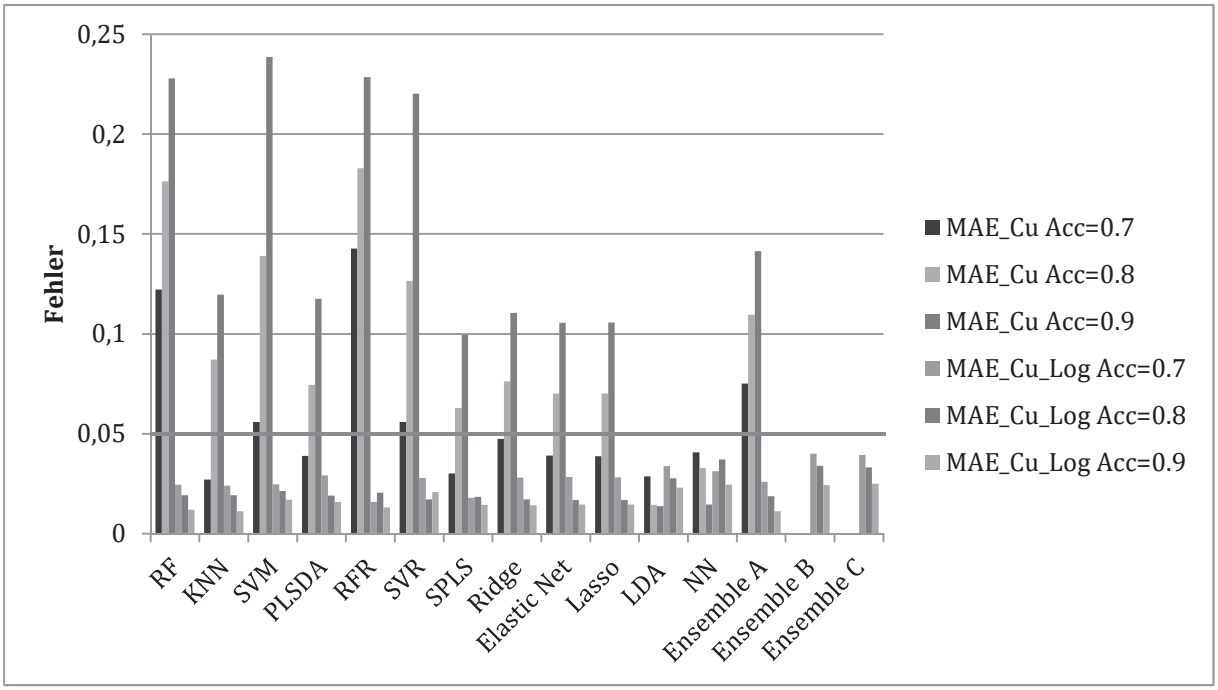


Abbildung 46: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf einem simulierten Datensatz mit 4000 Objekten, r=0 und einer variierenden Korrektklassifizierungsrate (Acc) von 0.7 bis 0.9. Kalibrierung führt bei allen Techniken, ausgenommen der LDA und den NNs zu einer Verbesserung. Alle kalibrierten Wahrscheinlichkeitsschätzer befinden sich nah an der wahren Wahrscheinlichkeit. Dies gilt ebenfalls für die Ensembles A, B, C. Bei diesen weist der MAE_Cu_Log des Typs A den niedrigsten Wert auf und ist somit am leistungsstärksten. Darüber hinaus kann beobachtet werden, dass der MAE_Cu mit steigender Korrektklassifizierungsrate (Acc) zunimmt und der MAE_Cu_Log abnimmt.

Abbildung 46 zeigt die Ergebnisse der Simulationsstudie für r=0. Es ist zu erkennen, dass sich die MAE_Cu bzw. MAE_Cu_Log-Werte der einzelnen Techniken mit denen aus den bisherigen Simulationsstudien decken. Kalibration führt bei allen Techniken, mit Ausnahme der LDA und den NN, zu einer Verbesserung. Mit steigender Korrektklassifizierungsrate (Acc) nimmt der MAE_Cu zu und der MAE_Cu_Log ab. Die Bildung eines Hetero-Ensembles führt im Fall des Ensembles A sowohl bei den MAE_Cu, als auch bei den MAE_Cu_Log-Werten, zu einer Verbesserung. Beim MAE_Cu_Log liegt der Fehler sogar auf Höhe des niedrigsten Einzelfehlers. Bei den Ensembles B und C liegen die Werte jedoch auf Höhe der höchsten Einzelwerte. Somit führt in diesem Beispiel die Mittelung der unkalibrierten Schätzwerte und die anschließende logistische Regression zu den besten Ergebnissen. Alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer liegen nah an der wahren Wahrscheinlichkeit liegen (unterhalb der roten Linie).

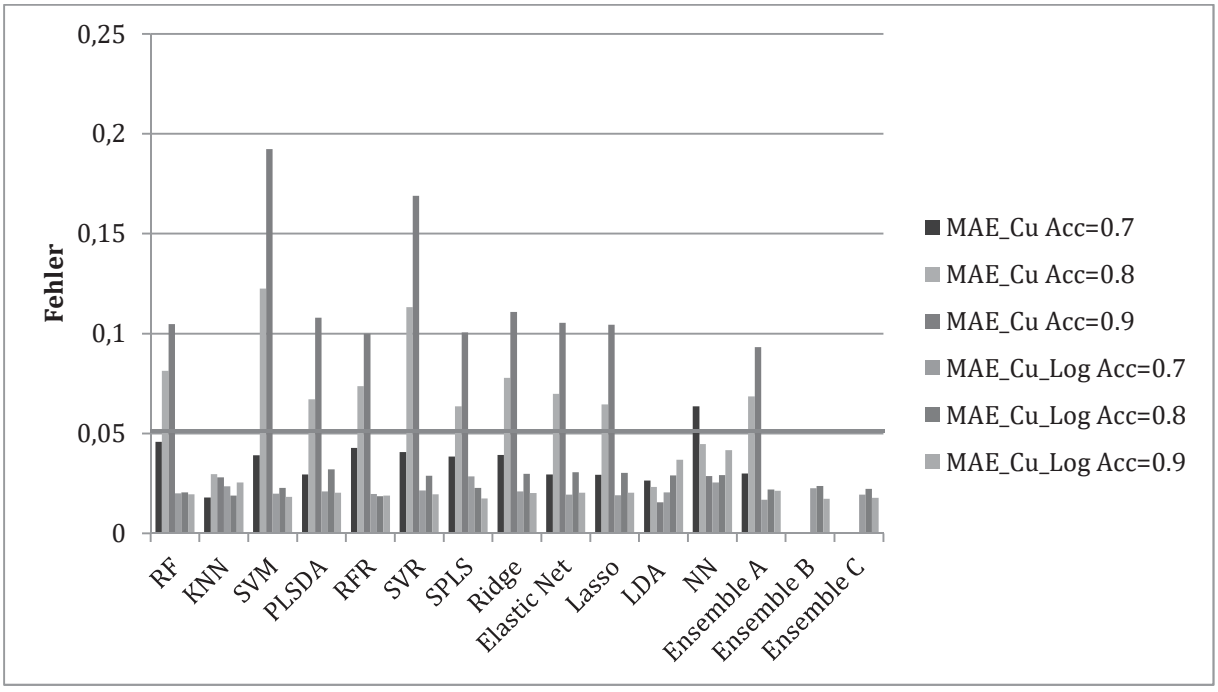


Abbildung 47: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf einem simulierten Datensatz mit 4000 Objekten, r=0.1 und einer variierenden Korrektklassifizierungsrate (Acc) von 0.7 bis 0.9. Kalibrierung führt bei allen Techniken, ausgenommen der LDA, KNN und NN (Acc=0.9), zu einer Verbesserung. Alle kalibrierten Wahrscheinlichkeitsschätzer befinden sich nah an der wahren Wahrscheinlichkeit. Dies gilt ebenfalls für die Ensembles A, B, C. Bei diesen sind die MAE_Cu_Log-Werte aller Typen sehr niedrig. Es kann somit kein „überlegener" Ensemble-Typ benannt werden. Darüber hinaus kann beobachtet werden, dass der MAE_Cu mit steigender Korrektklassifizierungsrate (Acc) zunimmt.

Abbildung 47 stellt die Ergebnisse für r=0.1 dar. Kalibration führt auch in diesem Fall, wie bereits aus den vorherigen Simulationen hervorging, in allen Fällen, mit Ausnahme der LDA, KNN und NN (Acc=0.9), zu einer Verbesserung. Alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer liegen nah an der wahren Wahrscheinlichkeit. Die Bildung von Hetero-Ensembles wirkt sich bei Ensemble A, B und C positiv aus. Der MAE_Cu-Wert des Ensembles A befindet sich für alle drei Korrektklassifizierungsrate (Acc) Werte ungefähr auf Höhe des niedrigsten Einzelwertes. Dasselbe gilt auch für den MAE_Cu_Log-Wert des Ensembles A, B und C. Folglich sind alle Ensemble-Typen gleich leistungsstark.

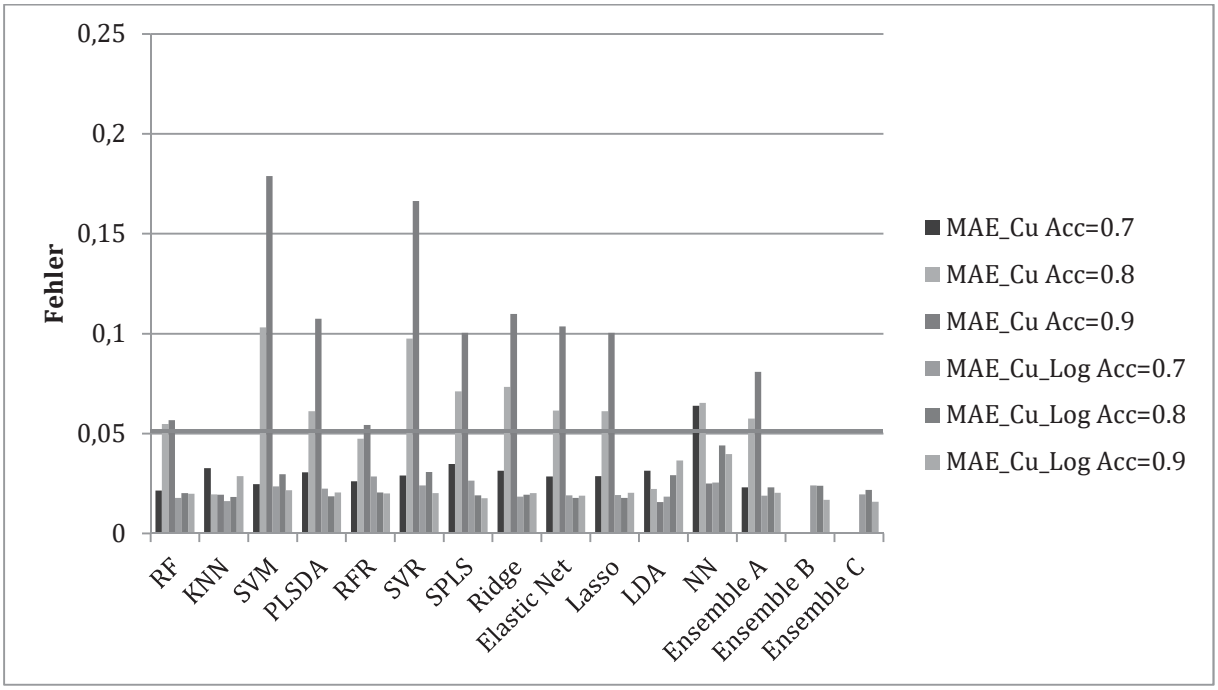


Abbildung 48: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf einem simulierten Datensatz mit 4000 Objekten, r=0.2 und einer variierenden Korrektklassifizierungsrate (Acc) von 0.7 bis 0.9. Kalibrierung führt bei allen Techniken, ausgenommen der LDA und KNN, zu einer Verbesserung. Alle kalibrierten Wahrscheinlichkeitsschätzer befinden sich nah an der wahren Wahrscheinlichkeit. Dies gilt ebenfalls für die Ensembles A, B, C. Bei diesen sind die MAE_Cu_Log-Werte aller Typen sehr niedrig. Es kann somit kein „überlegener" Ensemble-Typ benannt werden. Darüber hinaus kann beobachtet werden, dass der MAE_Cu mit steigender Korrektklassifizierungsrate (Acc) zunimmt.

Für r=0.2 (Abbildung 48) ähneln die Ergebnisse sehr stark denen von r=0.1. Denn die Bildung von Hetero-Ensembles führt auch hierbei bei allen drei Typen dazu, dass sowohl der MAE_Cu, als auch der MAE_Cu_Log ungefähr auf Höhe des niedrigsten Einzelwertes gehalten werden. Dies gilt selbst wenn, wie hier bei den NN bei einer Korrektklassifizie-

rungsrate (Acc=07 und 0.8), Ausreißer auftreten. Somit sind ebenfalls alle Ensemble-Typen gleich leistungsstark.

Zusammenfassend lässt sich sagen, dass die Bildung von Hetero-Ensemblen in den korrelierten Beispielen den MAE_Cu bzw. den MAE_Cu_Log stabilisiert, bzw. ungefähr auf die Höhe des niedrigsten Einzelwertes absenkt. Dabei gibt es keine großen Unterschiede zwischen den Typen A, B und C. Obwohl die Ensemble-Bildung den MAE_Cu senkt, ist dieser trotzdem meist höher als der kalibrierte Fehler. (Die Rohdaten, welche den Graphiken zu Grunde liegen, sind im Anhang in Kapitel 9.1.5 zu finden).

4.1.6 Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels realer Datensätze

In dieser Studie wurde analysiert, ob in den realen Datensatzbeispielen dieselben Beobachtungen gemacht werden können, wie in den Simulationsstudien. Zu diesem Zweck wurden die im Methodenteil bereits beschriebenen Datensätze mit variierender Korrektklassifizierungsrate (Acc) und Datensatzgröße verwendet. Zur Evaluierung wurde wie bereits im Methodenteil beschrieben (Kapitel 3.1.5) eine 50*50% LMO-CV verwendet. Jeder Versuch wurde einmal (bei allen Techniken gleicher random seed) durchgeführt. Es wurde ein Deskriptor (MACCS) verwendet, da aus vorherigen Studien (4.1.4) hervorging, dass die Verwendung unterschiedlicher Deskriptoren die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der einzelnen Techniken nur wenig beeinflusst. An dieser Stelle wird exemplarisch jeweils ein Datensatz mit niedriger, mittlerer und hoher Korrektklassifizierungsrate (Acc) gezeigt.

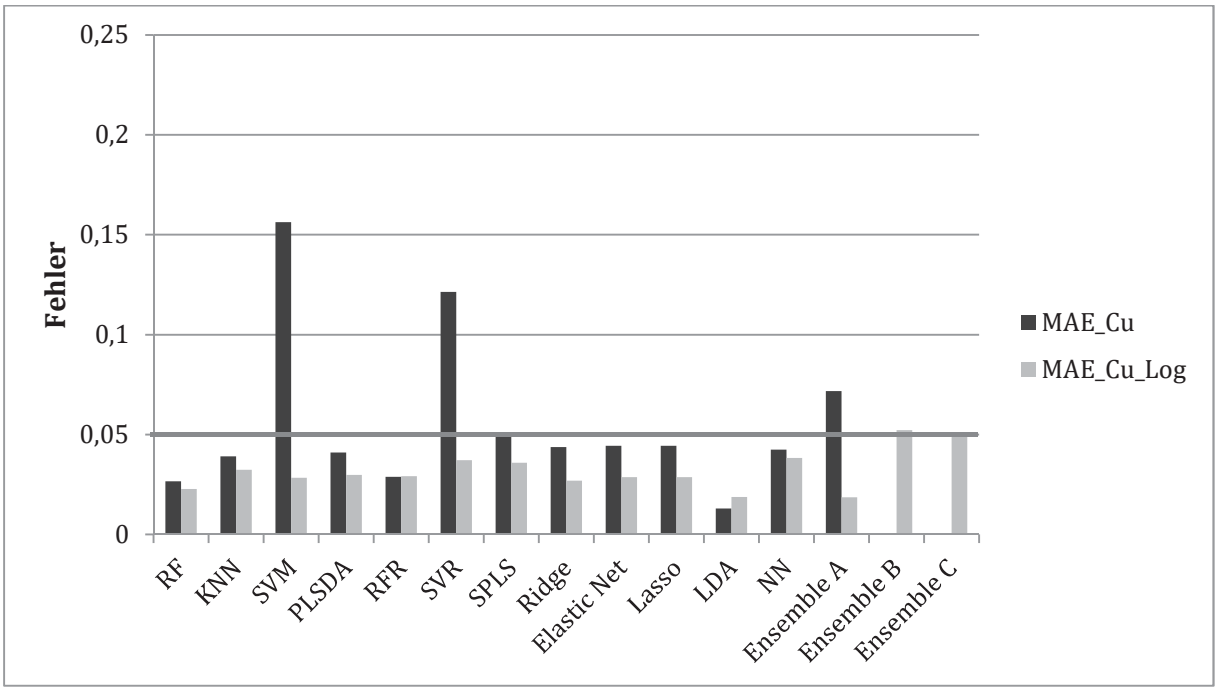


Abbildung 49: MAE_Cu und MAE_Cu_Log des CYP1A2 Datensatzes bestehend aus 7485 Molekülen (Acc (gemittelt)=0.79). Zum besseren Vergleich wurden die Ergebnisse der einzelnen Techniken noch einmal mit aufgeführt. Es wird beobachtet, dass der Ensemble Typ A deutlich bessere Ergebnisse hervorbringt. Die Typen B und C liefern schlechtere Klassenzugehörigkeits-Schätzer als die einzelnen Techniken. Allerdings befinden sich alle Schätzer noch unterhalb bzw. auf der Höhe der roten Linie und sind somit nah an der wahren Wahrscheinlichkeit.

Begonnen wird mit dem Datensatz mit mittlerer Korrektklassifizierungsrate (Acc) (Abbildung 49). Es ist zu erkennen, dass der Ensemble Typ A den anderen beiden Typen deutlich überlegen ist. Der MAE_Cu_Log-Wert des Typs A nimmt, verglichen mit den Einzelwerten der übrigen Techniken, den niedrigsten Wert an. Die MAE_Cu_Log-Werte der Typen B und C sind höher als die der Einzelwerte. Trotzdem liegen sie unterhalb der roten Linie und damit noch nah an der wahren Wahrscheinlichkeit.

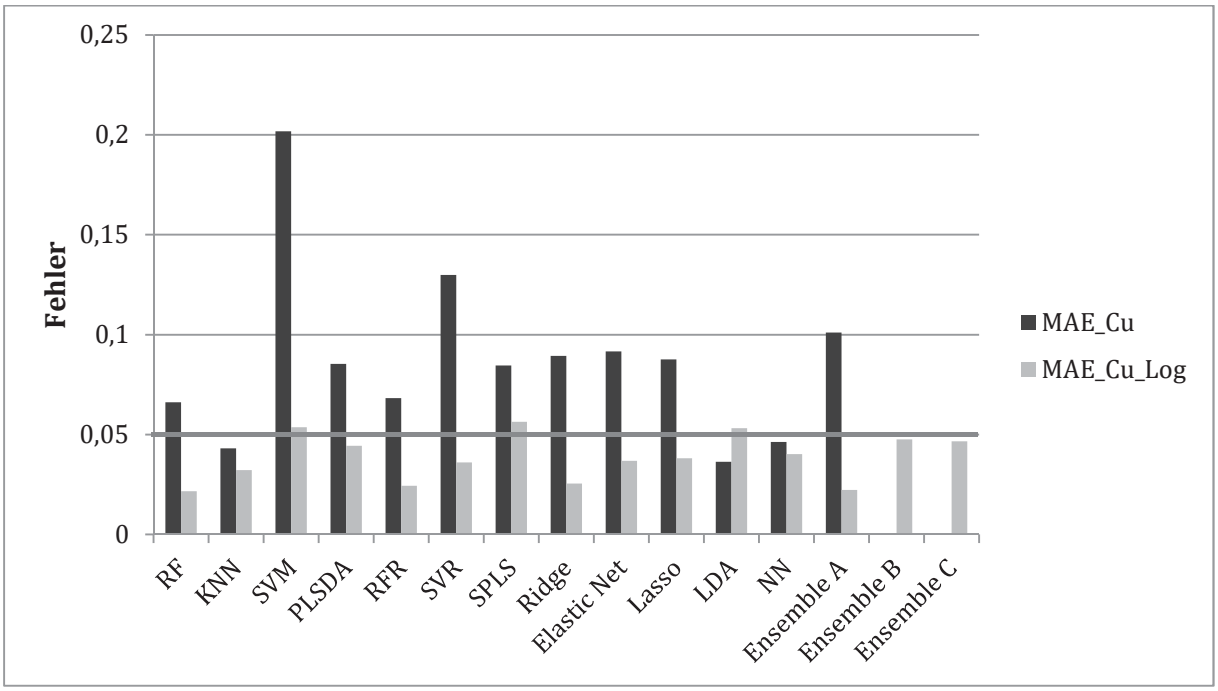


Abbildung 50: MAE_Cu und MAE_Cu_Log des Faktor Xa Datensatzes bestehend aus 435 Molekülen (Acc (gemittelt)=0.9). Zum besseren Vergleich wurden die Ergebnisse der einzelnen Techniken noch einmal mit aufgeführt. Analog zur vorherigen Abbildung wird beobachtet, dass der Ensemble Typ A deutlich bessere Ergebnisse hervorbringt. Alle Schätzer befinden sich unterhalb der Höhe der roten Linie und sind somit nah an der wahren Wahrscheinlichkeit.

Die Ergebnisse des Faktor Xa-Datensatzes (Abbildung 50) mit einer hohen Korrektklassifizierungsrate (Acc) decken sich mit denen des CYP1A2-Datensatzes. Der MAE_Cu_Log-Wert des Ensemble Typs A liegt auf gleicher Höhe mit dem niedrigsten Einzelwert und ist folglich den anderen beiden Typen überlegen. Dennoch liegen die MAE_Cu_Log-Werte der Typen B und C noch immer unterhalb der roten Linie. Beim Liver-Datensatz, welcher eine schlechte Korrektklassifizierungsrate (Acc) aufweist, unterscheiden sich die Ergebnisse. Der Ensemble Typ B weist den niedrigsten MAE_Cu_Log-Wert auf, allerdings gibt es zwischen den drei Typen keine großen Unterschiede. Bei Betrachtung des MAE_Cu fällt auf, dass sich der Wert des Ensemble Typs A auf gleicher Höhe mit dem niedrigsten Einzelwert befindet. Zusammenfassend lässt sich sagen, dass der Ensemble Typ A die niedrigsten MAE_Cu_Log-Werte hervorbringt und somit bevorzugt verwendet werden sollte. (Die Rohdaten, welche den Graphiken zu Grunde liegen, sind im Anhang in Kapitel 9.1.6 zu finden).

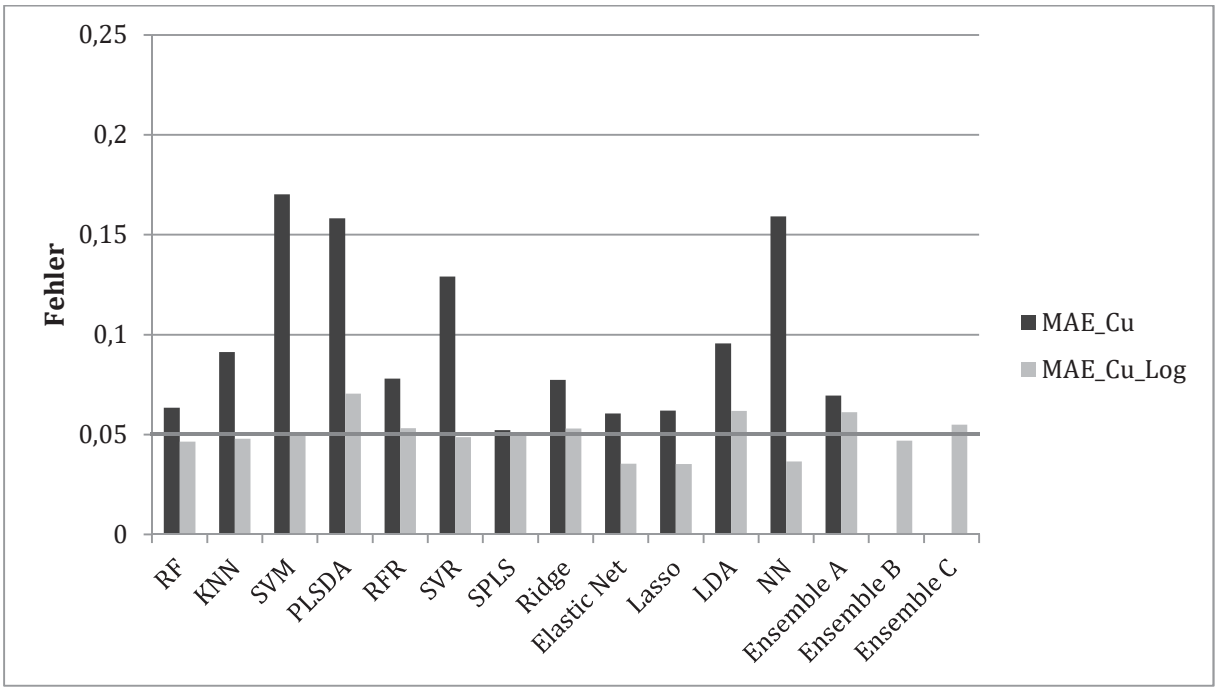


Abbildung 51: MAE_Cu und MAE_Cu_Log des Liver-Datensatzes bestehend aus 951 Molekülen (Acc (gemittelt)=0.66). Zum besseren Vergleich wurden die Ergebnisse der einzelnen Techniken noch einmal mit aufgeführt. Im Gegensatz zur vorherigen Abbildung wird beobachtet, dass der Ensemble Typ B bessere Ergebnisse hervorbringt. Allerdings unterscheiden sich drei Typen nicht besonders voneinander. Zusätzlich kann beobachtet werden, dass der MAE_Cu des Ensemble Typs A ungefähr auf gleicher Höhe liegt, wie der niedrigste Einzelwert der Techniken.

4.2 Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern versus CP

In dieser Studie wurde untersucht, wie viele Objekte durch den CP aus dem R-Paket conformal als uninformativ vorhergesagt werden, um ein bestimmtes Signifikanzlevel (1-Signifikanzlevel) zu halten. Verglichen damit, wurde untersucht, wie viele Objekte durch die Methode bestimmte Segmente sortierter Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer herauszuschneiden (Reject-Option), als uninformativ vorhergesagt werden. Insgesamt wurden drei Datensätze untersucht, welche alle eine mittlere bis hohe Korrektklassifizierungsrate (Acc) aufweisen, da mehrere Signifikanzlevel analysiert werden sollten. Exemplarisch werden an dieser Stelle die Ergebnisse des Ames-Datensatzes gezeigt. Die Ergebnisse für die übrigen Datensätze sind in Kapitel 9.2 im Anhang zu finden. Als Deskriptor wurden MACCS Deskriptoren und zur Validierung wurde eine 5-fache CV verwendet. Auch in dieser Studie wurde keine Optimierung von Hyperparametern vorgenommen. Aus Abbildung 52 geht hervor, dass beide Techniken

in der Lage sind, dass jeweils vorgegebene Signifikanzlevel zu halten, bzw. sogar noch darüber liegen. Allerdings muss an dieser Stelle berücksichtigt werden, dass die gesamt Korrektklassifizierungsrate (Acc) bei 0.8 liegt und es folglich von Anfang an leichter ist, das Signifikanzlevel zu halten. Der CP sagt im direkten Vergleich immer 5-10% mehr Objekte als uninformativ vorher, als die Technik, welche Segmente aus der Mitte entfernt. Diese Beobachtung wird auch bei den zwei weiteren Datensätzen gemacht, welche im Anhang (Kapitel 9.2) zu finden sind.

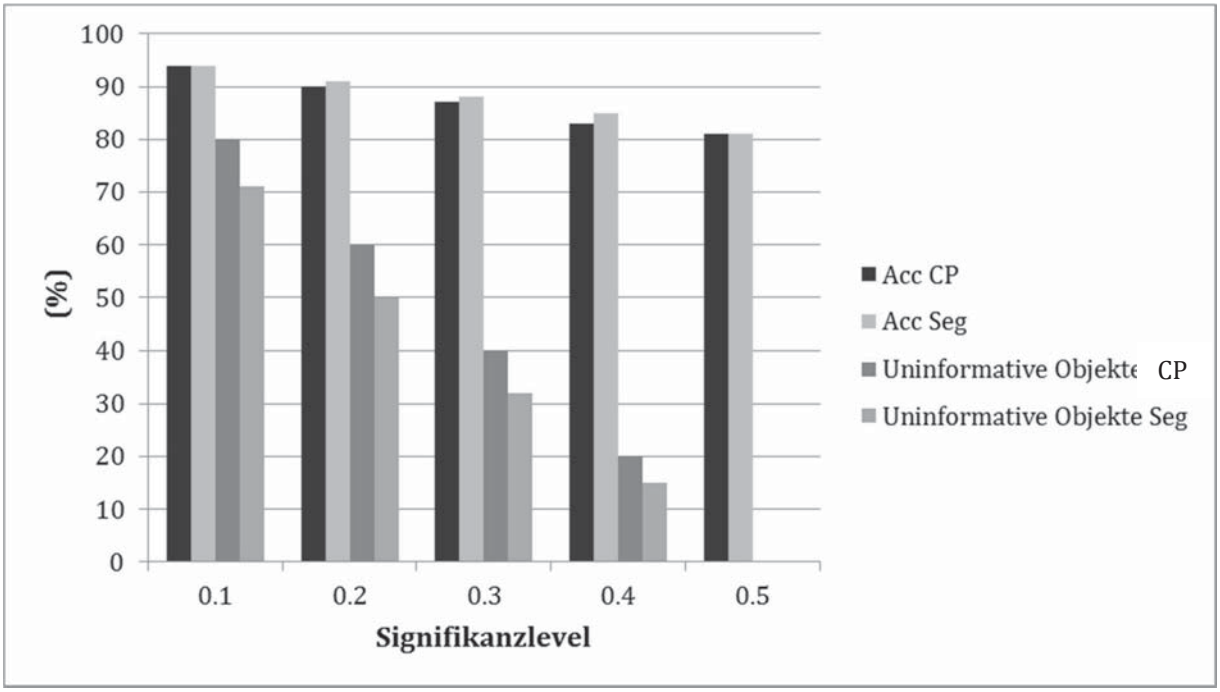


Abbildung 52: Korrektklassifizierungsrate (Acc) und die uninformativ vorhergesagten Objekte des CP und der Methode, welche Segmente entfernt auf Basis des Ames-Datensatzes. Beide Methoden können die vorgegebenen Signifikanzlevel halten. Der Ansatz, welcher Segmente aus der Mitte der sortierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer entfernt, sagt weniger Objekte uninformativ vorher.

5 Diskussion

5.1 Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeits-schätzern

5.1.1 Visuelle Analyse der Zuverlässigkeits-Diagramme und Histogramme von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerten unterschiedlicher Klassifikations-und Regressionstechniken vor und nach Kalibrierung

In den Zuverlässigkeits-Diagrammen zeigen die Klassenzugehörigkeits-Wahrscheinlichkeits-Kurven, sowohl von RF als auch KNN, einen sigmoiden Verlauf, welcher durch logistische Regression ausgeglichen wird. Bei detaillierterer Betrachtung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen wird erkannt, dass die Schätzwerte zur Mitte des Histogramms hin verschoben sind und dass Schätzwerte nahe 0 und 1 relativ selten sind. Einen Erklärungsansatz für dieses Phänomen liefern Niculescu-Mizil und Caruana [95]. Methoden wie der RF auf Basis von Bagging mitteln Vorhersagen aus einem Satz unterschiedlicher Modelle und haben deshalb Probleme Vorhersagen nahe 0 und 1 zu treffen, da die Varianz in den vorliegenden Modellen die Vorhersage, die sich nahe 0 und 1 befinden sollten verzerren (von diesen Werten weg) [95]. Die Richtung ist hierbei vorgegeben da Vorhersagen nur in dem Intervall zwischen 0 und 1 liegen können. Zur Verdeutlichung ein Beispiel: Ein beliebiges Modell soll für einen Fall 0 vorhersagen. Dies wäre für das Bagging nur möglich, wenn alle Bäume des Ensembles 0 vorhersagen würden. Wenn an dieser Stelle Rauschen zu den Bäumen addiert werden würde, dann würde dieses Rauschen dazu führen, dass einige Bäume für den betrachteten Fall einen Wert größer 0 vorhersagen würden. Nachdem über alle Bäume gemittelt werden würde, würde sich somit der Mittelwert der Vorhersage von 0 entfernen [95]. Der KNN kann ebenfalls nur Vorhersagen zwischen 0 und 1 treffen, die Vorhersage hängt von der Population der Nachbarschaft ab. Je mehr Nachbarn in Betracht gezogen werden, desto seltener ist es, dass alle Nachbarn derselben Klasse angehören, deshalb sind die Vorhersagen nahe 0 und 1 auch seltener. Durch die begrenzte Anzahl an Nachbarn sind auch nur begrenzte Schätzwerte möglich, deshalb entstehen zum Teil leere Segmente bei den Zuverlässigkeitsdiagrammen (siehe Kapitel 1.7.3). Um die Diversität zu erhöhen und somit die Anzahl leerer Segmente zu reduzieren, wurde immer mit k=15 gerechnet, unabhängig von der Klassifikationsleistung des KNN. Bei den SVM werden

ebenfalls Wahrscheinlichkeitsschätzer, welche von 0 und 1 zur Mitte hin verschoben sind, beobachtet. Diese lassen sich gleichermaßen mittels logistischer Regression rekalibrieren. An dieser Stelle muss bedacht werden, dass es sich bei den SVM nicht um Wahrscheinlichkeitsschätzer im eigentlichen Sinn handelt, sondern die Distanz der Objekte zur Entscheidungsebene. Das Phänomen der Verschiebung der Vorhersagen zur Mitte hin wird grundsätzlich bei allen Klassifikationsmethoden beobachtet, welche die Margin maximieren (siehe Kapitel 1.5.5) [95]. Nach der Kalibrierung wird dies behoben und die Wahrscheinlichkeitsschätzungen befinden sich nah der wahren Wahrscheinlichkeit. Die Zuverlässigkeits-Diagramme der LDA und des NBC zeigen für unkorrelierte Variablen, dass sich bereits die unkalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer nah an der wahren Wahrscheinlichkeit befinden und, dass Kalibrierung zu keiner Verbesserung führt. Die Begründung für die LDA ist, dass diese echte A-posteriori Wahrscheinlichkeiten schätzt [15]. Die Klassenzuweisung in der LDA erfolgt basierend auf der maximalen Klassenzugehörigkeits-Wahrscheinlichkeit für eine bestimmte Klasse. Dasselbe gilt auch für den NBC wenn seine Annahmen erfüllt sind [15]. Nach der Kalibrierung kommt es somit lediglich zu einer Verschlechterung der Wahrscheinlichkeitsschätzung. Bei Betrachtung der Zuverlässigkeits-Diagramme und Histogramme der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von einem Ensemble an NN im Klassifikationsmodus ist zu erkennen, dass die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzungen bereits vor der Kalibrierung gut sind und diese zu keiner Verbesserung führt [135]. Werden NN mit der Back-Propagation-Regel, welche die Summe quadrierter Fehler minimiert, in Form eines Feeforward NN mit Multilayer verwendet, dann schätzt diese Ausgabe (engl.: Output) Klassenzugehörigkeits-Wahrscheinlichkeiten [75]. Durch die Verwendung von Ensembles werden die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer zusätzlich verbessert. Die PLSDA kann, wie bereits in der Einleitung beschrieben, auf zwei unterschiedliche Weisen (siehe Kapitel 1.5.9) berechnet werden. Allerdings ist die bevorzugte Variante im Fall dicht besetzter Matrizen, die Verwendung der Klasseninformation im Training als abhängige Variable und die Vorhersage kontinuierlicher Werte für die betrachteteten Moleküle. Diesen wird anschließend diejenige Klasse zugewiesen, welche die geringste Differenz zum vorhergesagten, kontinuierlichen Wert aufweist. Bei Verwendung eines, in Hinblick auf den Rang, vollen PLSDA-Modells, bringt die PLSDA dasselbe Klassifikationsmodell hervor wie die LDA [136]. Um allerdings die Vorhersagen der Klassenzugehörigkeit zu Wahr-

scheinlichkeitsschätzern zu transformieren, wird die softmax Funktion verwendet, sodass die Vorhersagen Werte zwischen [0,1] annehmen und deren Zeilensummen 1 ergeben. (Bei der softmax Funktion handelt es sich um eine Generalisierung der logistischen Funktion [78]). Dies erklärt warum die PLSDA ebenfalls sich nah an der wahren Wahrscheinlichkeit befindende Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervorbringt und, dass eine anschließende, logistische Regression, durch die zuvorige Anwendung der softmax Funktion, nur einen geringen Einfluss hat.

Im Regressionsfall werden die vorhergesagten kontinuierlichen Werte zur Schätzung der Klassenzugehörigkeits-Wahrscheinlichkeit genutzt. Wie bereits im Methodenteil beschrieben minimieren Regressionsmethoden, im Gegensatz zu Klassifikationsmethoden, in der Regel die Summe der quadratischen Fehler. Diese können allerdings ebenfalls zur Lösung von Zwei-Klassen-Klassifikationsproblemen genutzt werden. Der Regressionsalgorithmus minimiert die quadratische Abweichung der angepassten Werte von den Klasseninformationen und unter diesen Voraussetzungen schätzt die Regressionsfunktion $\hat{f}(\boldsymbol{x_0})$ Klassenzugehörigkeits-Wahrscheinlichkeiten [15]. Im Falle der Techniken Lasso, Ridge und Elastic Net können allerdings Probleme auftreten, da $\hat{y}(\boldsymbol{x_0})$ nicht zwischen 0 und 1 liegen muss. Bei Betrachtung realer Fragestellungen ist es von Bedeutung, dass die Regressionsfunktion die Erwartungswerte $E(1|\boldsymbol{x_0})$ gut approximieren können muss, um gute Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer liefern zu können. Bei vielen nichtparametrischen Regressionstechniken, wie beim Random Forest im Regressionsmodus, schätzt $\hat{y}(\boldsymbol{x_0})$ $p = (1|\boldsymbol{x_0})$ konsistent [110], wenn die Stichprobengröße gegen unendlich strebt. Allerdings muss an dieser Stelle bedacht werden, dass Konsistenz alleine nichts über die Eigenschaften eines bestimmten Schätzers in kleinen Stichproben aussagt [112]. Bei den Techniken Lasso, Ridge und Elastic Net wird eine stärker ausgeprägte sigmoide Kurve im Zuverlässigkeits-Diagramm beobachtet. Der sigmoide Verlauf wird, wie bereits im Ergebnisteil gezeigt, durch logistische Regression behoben. Der stärkere sigmoide Verlauf könnte aus der Fähigkeit der Techniken hervorgehen, im Gegensatz zum RFR und zur SVR, über 0 und 1 extrapolieren zu können. Durch die Fähigkeit des Extrapolierens ergeben sich unterschiedliche Möglichkeiten die Wahrscheinlichkeitsschätzer auf eine Skala zwischen 0 und 1 zu bekommen. Auf diese Problematik wird in Kapitel 5.1.3 näher eingegangen.

5.1.2 Einfluss der Variablenanzahl des Datensatzes auf den Fehler sowie Beurteilung der Fehlermaße

Die Anzahl an Variablen beeinflusst hauptsächlich die Qualität der Wahrscheinlichkeitsschätzer des RF, wenn es sich um unkorrelierte Daten handelt. Je mehr Variablen verwendet werden, desto sicherer wird sich der RF, daher kommt es zu einer Verschiebung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von 0.5 in Richtung 0.4 oder 0.6. Dieses Phänomen ist in Abbildung 53 dargestellt. Allerdings ist diese Verschiebung, verglichen mit den kalibrierten Schätzwerten, welche deutlich näher an der wahren Wahrscheinlichkeit liegen, nicht ausreichend. Dies führt zu einer Vergrößerung des berechneten Fehlers.

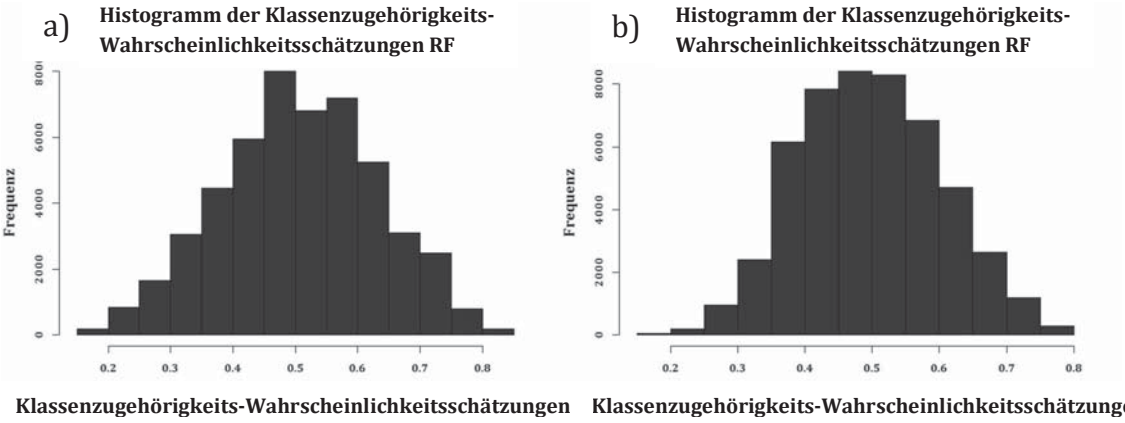


Abbildung 53: Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF bei einer Korrektklassifizierungsrate (Acc) von 0.7 und 2000 Objekten. Die Daten sind unkorreliert, das linke Histogramm (a) basiert auf der Verwendung von 20 Variablen und das rechte Histogramm (b) von 60 Variablen. Es ist zu erkennen, dass es zu einer leichten Verschiebung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von der Mitte zu den Rändern kommt.

Wenn korrelierte Variablen verwendet werden, ist die Abhängigkeit der Fehlergröße von der Anzahl verwendeter Variablen nicht mehr erkennbar. Des Weiteren ist der Fehler deutlich kleiner als bei unkorrelierten Variablen. In Abbildung 54 ist jeweils ein Histogramm der mit korrelierten und unkorrelierten Variablen bei konstanter Korrektklassifizierungsrate (Acc) und Datensatzgröße visualisiert.

(Die Histogramme, welche die unkalibrierten und kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer darstellen, unterscheiden sich hinsichtlich der counts (y-Achse). Der Grund hierfür ist das CV-Schema, durch die Verwendung einer 50*50% LO-CV stehen am Ende nur die Hälfte der unkalibrierten Klassenzugehörigkeits-

Wahrscheinlichkeitsschätzer, in Form von kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern, zur Verfügung).

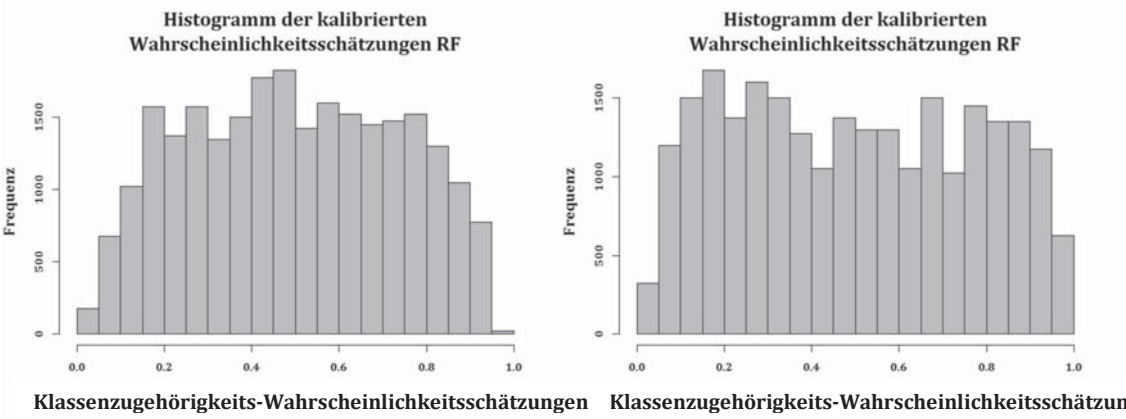


Abbildung 54: Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF bei einer Korrektklassifizierungsrate (Acc) von 0.7 und 2000 Objekten. Die Variablen sind korreliert, das linke Histogramm (a) basiert auf der Verwendung von 20 Variablen und das rechte Histogramm (b) von 60 Variablen. Es ist zu erkennen, dass es zu einer leichten Verschiebung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von der Mitte zu den Rändern kommt.

Auffallend ist, dass es zu einer starken Verschiebung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer zu höheren bzw. niedrigeren Werten kommt.

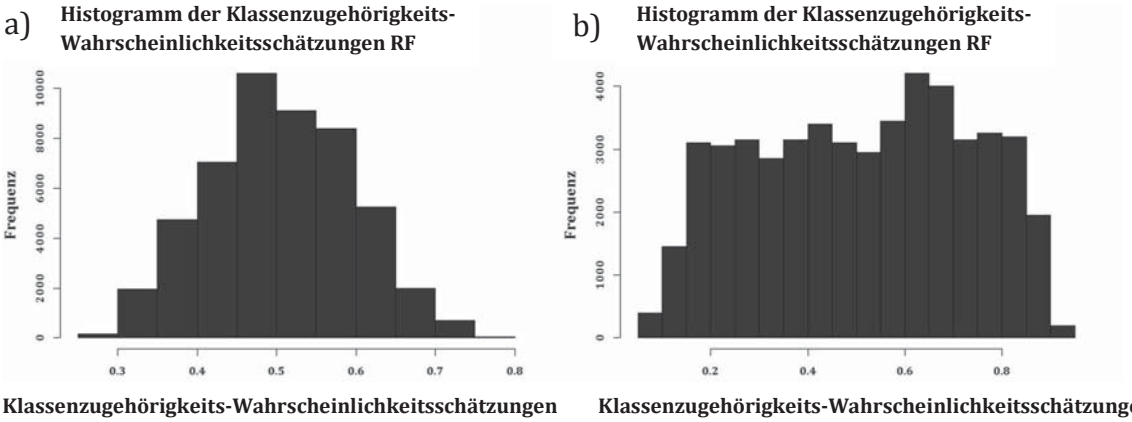


Abbildung 55: Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RF bei einer Korrektklassifizierungsrate (Acc) von 0.8, 40 Variablen und 2000 Objekten. Das linke Histogramm (a) basiert auf unkorrelierten Variablen und das rechte Histogramm (b) auf korrelierten Variablen (r=0.1). Es ist zu erkennen, dass es zu einer starken Verschiebung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer von der Mitte zu den Rändern kommt.

Eine mögliche Begründung liefert zum einen erneut der im letzten Abschnitt beschriebene Erklärungsansatz von Niculescu-Mizil und Caruana [95] und die generelle Funktionsweise des RF [63]. Je höher die Varianz im Vorhersagefehler des einzelnen Baumes,

desto stärker ist die Verschiebung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer zur Mitte hin, durch die festgelegte Ober- und Untergrenze von 0 und 1. Je mehr unkorrelierte Variablen verwendet werden, desto mehr Möglichkeiten bekommt der RF, sofern die Korrektklassifizierungsrate (Acc) konstant gehalten wird. Somit kommt es zu einer Erhöhung der Varianz und somit auch zu einer Erhöhung des unkalibrierten Fehlers. An dieser Stelle stellt sich darüber hinaus noch die Frage, warum in dieser Arbeit die Wahrscheinlichkeitsschätzer „gewöhnlicher" RFs im Klassifikationsmodus untersucht wurden und keine speziellen „Probability Estimation Trees" (PET) [137]. Der Grund hierfür ist, dass spezielle PET keinen Vorteil gegenüber gewöhnlichen CART im Bereich der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer liefern [138].

Zur Evaluation der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer wurden der MAE, MSE, MAE_w, MSE_w, MAE_Cu, MSE_Cu, sowie der Brier-Score berechnet, da diese in der Literatur die gängigste Methode ist [145]. Die Fehlermaße wurden sowohl für die kalibrierten, als auch für die unkalibrierten Ergebnisse berechnet. Da der MSE größere Abweichungen stärker bestraft als der MAE, lag die Vermutung nahe, dass die Berechnung beider Fehlermaße vorteilhaft ist zur Erkennung von Ausreißern. Allerdings zeigt sich in den Ergebnissen, dass diese Unterscheidung praktisch nicht relevant ist. Da weder in den MAE noch in den MSE die Besetzungsstatistik der Segmente im Zuverlässigkeits-Diagramm eingeht, wurde zusätzlich der MAE_w bzw. MSE_w berechnet, bei welchem die jeweiligen Segmentmittelwerte noch einmal durch die Anzahl der Objekte geteilt werden. Darüber hinaus wurde von Caruana und Niculescu-Mizil noch eine weitere Variante berechnet, bei welcher anstelle der zehn Mittelwerte der Segmente, immer 100 Objekte in ein Segment gepackt werden und dann von (Sx) {S1=1-100, S2=2-101,S3=3-102 etc. } der gleitenden Mittelwert berechnet wird (MAE_Cu bzw. MSE_Cu) [102]. Alle Fehlermaße, ausgenommen der Brier-Score, korrelieren stark miteinander. Dies ist aufgrund der Berechnungsweise auch nicht überraschend, da in den betrachteten Simulationsstudien und Realdatensätzen Ausreißer offensichtlich eine untergeordnete Rolle spielen. Aus diesem Grund ist folglich ein Fehlermaß für die kalibrierten und unkalibrierten Ergebnisse ausreichend. Hierfür wurde das Fehlermaß von Caruana und Niculescu-Mizil ausgewählt, da es einerseits, im Gegensatz zur Brier-Score, den „reinen Kalibrierfolg" (Calibration Loss (siehe Kapitel 3.1.4)) misst und andererseits durch die schrittweise Segmentbildung keine leeren Segmente hervorbringt. Die ungewichteten

und gewichteten MAE bzw. MSE Werte können in einigen Fällen aufgrund leerer Segmente nicht berechnet werden.

5.1.3 Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels Simulationsstudien und Realdatensätzen

Bei der SVM ist eine Abhängigkeit der unkalibrierten Klassenzugehörigkeits-Schätzwerte von der Korrektklassifizierungsrate (Acc) und der Korrelation erkennbar. Im Detail sind die Histogramme der Schätzwerte in Abbildung 56 dargestellt. Es ist zu erkennen, dass sich die Schätzwerte mit steigender Korrektklassifizierungsrate (Acc) von der Mitte in Richtung Ränder des Histogramms verschieben, allerdings nur leicht, sodass es zunächst nicht, wie nach Kalibrierung, zu einer Verringerung des Fehlers, sondern zu einer Vergrößerung kommt. Bei den Klassenzugehörigkeits-Schätzwerten handelt es sich, wie bereits im Methodenteil beschrieben, um Werte, welche die Nähe zur Entscheidungsebene charakterisieren. Je größer diese sind, desto verlässlicher ist die Klassifikation. Wenn die Korrektklassifizierungsrate (Acc) steigt ist sich die Technik sicherer und somit nehmen die Klassenzugehörigkeits-Schätzwerte größere Werte an. Die Korrektklassifizierungsrate (Acc) verändert somit den sigmoidalen Verlauf der Kurve [96]. In allen Fällen führt die logistische Regression zu einer Verschiebung Richtung Winkelhalbierenden und somit zur wahren Wahrscheinlichkeit. Für den RF wurden bereits im Kapitel 5.1.2 alle Einflussfaktoren, mit Ausnahme der Korrektklassifizierungsrate (Acc) diskutiert. Beim RF werden für die Korrektklassifizierungsrate (Acc) dieselben Beobachtungen gemacht, wie bereits bei der SVM. Die Begründung deckt sich mit derer der SVM. Durch Kalibrierung kommt es beim RF in allen Fällen mit Ausnahme einer Korrektklassifizierungsrate (Acc) von 0.7 und korrelierten Daten zu einer Verbesserung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Denn durch logistische Regression werden die Schätzwerte von der Mitte (0.5) zu den Rändern geschoben (auch nahe 0 und 1), die der RF aufgrund seiner Funktionsweise schlecht erreichen kann. Je höher die Korrektklassifizierungsrate (Acc), desto mehr Schätzwerte sollten nahe 0 und 1 liegen. Somit ist der Effekt der Kalibrierung in diesen Fällen am größten. Bei korrelierten Daten kommt es von Anfang an zu einer breiteren Verteilung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer (Abbildung 56). Liegt die Korrektklassifizierungsrate (Acc) unter diesen Bedingungen lediglich bei 0.7, dann liegen ohnehin nicht viele Schätzwerte

nahe 0 und 1 und somit ist der Kalibriererfolg geringer. Eine detailliertere Analyse der Zunahme des MAE mit steigender Korrektklassifizierungsrate (Acc) wird im unteren Abschnitt am Beispiel des Lassos durchgeführt, bei welchem dieses Phänomen unabhängig von der Korrelation der Daten, am ausgeprägtesten beobachtet wird.

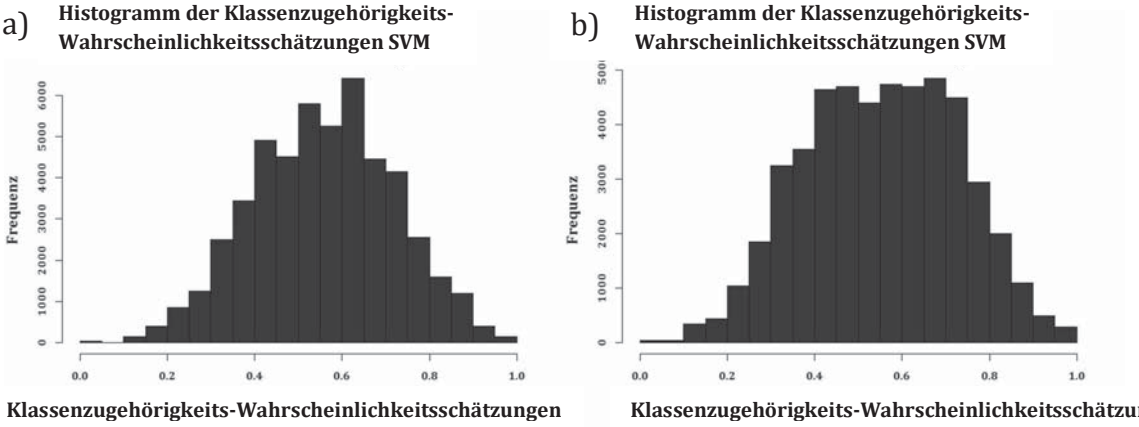


Abbildung 56: Unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte der SVM (decision values) für 2000 Objekte bei unkorrelierten Daten. Das linke Histogramm (a) zeigt die Klassenzugehörigkeits-Schätzwerte bei einer Korrektklassifizierungsrate (Acc) von 0.7 und das rechte Histogramm (b) bei 0.8. Es ist zu erkennen, dass sich die Schätzwerte von der Mitte in Richtung Ränder verschieben.

Die Ergebnisse des KNN lassen sich auf dieselbe Weise wie beim RF begründen, da im Falle des KNN 15 Nachbarn ausgewählt wurden, welche ebenfalls über eine Mehrheitsentscheidung das Ergebnis bestimmen. Somit ist es für den KNN ebenfalls schwieriger Vorhersagen nahe 0 und 1 zu tätigen. Die NN zeigen keine Abhängigkeit von der Korrelation der Daten und nur eine sehr schwache Abhängigkeit der Korrektklassifizierungsrate (Acc). Der Fehler nimmt mit steigender Korrektklassifizierungsrate (Acc) ab. NN bringen, wie bereits in der Einleitung beschrieben, unkalibriert gute Wahrscheinlichkeitsschätzer hervor [135]. Dasselbe gilt auch für die LDA. Es ist weder eine Abhängigkeit von der Korrektklassifizierungsrate (Acc) noch von der Korrelation erkennbar [15, 78, 48].

5.1.3.1 Detailliertere Betrachtung der Kalibrierung des NBC

Im Fall des NBC wurde beobachtet, dass der MAE stark ansteigt, sobald die Annahmen des NBC nicht mehr erfüllt sind (Annahmen wurden durch Korrelation der Daten verletzt). Nach der Kalibrierung mit der isotonischen Regression resultierten wieder gute Wahrscheinlichkeitsschätzer. Die Begründung hierfür ist, dass der NBC bei Verletzung

der Annahmen dazu neigt die Wahrscheinlichkeiten in Richtung 0 oder 1 zu schieben [79]. Die isotonische Regression ist der logistischen Regression in diesem Fall überlegen, denn die logistische Regression passt eine sigmoide Form an und die isotonische eine willkürlich steigende (isotonische) Funktion [16]. Dadurch hat diese einen schwächer ausgeprägten systematischen Fehler und kann somit breitflächiger angewendet werden. Der Nachteil ist allerdings, dass die isotonische Regression durch den schwächeren ausgeprägten systematischen Fehler mehr Daten zur Kalibrierung benötigt [95].

(Ende des Unterkapitels 5.1.3.1)

Die Ergebnisse des RFR und der SVR gleichen denen der Klassifikationstechniken. Aus diesem Grund wird an dieser Stelle nicht näher auf diese eingegangen. Die übrigen Regressionstechniken können, im Gegensatz zum RFR und der SVR, über 0 und 1 hinaus extrapolieren. Aus diesem Grund ergaben sich zwei unterschiedliche Möglichkeiten die ausgegebenen Regressionswerte wieder auf eine Skala zwischen 0 und 1 zu bringen. Die erste Möglichkeit ist diese so skalieren, dass der minimalste Wert 0 und der maximalste Wert 1 ist. Die zweite Möglichkeit ist es die Werte oberhalb von 1 abzuschneiden. Die Ergebnisse zeigen, dass die Skalierung zwischen 0 und 1 zu deutlich schlechteren unkalibrierten Klassenzugehörigkeits-Schätzwerten führt. Eine Erklärung hierfür ist, dass Regressionsfunktionen, die die quadratische Abweichung der angepassten Werte von den Klasseninformationen minimieren, Klassenzugehörigkeits-Wahrscheinlichkeiten schätzen [15]. Dies gilt allerdings nur für Werte zwischen 0 und 1. Folglich wären alle Werte über 1 mit einem Wert gleich 1 gleichzusetzten. Wenn diese Werte skaliert werden, kommt es zu einer „falschen“ Verschiebung aller Klassenzugehörigkeits-Schätzwerte Richtung 0 und somit verschlechtern sich alle Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer und der Gesamtfehler steigt an. Die Regressionstechniken zeigten eine Zunahme des unkalibrierten Fehlers bei steigender Korrektklassifizierungsrate (Acc) unabhängig von der Korrelation der Daten. An dieser Stelle werden exemplarisch, anhand der Histogramme der unkalibrierten und kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des Lassos, vergleichbar mit dem RF, die Abhängigkeit der Fehlerschätzung von der Korrektklassifizierungsrate (Acc) erklärt. Abbildung 57 zeigt die Histogramme der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des Lassos bei einer Korrektklassifizierungsrate (Acc) von 0.7. Sowohl der MAE_Cu als auch der MAE_Cu_Log liegen bei ca. 0.03. Die Fehler sind sehr gering, folglich liegen die Klas-

senzugehörigkeits-Wahrscheinlichkeitsschätzer nah an den wahren Wahrscheinlichkeitsschätzern. Beide Histogramme zeigen eine vergleichbare Form. Durch die vorgegebene Korrektklassifizierungsrate (Acc) von 0.7 liegen die Schätzwerte hauptsächlich zwischen 0.3 und 0.7, da viele Moleküle vergleichbar unsicher vorhergesagt werden. Mit steigender Korrektklassifizierungsrate (Acc) sollten die Schätzwerte von der Mitte zu den Rändern (Bereichen mit höherer bzw. niedrigerer Wahrscheinlichkeit) verschoben werden, da die Klassifikationstechnik weniger Fehler macht und somit die Zuverlässigkeit zunimmt.

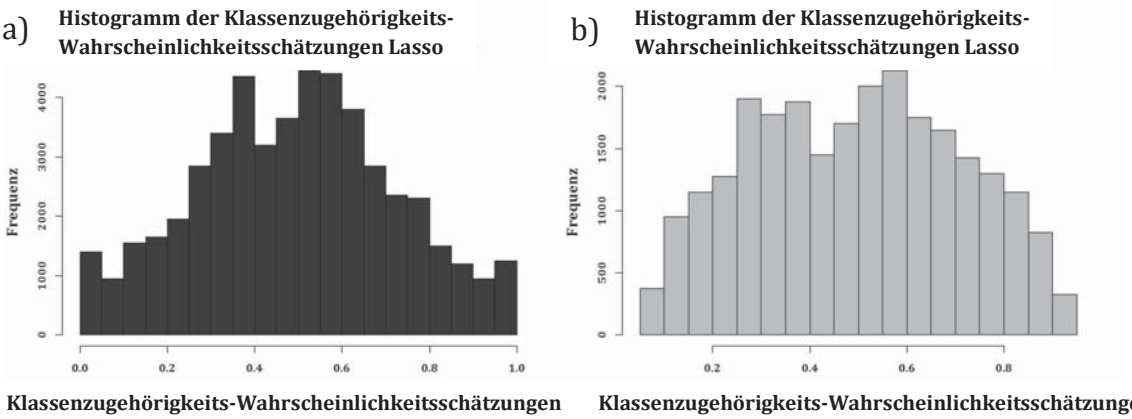


Abbildung 57: Unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitschätzwerte des Lassos bei einer Korrektklassifizierungsrate (Acc) von 0.7, für 2000 Objekte und für unkorrelierte Daten. Das linke Histogramm (a) zeigt die unkalibrierten Klassenzugehörigkeits-Schätzwerte und das rechte Histogramm (b) die kalibrierten Schätzwerte. Beide Histogramme ähneln sich.

Die nachfolgende Abbildung 58 zeigt die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte bei einer Korrektklassifizierungsrate (Acc) von 0.8. Der MAE_Cu liegt in diesem Fall oberhalb des MAE_Cu_Log. Der MAE_Cu_Log befindet sich unterhalb von 0.05, folglich nah an der wahren Wahrscheinlichkeit. Es ist zu erkennen, dass es zu einer Verschiebung der Wahrscheinlichkeitsschätzer zu den äußeren Bereichen der Histogramme kommt. Allerdings ist diese Verschiebung im unkalibrierten Fall, verglichen zu dem kalibrierten Fall, nicht ausreichend, was auch den höheren MAE_Cu erklärt. Somit führt die logistische Regression zu einer besseren Annäherung an die wahre Wahrscheinlichkeit. (Die hohen Werte in Abbildung 58 a), >1 und <0, sind Artefakte durch das Abschneiden der Werte. Der Rest ist immer noch gleichverteilt).

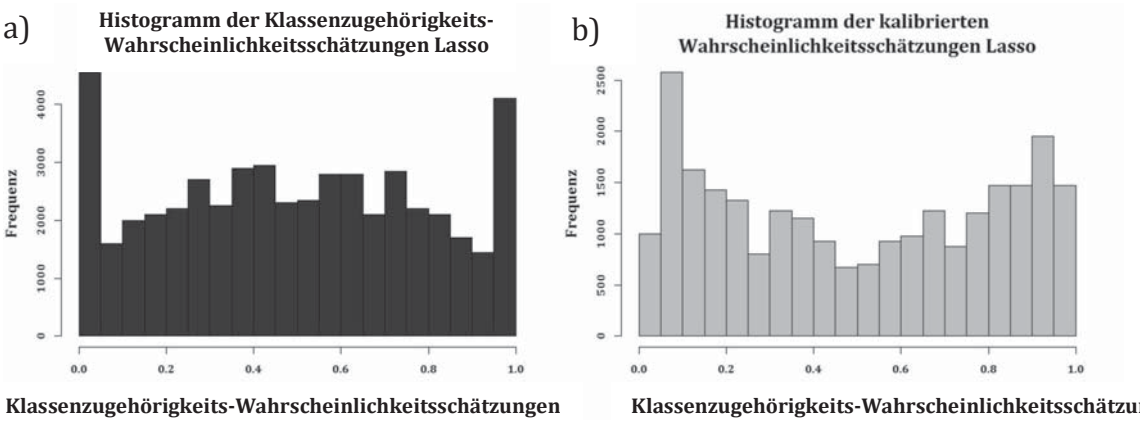


Abbildung 58: Unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte des Lassos bei einer Korrektklassifizierungsrate (Acc) von 0.8, für 2000 Objekte und für unkorrelierte Daten. Das linke Histogramm (a) zeigt die unkalibrierten Klassenzugehörigkeits-Schätzwerte und das rechte Histogramm (b) die kalibrierten Schätzwerte. Mit steigender Korrektklassifizierungsrate (Acc) kommt es zu einer Verschiebung der Wahrscheinlichkeitsschätzer zu den Randbereichen der Histogramme, da die Klassifikationsleistung sich verbessert und die Vorhersagen somit zuverlässiger werden. Die Schätzwerte des rechten Diagramms liegen näher an der wahren Wahrscheinlichkeit, als die des linken Histogramms. Folglich ist der Fehler des linken Diagramms höher.

Bei einer Korrektklassifizierungsrate (Acc) von 0.9 (siehe Abbildung 59) verschieben sich die Wahrscheinlichkeitsschätzwerte noch weiter in Richtung der Randbereiche der Histogramme, da die Klassifikationsleistung noch weiter verbessert wird und folglich die Zuverlässigkeit der Vorhersagen weiter zunimmt. Wie schon bei der Korrektklassifizierungsrate (Acc) von 0.8 beobachtet, ist diese Verschiebung bei den unkalibrierten Schätzwerten nicht ausreichend und der MAE_Cu spiegelt dies mit einem Wert oberhalb von 0.1 wieder. Der MAE_Cu_Log der kalibrierten Schätzwerte hingegen befindet sich unterhalb von 0.05. Die kalibrierten Schätzwerte sind somit nah an der wahren Wahrscheinlichkeit.

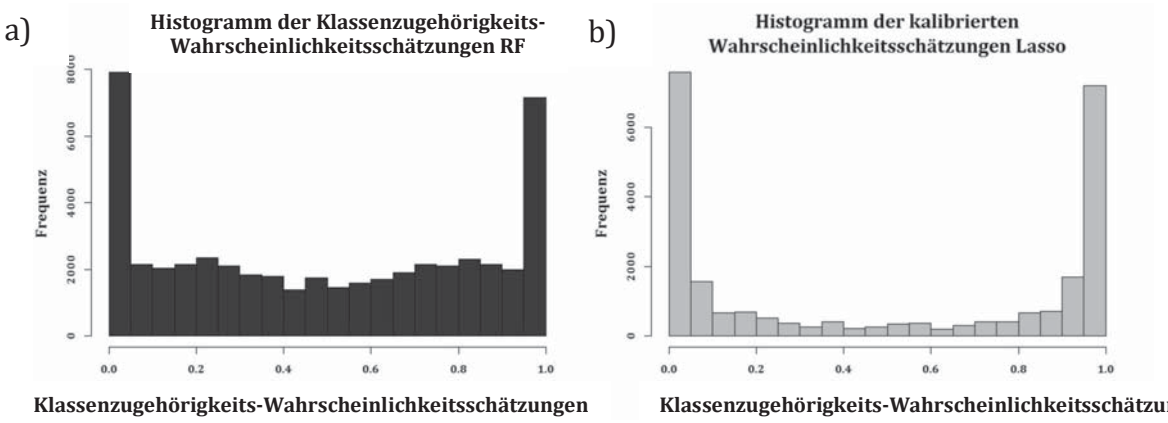


Abbildung 59: Unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte des Lassos bei einer Korrektklassifizierungsrate (Acc) von 0.9, für 2000 Objekte und für unkorrelierte Daten. Das linke Histogramm (a) zeigt die unkalibrierten Klassenzugehörigkeits-Schätzwerte und das rechte Histogramm (b) die kalibrierten Schätzwerte. Mit steigender Korrektklassifizierungsrate (Acc) kommt es zu einer Verschiebung der Wahrscheinlichkeitsschätzer zu den Randbereichen der Histogramme, da die Klassifikationsleistung sich verbessert und die Vorhersagen somit zuverlässiger werden. Die Schätzwerte des rechten Diagramms liegen näher an der wahren Wahrscheinlichkeit, als die des linken Histogramms. Folglich ist der Fehler des linken Diagramms höher.

Nachdem alle Einflussfaktoren, mit Ausnahme der Anzahl an Objekten analysiert wurden, wird an dieser Stelle kurz auf diesen Faktor eingegangen. Wie bereits im Vorfeld erwartet wurde haben alle Techniken gemeinsam, dass sich eine steigende Anzahl an Objekten positiv auf die Fehlerschätzung auswirkt, das bedeutet, dass sich der Fehler reduziert. Darüber hinaus führt eine steigende Anzahl an Objekten zu einer Reduktion der Standardabweichung.

Die Ergebnisse der Realdatensatzbeispiele decken sich mit den Ergebnissen aus den Simulationsstudien. Daher werden diese nicht wiederholt diskutiert. Die Verwendung unterschiedlicher Deskriptoren hat lediglich einen geringen Einfluss, bzw. dieser Einfluss lässt sich ebenfalls auf die schon zuvor analysierten Faktoren: Anzahl an Variablen, Korrelation, Korrektklassifizierungsrate (Acc) und Datensatzgröße zurückführen.

5.1.4 Analyse des Einflusses von Hetero-Ensemblen auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations-und Regressionstechniken mittels Simulationsstudien und Realdatensätzen

In vielen Studien konnte bisher gezeigt werden, dass Ensembles von Klassifikatoren die Fehlerrate eines Klassifikationsproblems, verglichen mit einzelnen Klassifikatoren, senken können. Besonders die Kombination unterschiedlicher Klassifikationstechniken,

auch bekannt als „Hetero-Ensembles“, führt zu einer Stabilisierung der Vorhersagen, dadurch, dass Schwierigkeiten einzelner Klassifikationstechniken auf bestimmten Datensätzen ausgeglichen werden können [139, 140].

Dies spiegelt sich auch in den Ergebnissen wieder, die Bildung von Hetero-Ensemblen stabilisiert den MAE_Cu bzw. den MAE_Cu_Log. Die Fehlerhöhe wird in den meisten Fällen auf die Höhe der niedrigsten Einzelfehlerwerte reduziert, auch wenn einige Klassifikationstechniken einen hohen MAE_Cu aufweisen. Kalibrierung führt in allen Fällen zu einer Verringerung der Fehlerhöhe. Aus den Simulationsstudien und besonders aus den Realdatensatzbeispielen ging zusammenfassend hervor, dass am besten zuerst die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte gemittelt werden sollten und danach kalibriert werden sollten. Diese Vorgehensweise führt zu dem niedrigsten MAE_Cu_Log-Werten und folglich befinden sich die so entstandenen Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte am nächsten an der wahren Wahrscheinlichkeit. Diese Reihenfolge wird auch von A. Bella et al. vorgeschlagen, welche eine Vergleichsstudie auf Basis der UCI repository Datensätze durchgeführt hatten [141]. Allerdings muss an dieser Stelle bedacht werden, dass die Unterschiede zwischen den Varianten: zuerst kalibrieren und dann kombinieren, zuerst alle mit „schlechten“ unkalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern kalibrieren und dann kombinieren oder erst kombinieren und dann kalibrieren, gering waren. Die Abhängigkeit des unkalibrierten Ensemble MAE_Cu-Wertes von der Korrektklassifizierungsrate (Acc), der Korrelation und der Datensatzgröße, lassen sich auf die bereits im letzten Kapitel (5.1.3) diskutierten Effekte zurückführen.

5.2 Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern versus CP

Aus den Realdatensatzbeispielen, berechnet mit dem RF, ging hervor, dass der CP mehr Moleküle als uninformativ vorhersagt als die Reject-Option, welche Segmente von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern aus der Mitte von Zuverlässigkeitssdiagrammen herausnimmt. Allerdings konnten beide Techniken, bei den betrachteten Datensätzen, die vorgegebenen Signifikanzlevel einhalten.

Jedoch hat der CP den Vorteil, dass dieser per Definition unter bestimmten Bedingungen das vorgegebene Signifikanzlevel automatisch einhält [107, 142]. Um dies zu garantieren, kommt es zu der gehäuften Vorhersage von uninformativen Molekülen. Die Herausnahme von Segmenten von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern aus der Mitte der Zuverlässigkeitsdiagramme hingegen garantiert nur ein bestimmtes Signifikanzlevel, wenn die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte mit den wahren Wahrscheinlichkeitswerten übereinstimmen. Dies ist allerdings, wie bereits in den letzten Kapiteln beschrieben, oftmals nicht zutreffend. Durch Kalibrierung können in vielen Fällen die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer näher an die wahre Wahrscheinlichkeit herangeführt werden und dann wären die erhaltenen kalibrierten Schätzwerte besser dazu geeignet in dieser Technik verwendet werden zu können. Durch die Unsicherheit der möglicherweise schlecht kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte scheint diese Technik zunächst dem CP unterlegen zu sein. Jedoch muss an dieser Stelle bedacht werden, dass es sich bei den betrachteten Anwendungsfällen um abgeschlossene Datensätze handelt und zusätzlich wurde zur Validierung die Kreuzvalidierung verwendet. Die Annahmen des CP gelten allerdings nur im „Online“ Modus, das beudet, dass nach jeder Vorhersage die wahre Klasse ermittelt wird und der Klassifikator dieses neue Objekt fortan für seine weiteren Vorhersagen, zum Beispiel durch Aktualisieren des Modells, nutzen kann. Weiter verbreitet als das Online-Lernen im Bereich der Chemieinformatik ist jedoch das Chargenweise lernen. Für diesen Lernmodus, sowie für das Ermitteln der p-Werte durch Kreuzvalidierung, gilt die garantierte Erreichung des Signifikanzlevels nicht unbedingt [108]. Allerdings funktioniert der CP im Mittel in diesen Fällen auch gut, jedoch ohne garantiertes Signifikanzlevel. Somit zeigen die Ergebnisse, dass die Verwendung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern eine gute und schnelle Alternative darstellt um einen AB zu definieren, da beide Techniken mit einer Unsicherheit behaftet sind.

Aus der in Kapitel 2.1 vorgestellten Studie zur Beurteilung der Effizienz von AB-Maßen [109] wurde gezeigt, dass die meisten Klassifikationsfehler nahe der Entscheidungsebene gemacht werden und somit die Vertrauens-Maße besser dazu geeignet sind einen AB zu definieren. Da es sich bei allen Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern um Vertrauens-Maße handelt, ist es folglich auch für alle Techniken interesssant, wie ein Grenzwert zur Definition eines AB gesetzt werden kann. Ein tatsächlich erreichtes Signifikanzlevel wäre ein idealer Grenzwert um Entscheidungen treffen zu können.

Darüber hinaus können die hier durchgeführten Studien zur Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern dazu verwendet werden, die Unsicherheit der Reject-Option-Technik zu reduzieren. Denn in diesen Studien wurden Informationen darüber generiert, bei welchen Techniken, unter welchen Bedingungen, Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer erhalten werden, welche nah an der wahren Wahrscheinlichkeit liegen.

6 Zusammenfassung und Schlussfolgerung

In dieser Arbeit wurden Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken untersucht. Um Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer zum Treffen von Entscheidungen zu verwenden, sollten diese möglichst nah an der wahren Wahrscheinlichkeit liegen und Charakteristika dieser Schätzer sollten bekannt sein. Aus diesem Grund wurden in dieser Arbeit unterschiedliche Regressions-und Klassifikationstechniken, hinsichtlich ihrer Fähigkeit analysiert, Klassenzugehörigkeits-Wahrscheinlichkeiten möglichst exakt schätzen zu können. Zusätzlich wurde der Effekt der Kalibrierung mittels logistischer Regression, sowie die Einflussfaktoren Korrektklassifizierungsrate (Acc), Korrelation und Datensatzgröße untersucht. Das Ergebnis ist, dass alle untersuchten Techniken (RF, RFR, SVM, SVR, KNN, PLSDA, SPLS, Ridge, Elastic Net, Lasso) mit Ausnahme der LDA, NN und NBC von der Kalibrierung mittels logistischer Regression profitieren. Die erhaltenen Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer befinden sich danach näher an der wahren Wahrscheinlichkeit. Die größten Einflussfaktoren sind die Korrektklassifizierungsrate (Acc) und die Korrelation. Bei einer Vielzahl der Techniken führt eine steigende Korrektklassifizierungsrate (Acc) und eine abnehmende Korrelation zu schlechteren Schätzwerten. Die Bildung von Hetero-Ensemblen führt zu stabileren Schätzwerten. Bei der Bildung von Hetero-Ensemblen sollten zunächst die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerte gemittelt werden und danach sollte kalibriert werden. Auf diese Weise werden die am nahesten an der wahren Wahrscheinlichkeit liegende Schätzwerte erhalten.

Diese Schätzwerte können danach verwendet werden um einen AB für das betrachtete Modell zu definieren. Die Verwendung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern zur Definition eines AB (Reject-Option) wurde verglichen mit dem Ansatz des CP. Der wesentliche Unterschied zwischen diesen beiden Ansätzen ist, dass der CP p-Werte berechnet. Ein Vorteil des Reject-Option-Ansatzes ist, dass exakte Wahrscheinlichkeiten erhalten werden, welche besser geeignet sind um Entscheidungen zu treffen, als irgendeinen beliebiger Score. Für diesen Score müsste, zur Einordnung, ohnehin erst wieder eine Perzentile berechnet werden, wie auch für die p-Werte des CP.

Die Idee hinter beiden Ansätzen ist es ein bestimmtes Signifikanzlevel vorzugeben, indem unzuverlässige Moleküle als solche gekennzeichnet werden und separiert werden. Separierte Moleküle befinden sich dann folglich außerhalb des AB. Beide Methoden haben den Nachteil kein exaktes Konfidenzlevel zu ermöglichen, da Annahmen verletzt werden. Allerdings kann anhand von Realdatensätzen gezeigt werden, dass in vielen Fällen das vorgegebene Konfidenzlevel dennoch gut erhalten werden kann, wenn entsprechend ausreichend Moleküle separiert werden. Es konnte gezeigt werden, dass die Technik, welche Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer verwendet, weniger Moleküle separieren musste um das vorgegebene Signifikanzlevel zu halten und somit etwas effizienter ist.

Zusammenfassend lässt sich sagen, dass gut kalibrierte Vertrauens-Maße dazu in der Lage sind besonders effizient, ohne viele Objekte zu verlieren, einen AB definieren zu können.

7 Ausblick

Die Charakterisierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sollte in Zukunft noch auf weitere Techniken, wie beispielsweise auf „bagged oder boosted trees“ oder auch „bagged oder boosted stumps“ ausgeweitet werden [143]. Darüber hinaus wurden in den Simulationsstudien bisher nur balancierte Datensätze betrachtet. In der Realität sind die meisten Datensätze allerdings nicht ausgeglichen. In diesem Fall müsste die Fehlerberechnung angepasst werden [144, 145].

Des Weiteren ist der Vergleich zwischen CP und der Verwendung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern zur Definition eines AB noch nicht ausreichend, sondern als Pilotprojekt zu sehen. Es gibt bereits aktuellere Varianten des CP (Aggregated Conformal Prediction) [146] welche dessen Effizienz verbessern sollen. Diese sollten zum Vergleich herangezogen werden. Danach sollte die Studie auf weitere Techniken, wie bespielsweise NN und SVM, ausgeweitet werden, denn bisher wurde ausschließlich der RF verwendet. Außerdem sollten mehr stark unbalacierte Datensätze untersucht werden, wie sie beispielsweise im Bereich des Virtuellen Screenings üblich sind. Dies würde Schwierigkeiten bei den $\hat{p}$ der kleineren Klasse verursachen. Hierfür gibt es aber bereits CP’s (Mondrian CP), mit welchem die Reject Option verglichen und gegebenenfalls angepasst werden sollte.

8 References

1. Kubinyi H (1997) QSAR and 3D QSAR in drug design Part 1: methodology. Drug Discov. Today 2(11): 457–467. DOI:10.1016/S1359-6446(97)01079-9
2. Hansch C, Fujita T (1964) p -σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J. Am. Chem. Soc. 86(8): 1616–1626. DOI:10.1021/ja01062a035
3. Free SM, Wilson JW (1964) A Mathematical Contribution to Structure-Activity Studies. J. Med. Chem. 7(4): 395–399. DOI:10.1021/jm00334a001
4. Böhm H-J, Klebe G, Kubinyi H (1996) Wirkstoffdesign. Der Weg zum Arzneimittel. Spektrum Akademischer Verlag, Heidelberg
5. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R et al. (2014) QSAR Modeling: Where Have You Been? Where Are You Going To? J. Med. Chem. 57(12): 4977–5010. DOI:10.1021/jm4004285
6. Klebe G (2009) Entwurf und Wirkung von Arzneistoffen. Spektrum Akademischer Verlag, Heidelberg
7. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics, 2nd ed. Wiley, Weinheim
8. Bawden D (1983) Computerized chemical structure-handling techniques in structure-activity studies and molecular property prediction. J. Chem. Inf. Model. 23(1): 14–22. DOI:10.1021/ci00037a003
9. Fujita T, Ban T (1971) Structure-activity relation. 3. Structure-activity study of phenethylamines as substrates of biosynthetic enzymes of sympathetic transmitters. J. Med. Chem. 14(2): 148–152. DOI:10.1021/jm00284a016
10. Chu KC, Feldmann RJ, Shapiro MB, Hazard GF, Geran RI (1975) Pattern recognition and structure-activity relation studies. Computer-assisted prediction of antitumor activity in structurally diverse drugs in an experimental mouse brain tumor system. J. Med. Chem. 18(6): 539–545. DOI:10.1021/jm00240a001
11. Hodes L, Hazard GF, Geran RI, Richman S (1977) A statistical-heuristic method for automated selection of drugs for screening. J. Med. Chem. 20(4): 469–475. DOI:10.1021/jm00214a002

12. Devillers J, Balaban AT (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon & Breach, Amsterdam
13. Katritzky AR, Gordeeva EV (1993) Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. J. Chem. Inform. Comput. Sci. 33(6): 835–857
14. Hansen PJ, Jurs PC (1988) Chemical applications of graph theory. Part I. Fundamentals and topological indices. J. Chem. Educ. 65(7): 574. DOI:10.1021/ed065p574
15. Hastie T, Tibshirani R, Friedman JH The elements of statistical learning. Data mining, inference, and prediction, 2nd ed. Springer , New York
16. James G (2013) An introduction to statistical learning. With applications in R. Springer, New York
17. Briscoe E, Feldman J (2011) Conceptual complexity and the bias/variance tradeoff. Cognition 118(1): 2–16. DOI:10.1016/j.cognition.2010.10.004
18. Aptula AO, Jeliazkova NG, Schultz TW, Cronin, Mark T. D. (2005) The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set? QSAR Comb. Sci. 24(3): 385–396. DOI:10.1002/qsar.200430909
19. Breiman L (1993) Classification and regression trees. Chapman & Hall, New York
20. Geisser S (1975) The Predictive Sample Reuse Method with Applications. J. Am. Stat. Assoc. 70(350): 320. DOI:10.2307/2285815
21. Baumann K (2003) Cross-validation as the objective function for variable-selection techniques. Trends Anal. Chem. 22(6): 395–406. DOI:10.1016/S0165-9936(03)00607-1
22. Filzmoser P, Gschwandtner M, Todorov V (2012) Review of sparse methods in regression and classification with application to chemometrics. J. Chemometr. 26(3-4): 42–51. DOI:10.1002/cem.1418
23. Dreger C, Kosfeld R, Eckey H-F (2014) Ökonometrie. Springer Fachmedien, Wiesbaden
24. Naes T, Mevik B-H (2001) Understanding the collinearity problem in regression and discriminant analysis. J. Chemometr. 15(4): 413–426. DOI:10.1002/cem.676
25. Tu Y-K, Clerehugh V, Gilthorpe MS (2004) Collinearity in linear regression is a serious problem in oral health research. Eur. J. Oral. Sci. 112(5): 389–397. DOI:10.1111/j.1600-0722.2004.00160.x

26. Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12(1): 55–67. DOI:10.1080/00401706.1970.10488634
27. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. J. Royal Statistical Soc. B 58: 267–288
28. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J. Royal Statistical Soc. B 67(2): 301–320. DOI:10.1111/j.1467-9868.2005.00503.x
29. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2(1-3): 37–52. DOI:10.1016/0169-7439(87)80084-9
30. Wold H (2004) Partial Least Squares. Encyclopedia of Statistical Sciences. Wiley, New York
31. Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemometrics Intell. Lab. Sys. 58(2): 109–130. DOI:10.1016/S0169-7439(01)00155-1
32. Varmuza K, Filzmoser P (2009) Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton
33. Chun H, Keles S (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. Royal Statistical Soc. B 72(1): 3–25. DOI:10.1111/j.1467-9868.2009.00723.x
34. Jong S de (1993) SIMPLS: An alternative approach to partial least squares regression. Chemometrics Intell. Lab. Sys. 18(3): 251–263. DOI:10.1016/0169-7439(93)85002-X
35. Jolliffe IT, Trendafilov NT, Uddin M (2003) A Modified Principal Component Technique Based on the LASSO. J. Comput. Graph. Stat. 12(3): 531–547. DOI:10.1198/1061860032148
36. Zou H, Hastie T, Tibshirani R (2006) Sparse Principal Component Analysis. J. Comput. Graph. Stat. 15(2): 265–286. DOI:10.1198/106186006X113430
37. Hand DJ, Mannila H, Smyth P (2001) Principles of data mining. Adaptive computation and machine learning. MIT Press, Cambridge
38. Kotsiantis SB, Zaharakis ID, Pintelas PE (2006) Machine learning: a review of classification and combining techniques. Artif. Intell. Rev. 26: 159. DOI:10.1007/s10462-007-9052-3

39. Baumann D, Baumann K (2014) Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. J. Chemoinform. 6(1): 47. DOI:10.1186/s13321-014-0047-1
40. Mathea M, Klingspohn W, Baumann K (2016) Chemoinformatic Classification Methods and their Applicability Domain. Mol. Inf. 35(5): 160–180. DOI:10.1002/minf.201501019
41. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30(7): 1145–1159. DOI:10.1016/S0031-3203(96)00142-2
42. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit. Lett. 27(8): 861–874. DOI:10.1016/j.patrec.2005.10.010
43. Provost F (2000) Machine learning from inbalanced data sets. Proc. AAAI workshop on learning from inbalanced data sets. 625–632
44. Kubinyi H (1998) Similarity and Dissimilarity: A Medicinal Chemist's View. Perspectives in Drug Discovery and Design 9/11: 225–252. DOI:10.1023/A:1027221424359
45. Kauffman GW, Jurs PC (2001) QSAR and k-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. J. Chem. Inform. Comput. Sci. 41(6): 1553–1560
46. Ajmani S, Jadhav K, Kulkarni SA (2006) Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. J. Chem. Inf. Model. 46(1): 24–31. DOI:10.1021/ci0501286
47. Zheng W, Tropsha A (2000) Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. J. Chem. Inf. Model. 40(1): 185–194. DOI:10.1021/ci980033m
48. Krzanowski WJ (2000) Principles of multivariate analysis. A user's perspective. Oxford University Press, Oxford
49. Mahalanobis PC (1936) On the generalised distance in statistics. Proc. Nat. Inst. Sci. 2(1): 49–55
50. Krause EF (1987) Taxicab Geometry. An adventure in non-Euclidean geometry. Dover Publications, New York
51. Rogers DJ, Tanimoto TT (1960) A Computer Program for Classifying Plants. Science 132(3434): 1115–1118. DOI:10.1126/science.132.3434.1115

52. Leach AR, Gillet VJ (2007) An introduction to chemoinformatics. Springer, London
53. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat. Rev. Cancer 8(1): 37–49. DOI:10.1038/nrc2294
54. Aggarwal CC (2001) Re-designing distance functions and distance-based applications for high dimensional data. SIGMOD Rec. 30(1): 13–18. DOI:10.1145/373626.373638
55. Aggarwal CC, Hinneburg A, Keim DA (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. Springer, Berlin
56. Aggarwal CC (2013) Outlier analysis. Springer, New York
57. Zimek A, Schubert E, Kriegel H-P (2012) A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analy. Data Mining 5(5): 363–387. DOI:10.1002/sam.11161
58. Rupp M, Schneider P, Schneider G (2009) Distance phenomena in high-dimensional chemical descriptor spaces: Consequences for similarity-based approaches. J. Comput. Chem.: NA. DOI:10.1002/jcc.21218
59. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When Is "Nearest Neighbor" Meaningful? Springer, Berlin
60. Willett P, Barnard JM, Downs GM (1998) Chemical Similarity Searching. J. Chem. Inf. Model. 38(6): 983–996. DOI:10.1021/ci9800211
61. Todeschini R, Consonni V, Xiang H, Holliday J, Buscema M, Willett P (2012) Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets. J. Chem. Inf. Model. 52(11): 2884–2901. DOI:10.1021/ci300261r
62. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Chemoinform. 7: 20. DOI:10.1186/s13321-015-0069-3
63. Breiman L (2001) Random Forests. Mach. Learn. 45(1): 5–32. DOI:10.1023/A:1010933404324
64. Svetnik V, Liaw A, Tong C, Wang T (2004) Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. Springer, Berlin

65. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Model. 43(6): 1947–1958. DOI:10.1021/ci034160g
66. Raschka S (2017) Machine Learning mit Python. Das Praxis-Handbuch für Data Science, Predictive Analytics und Deep Learning, mitp, Frechen
67. Burges CJ (1998) A Tutorial on Support Vector Machines for Pattern Recognition. Data Min. Knowl. Discov. 2(2): 121–167. DOI:10.1023/A:1009715923555
68. Cortes C, Vapnik V (1995) Machine Learning. Mach. Learn. 20(3): 273–297. DOI:10.1023/A:1022627411411
69. Schölkopf B, Smola AJ (2002) Learning with kernels. Support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning. MIT Press, Cambridge
70. Müller KR, Mika S, Ratsch G, Tsuda K, Schölkopf B (2001) An introduction to kernel-based learning algorithms. IEEE Trans. Neural Netw. 12(2): 181–201. DOI:10.1109/72.914517
71. Murphy KP (2012) Machine learning. A probabilistic perspective. Adaptive computation and machine learning series. MIT Press, Cambridge
72. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
73. Hinton GE (1986) Learning distributed representation of concepts. Proc. 8th Ann. Conf. CogSci 1–12
74. Werbos PJ (1982) Applications of advances in nonlinear sensitivity analysis. Springer, Berlin
75. Lippmann RP (1989) Pattern classification using neural networks. IEEE Commun. Mag. 27(11): 47–63. DOI:10.1109/35.41401
76. Rokach L (2010) Pattern classification using ensemble methods. Series in machine perception and artificial intelligence. World Scientific Pub. Co., Singapore
77. Breiman L (1996) Bagging predictors. Mach. Learn. 24(2): 123–140. DOI:10.1007/BF00058655
78. Duda RO, Hart PE, Stork DG (2001) Pattern classification, Wiley, New York
79. Domingos P, Pazzani M (1997) On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Mach. Learn. 29(2/3): 103–130. DOI:10.1023/A:1007413511361

80. Barker M, Rayens W (2003) Partial least squares for discrimination. J. Chemometr. 17(3): 166–173. DOI:10.1002/cem.785
81. Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7(1): 1–26. DOI:10.1214/aos/1176344552
82. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Machine Intell. 20(8): 832–844. DOI:10.1109/34.709601
83. Skurichina M, Duin RPW (2002) Bagging, Boosting and the Random Subspace Method for Linear Classifiers. Pattern Anal. Appl. 5(2): 121–135. DOI:10.1007/s100440200011
84. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA et al. (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. Altern. Lab .Anim. 33(2): 155–173
85. Hodge VJ, Austin Jim (2004) A Survey of Outlier Detection Methodologies. Artif. Intell. Rev. 22: 85–126
86. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection. ACM Comput. Surv. 41(3): 1–58. DOI:10.1145/1541880.1541882
87. Markou M, Singh S (2003) Novelty detection: a review—part 1: statistical approaches. Signal Process. 83(12): 2481–2497. DOI:10.1016/j.sigpro.2003.07.018
88. Pimentel MA, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. Signal Process. 99: 215–249. DOI:10.1016/j.sigpro.2013.12.026
89. Vanderlooy S, Maaten L, Sprinkhuizen-Kuyper I (2007) Off-Line Learning with Transductive Confidence Machines: An Empirical Evaluation. Springer, Berlin
90. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R (2004) Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environ. Health Perspect. DOI:10.1289/txg.7125
91. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J, Gramatica P, Hansen K, Schroeter T, Müller K-R, Xi L, Liu H, Yao X, Öberg T, Hormozdiari F, Dao P, Sahinalp C, Todeschini R, Polishchuk P, Artemenko A, Kuz'min V, Martin TM, Young DM, Fourches D, Muratov E, Tropsha A, Baskin I, Horvath D, Marcou G, Muller C, Varnek A, Prokopenko VV, Tetko IV (2010) Applicability domains for classification

problems: benchmarking of distance to models for Ames mutagenicity set. J. Chem. Inf. Model. 50:2094–2111. DOI: 10.1021/ci100253r

92. Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, Tetko IV (2010) Applicability domain for in silico models to achieve accuracy of experimental measurements. J. Chemometr. 24(3-4): 202–208. DOI:10.1002/cem.1296
93. Dragos H, Gilles M, Alexandre V (2009) Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J. Chem. Inf. Model. 49(7): 1762–1776. DOI:10.1021/ci9000579
94. Ferri C, Hernández-Orallo J, Modroiu R (2009) An experimental comparison of performance measures for classification. Pattern Recogn. Lett. 30(1): 27–38. DOI:10.1016/j.patrec.2008.08.010
95. Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. Proc. 22nd Int. Conf. Mach. Learn., ICML. 625–632
96. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers: 61–72
97. Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. Proc. 18th Int. Conf. Mach. Learn.,ICML. 609-616
98. Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. Proc. 8th Int. Conf. Knowl. Discov.Data Mining, KDD. 694–699.
99. Robertson T, Wright FT, Dykstra R (1988) Order restricted statistical inference. Wiley, New York
100. Ayer M, Brunk HD, Ewing GM, Reid WT, Silverman E (1955) An Empirical Distribution Function for Sampling with Incomplete Information. Ann. Math. Statist. 26(4): 641–647
101. DeGroot M, Fienberg S (1982) The comparison and evaluation of forecasters. Statistician(32): 12–22
102. Caruana R, Niculescu-Mizil A (2004) Data mining in metric space. Proc. ACM SIGKDD. 69
103. Papadopoulos H, Vovk V, Gammerman A (2015) Guest editors' preface to the special issue on conformal prediction and its applications. Ann. Math. Artif. Intell. 74(1-2): 1–7. DOI:10.1007/s10472-014-9429-3

104. Vovk V, Gammerman A, Shafer G (2005) Algorithmic learning in a random world. Springer, New York
105. Eklund M, Norinder U, Boyer S, Carlsson L (2013) The application of conformal prediction to the drug discovery process. Ann. Math. Artif. Intell. DOI:10.1007/s10472-013-9378-2
106. Norinder U, Carlsson L, Boyer S, Eklund M (2014) Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. J. Chem. Inf. Model. 54(6): 1596–1603. DOI:10.1021/ci5001168
107. Vovk V (2013) Conditional validity of inductive conformal predictors. Mach. Learn. 92(2-3): 349–376. DOI:10.1007/s10994-013-5355-6
108. Vovk V (2015) Cross-conformal predictors. Ann. Math. Artif. Intell. 74(1-2): 9–28. DOI:10.1007/s10472-013-9368-4
109. Klingspohn W, Mathea M, Ter Laak A, Heinrich N, Baumann K (2017) Efficiency of different measures for defining the applicability domain of classification models. J. Cheminf. 9(1): 1616. DOI:10.1186/s13321-017-0230-2
110. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A (2012) Probability machines: consistent probability estimation using nonparametric learning machines. Methods Inf. Med. 51(1): 74–81. DOI:10.3414/ME00-01-0052
111. Devroye L, Györfi L, Lugosi G (1996) A probabilistic theory of pattern recognition. Springer, New York
112. Simon R (2014) Class probability estimation for medical studies. Biom. J. 56(4): 597–600. DOI:10.1002/bimj.201300296
113. Chang C-C, Lin C-JL LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol. 2:1–27. DOI:10.1145/1961189.1961199
114. Brier GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review(78): 1–3
115. Kattan MW, Cowen ME (2009) Encyclopedia of medical decision making. SAGE Publications, Thousand Oaks
116. Hand DJ (1997) Construction and assessment of classification rules. Wiley, New York
117. Zadrozny B, Elkan C (2001) Learning and making decisions when costs and probabilities are both unknown. Proc. 7th ACM SIGKDD. 204–213

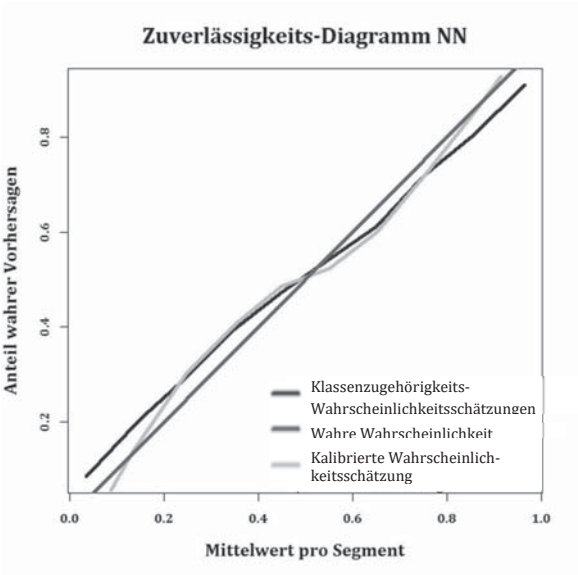
118. Good IJ (1952) Rational Decisions. J. Royal Stat. Soc. B 14(1): 107–114
119. Good IJ (1968) Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor. Br. J. Philos. Sci. 19(2): 123–143. DOI:10.1093/bjps/19.2.123
120. Piatetsky-Shapiro G, Masand B (1999) Estimating campaign benefits and modeling lift. Proc. 5th ACM SIGKDD. 185–193
121. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1): 29–36. DOI:10.1148/radiology.143.1.7063747
122. Provost F, Domingos P (2000) Well-Trained PETs: Improving Probability Estimation Trees. Technical Report CDER(#00-04-IS)
123. Caruana R, Karampatziakis N, Yessenalina A (2008) An empirical evaluation of supervised learning in high dimensions. Proc. 25th Int. Conf. Mach. Learn, ICML. 96–103
124. Gavin C, Cawley NLC, Talbot G (2010) On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research 11: 2079–2107
125. Dietterich T, Jain A, Lathrop R, Lozano-Perez T (1994) A comparison of dynamic reposing and tangent distance for drug activity prediction. Advances in Neural Information Processing Systems 6. 216–223
126. Alcalá-Fdez J, Fernández A, Luengo J, Derrac J, García S, Sánchez L, Herrera F (2011) KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework 17(2-3): 255–287
127. Mansouri K, Ringsted T, Ballabio D, Todeschini R, Consonni V (2013) Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals. J. Chem. Inf. Model. 53(4): 867–878. DOI:10.1021/ci4000213
128. Asuncion A, Newman DJ (2007) UCI Machine Learning Repository
129. Doniger S, Hofmann T, Yeh J (2002) Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. J. Comput. Biol. 6: 849–864. DOI:10.1089/10665270260518317
130. Symyx (2005) MACCS Structural Keys, MDL Information Systems Inc., San Ramon
131. Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, http://www.chemcomp.com/

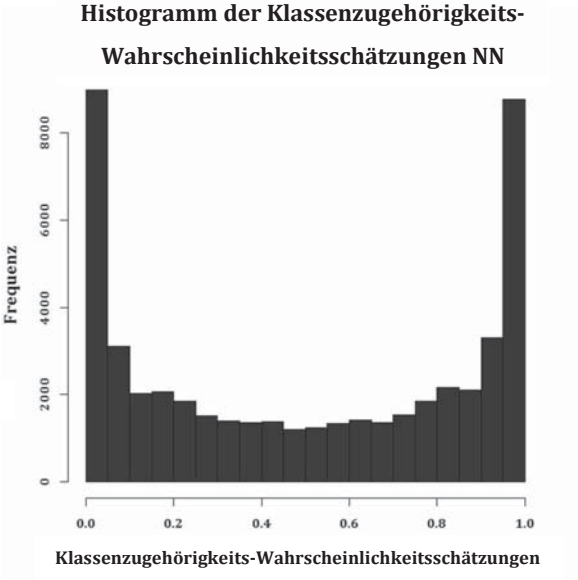
132. Chow C (1970) On optimum recognition error and reject tradeoff. IEEE Trans. Inform. Theory 16(1): 41–46. DOI:10.1109/TIT.1970.1054406
133. Devetyarov D, Nouretdinov I (2010) Prediction with Confidence Based on a Random Forest Classifier. Springer, Berlin
134. Isidro C conformal. https://cran.r-project.org/web/packages/conformal/conformal.pdf
135. Hung MS, Hu MY, Shanker MS, Patuwo BE (1996) Estimating posterior probabilities in classification problems with neural networks. International Journal of Computational Intelligence and Organizations, 1(1): 49–60
136. Brereton RG, Lloyd GR (2014) Partial least squares discriminant analysis: taking the magic away. J. Chemometr. 28(4): 213–225. DOI:10.1002/cem.2609
137. Provost F, Domingos P (2003) Tree Induction for Probability-Based Ranking. Mach. Learn. 52(3): 199–215. DOI:10.1023/A:1024099825458
138. Bostrom H (2007) Estimating class probabilities in random forests. Proc. 6th Int. Conf. Mach. Learn. Appl. 211–216
139. Moreno-Seco F, Iñesta JM, León PJP de, Micó L (2006) Comparison of Classifier Fusion Methods for Classification in Pattern Recognition Tasks. Springer, Berlin
140. Dietterich TG (2000) Ensemble Methods in Machine Learning. Springer, Berlin
141. Bella A, Ferri C, Hernández-Orallo J, Ramírez-Quintana MJ (2013) On the effect of calibration in classifier combination. Appl. Intell. 38(4): 566–585. DOI:10.1007/s10489-012-0388-2
142. Shafer G, Vovk V A (2008) Tutorial on Conformal Prediction. The Journal of Machine Learning Research. 9:371-421
143. Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. Proc. 23rd Int. Conf. Mach. Learn. 161–168. DOI: 10.1145/1143844.1143865
144. Wallace BC, Dahabreh IJ (2012) Class Probability Estimates are Unreliable for Imbalanced Data (and How to Fix Them). Proc. 12th Int. Conf. Data Mining, IEEE. 695–704
145. Wallace BC, Dahabreh IJ (2014) Improving class probability estimates for imbalanced data. Knowl. Inf. Syst. 41(1): 33–52. DOI:10.1007/s10115-013-0670-6
146. Carlsson L, Eklund M, Norinder U (2014) Aggregated Conformal Prediction. Springer, Berlin

9 Anhang

9.1 Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern

9.1.1 Visuelle Analyse der Zuverlässigkeits-Diagramme und Histogramme von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerten unterschiedlicher Klassifikations-und Regressionstechniken vor und nach Kalibrierung





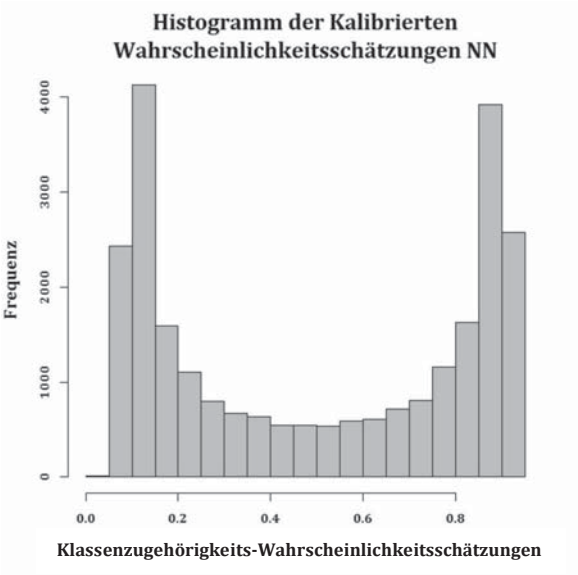
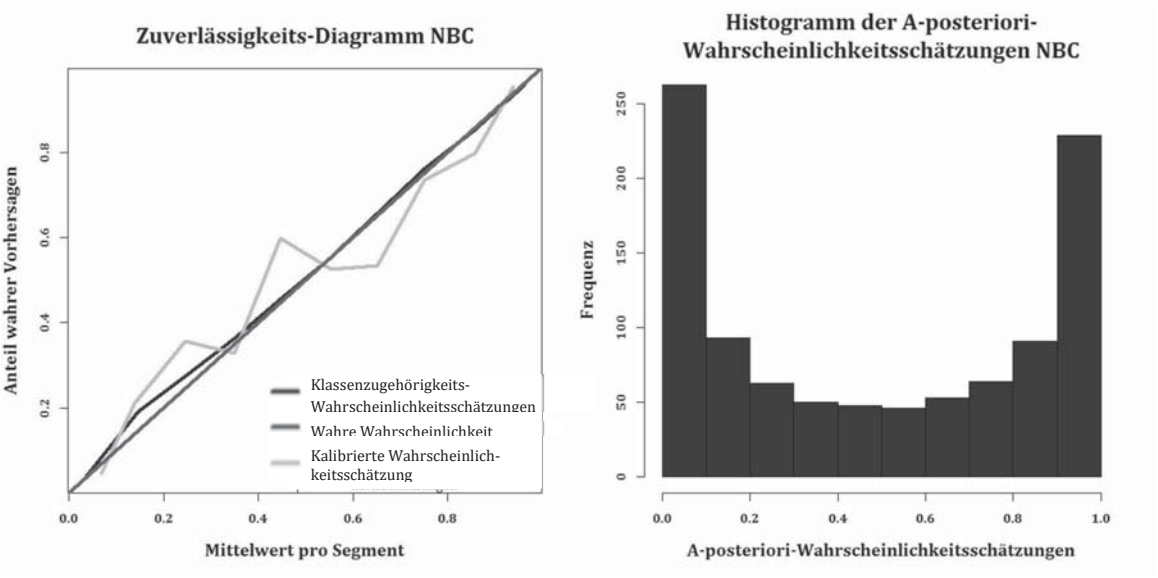


Abbildung 60: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der NNs. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Es ist zu erkennen, dass sich die unkalibrierten Schätzer bereits sehr nah an der wahren Wahrscheinlichkeit befinden. Logistische Regression führt zu keiner Verbesserung der Schätzwerte. Dies geht sowohl aus den Zuverlässigkeits-Diagrammen, als auch aus den Histogrammen hervor.



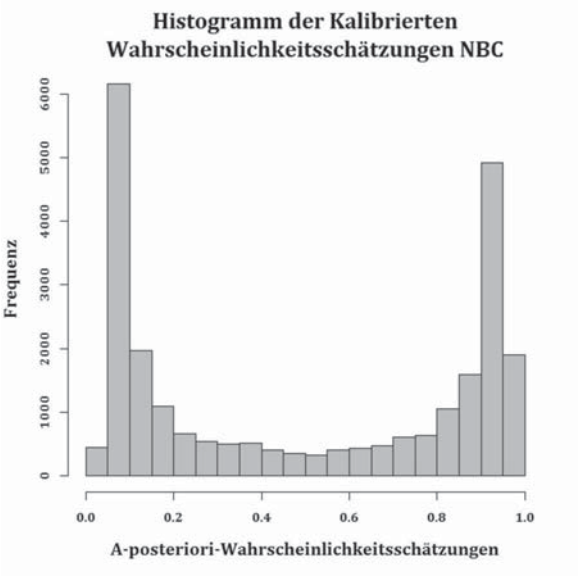
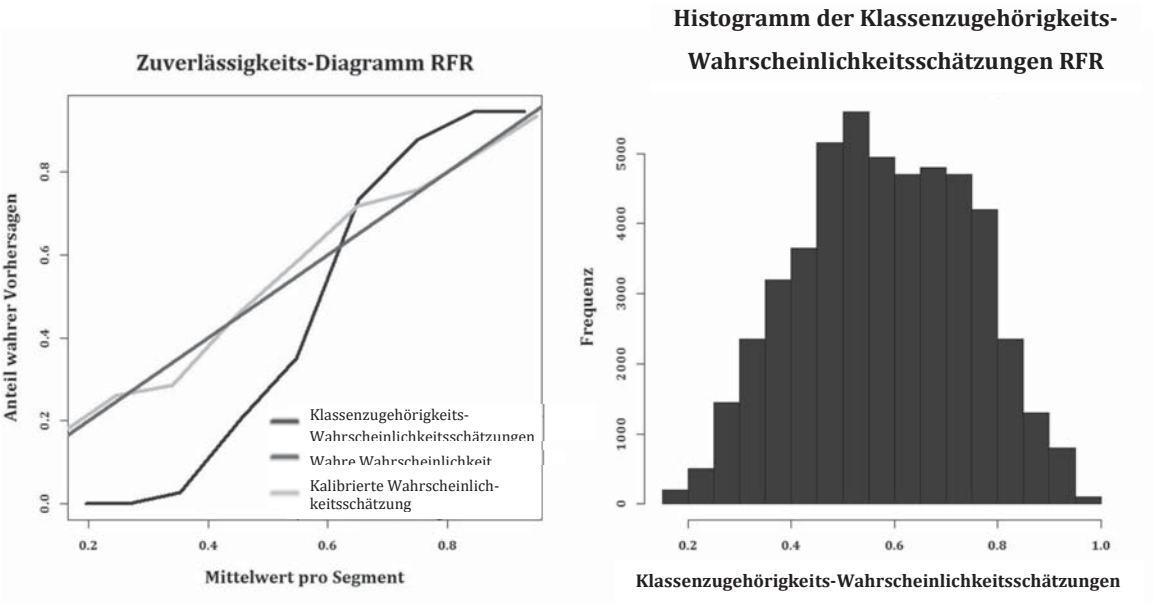


Abbildung 61: Kalibrierte und unkalibrierte A-posteriori-Wahrscheinlichkeitsschätzer des NBC. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Es ist zu erkennen, dass sich die unkalibrierten Schätzer bereits sehr nah an der wahren Wahrscheinlichkeit befinden. Logistische Regression führt zu keiner Verbesserung der Schätzwerte. Dies geht sowohl aus den Zuverlässigkeits-Diagrammen, als auch aus den Histogrammen hervor.



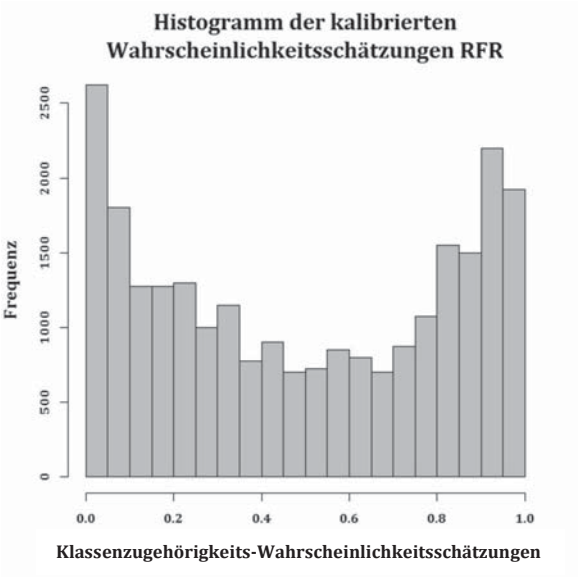
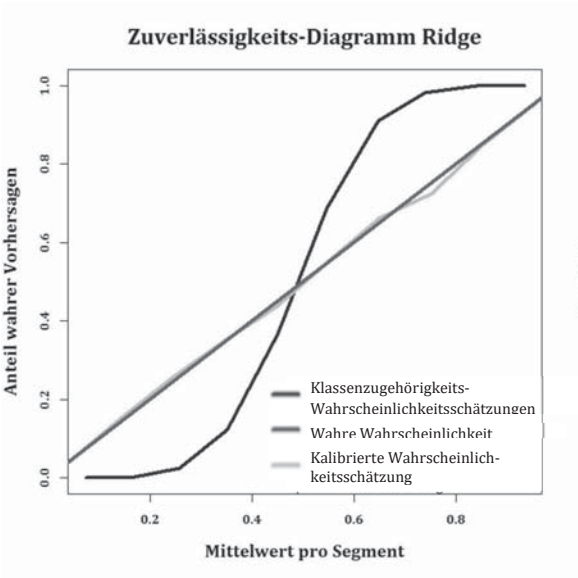
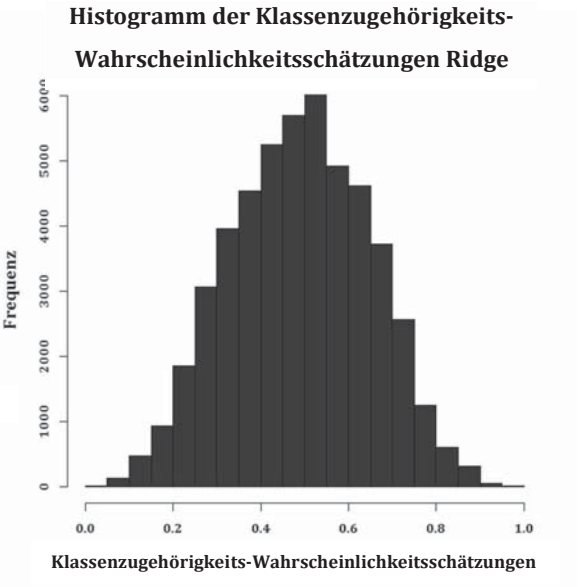


Abbildung 62: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des RFR. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert wird. Bei den unkalibrierten Schätzwerten handelt es sich um die kontinuierlichen Vorhersagen der Regressionstechnik. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.





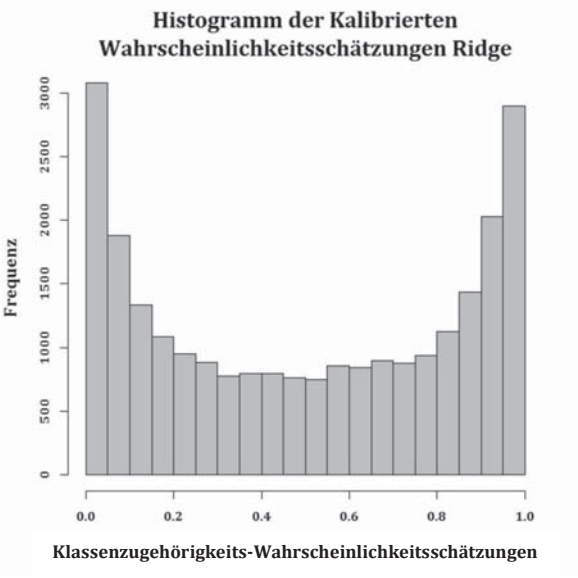
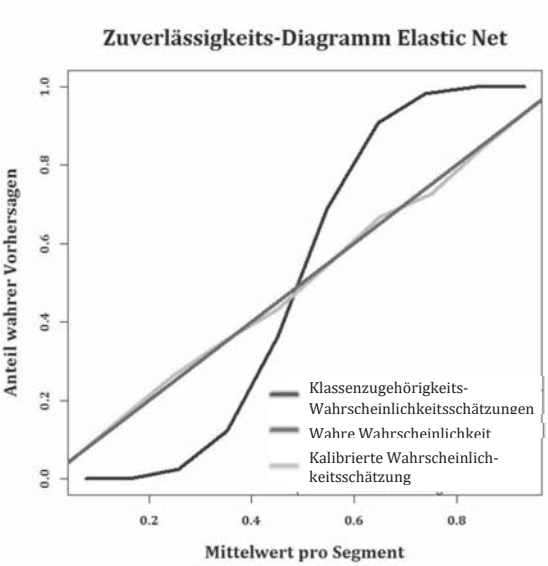
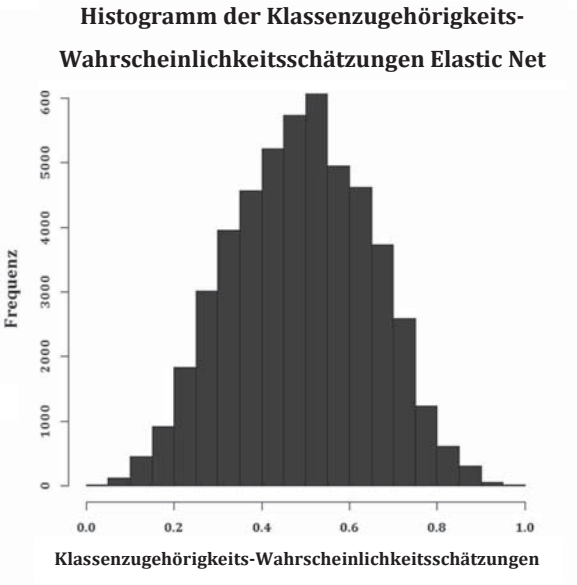


Abbildung 63: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der Ridge. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert wird. Bei den unkalibrierten Schätzwerten handelt es sich um die kontinuierlichen Vorhersagen der Regressionstechnik, welche skaliert wurden, sodass diese zwischen 0 und 1 liegen. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.





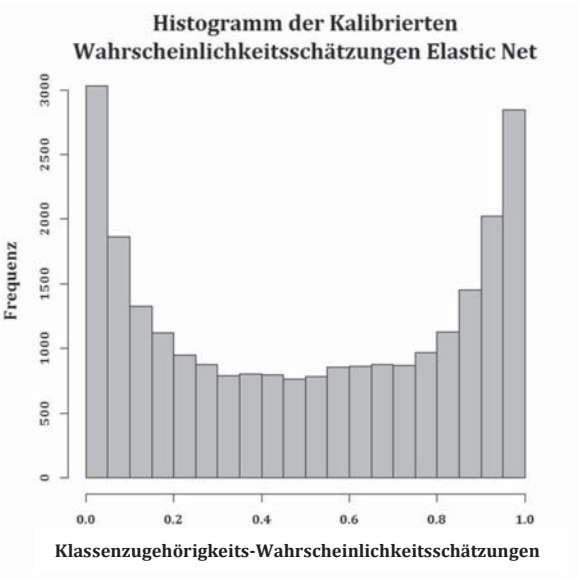
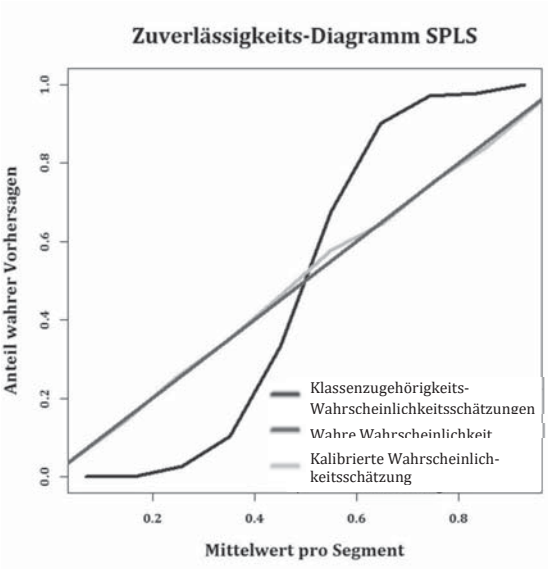
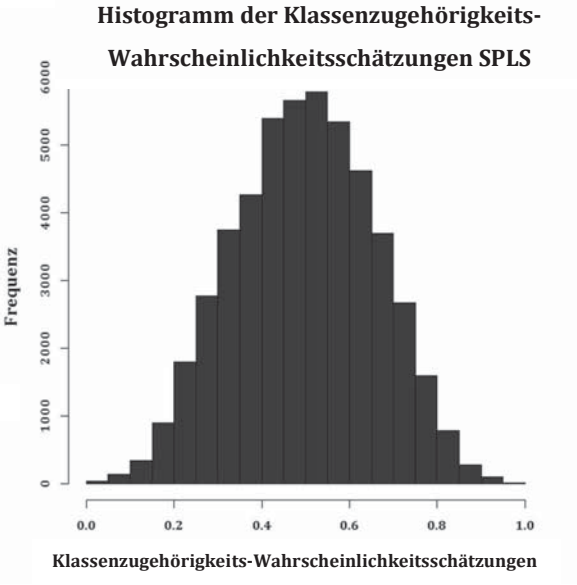


Abbildung 64: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des Elastic Net. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert wird. Bei den unkalibrierten Schätzwerten handelt es sich um die kontinuierlichen Vorhersagen der Regressionstechnik, welche skaliert wurden, sodass diese zwischen 0 und 1 liegen. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.





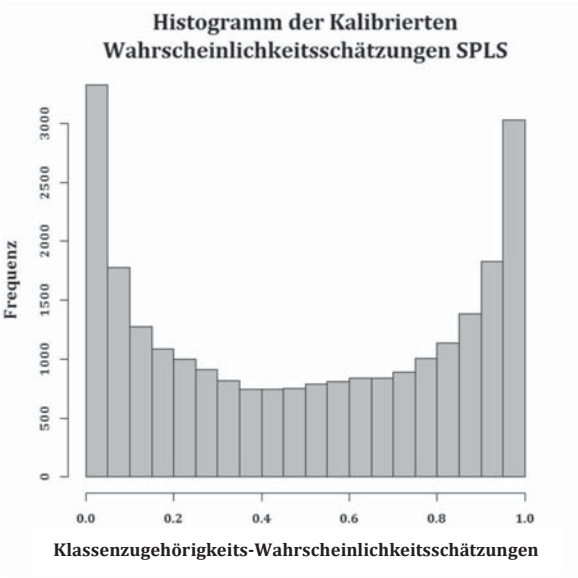


Abbildung 65: Kalibrierte und unkalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der SPLS. Die Simulationsstudie wurde mit 2000 Objekten durchgeführt und die Korrektklassifizierungsrate wurde auf 0.8 eingestellt. Auf dem Zuverlässigkeits-Diagramm ist zu erkennen, dass die unkalibrierten Schätzer einen sigmoiden Verlauf zeigen, welcher durch logistische Regression der wahren Wahrscheinlichkeit angenähert wird. Bei den unkalibrierten Schätzwerten handelt es sich um die kontinuierlichen Vorhersagen der Regressionstechnik, welche skaliert wurden, sodass diese zwischen 0 und 1 liegen. Die Histogramme zeigen, dass die unkalibrierten Schätzer zur Mitte verschoben sind, was durch logistische Regression wieder korrigiert wird.

9.1.2 Vorversuch: Einfluss der Variablenanzahl des Datensatzes auf den Fehler sowie Beurteilung der Fehlermaße

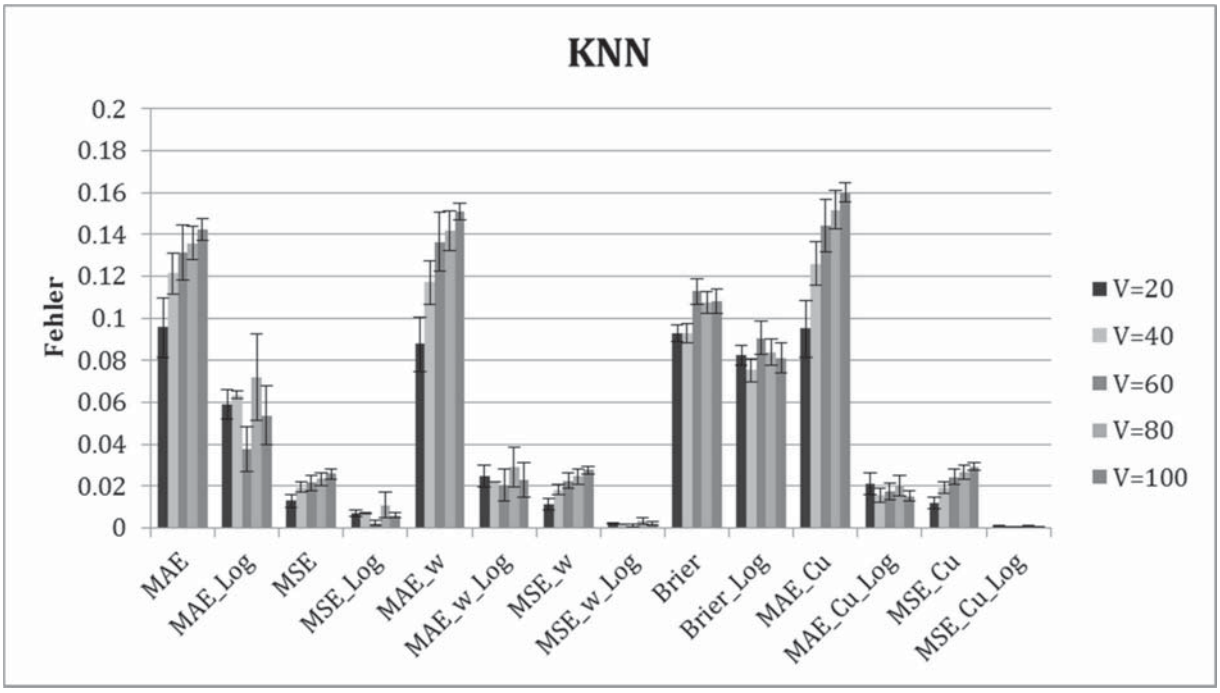


Abbildung 66: Ergebnisse der Simulationsstudie, welche den Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des KNN untersucht hat. Bei unkorrelierten Daten steigt der Fehler vor logistischer Regression mit Zunahme der Variablenanzahl an. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

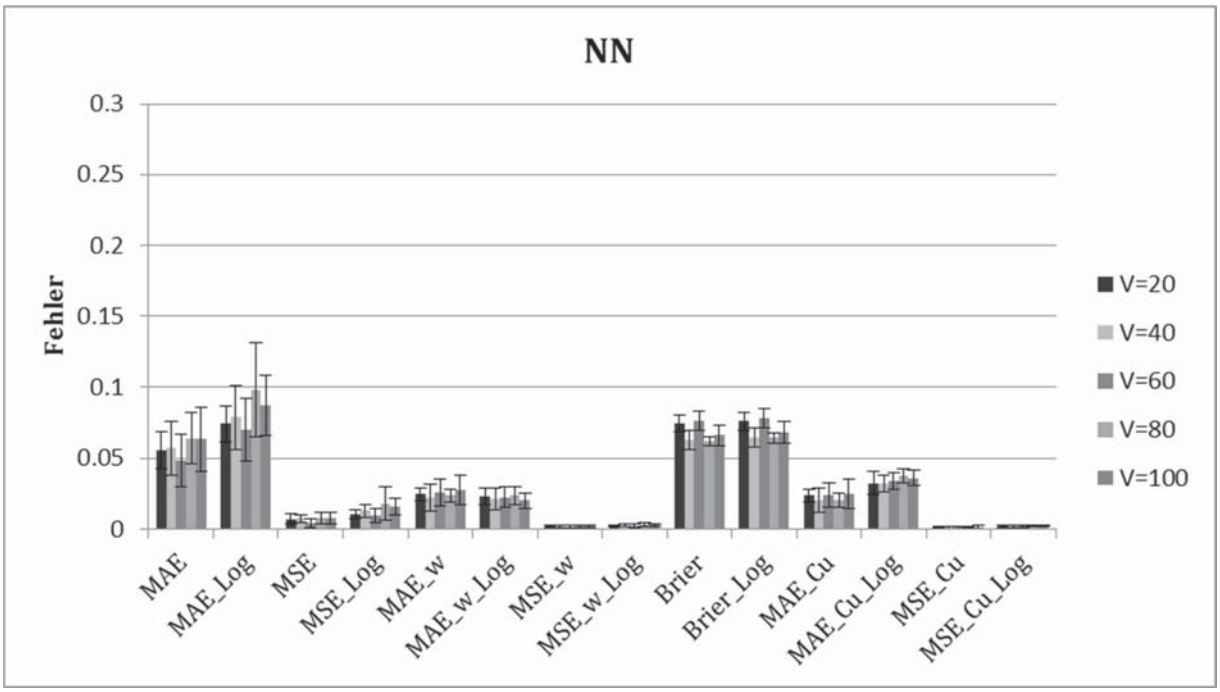


Abbildung 67: Ergebnisse der Simulationsstudie, welche den Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der NN untersucht hat. Es ist keine Zunahme des Fehlers mit steigender Variablenanzahl zu erkennen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

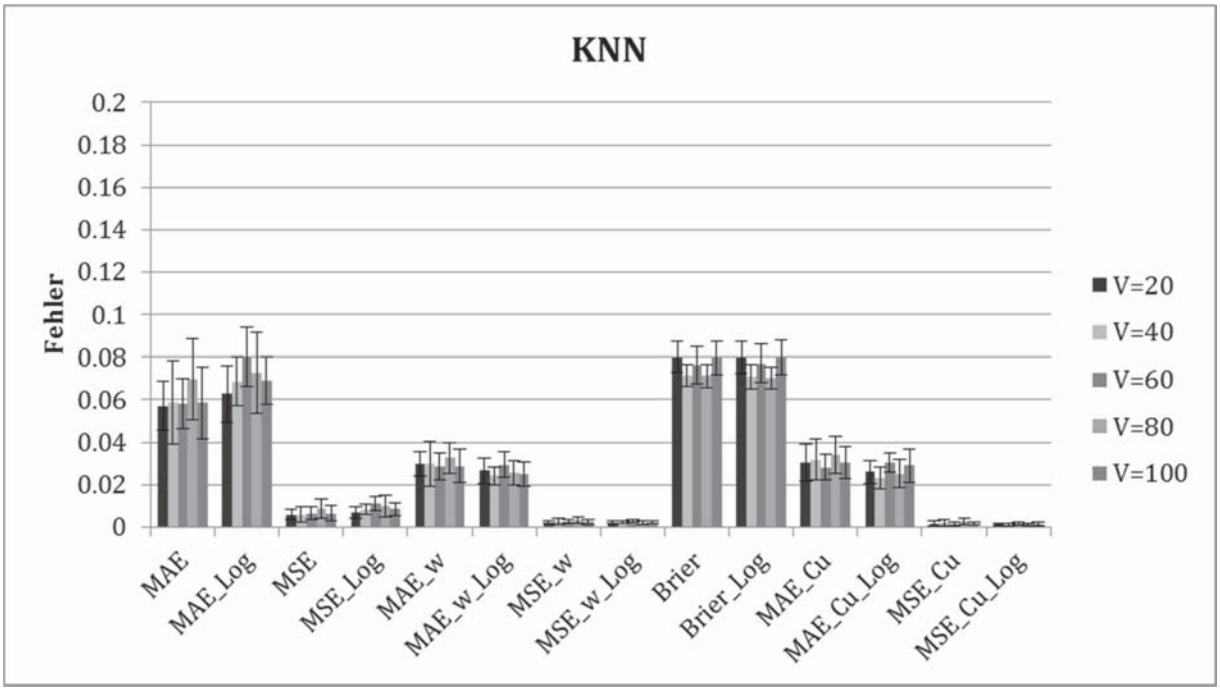


Abbildung 68: Ergebnisse der Simulationsstudie, welche den Einfluss der Variablenanzahl auf die Exaktheit der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer des KNN mit korrelierten Daten untersucht hat. Es ist keine Zunahme des Fehlers mit steigender Variablenanzahl zu erkennen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie die Standardabweichung der Einzelwerte.

Tabelle 7: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.58). Aufgelistet sind die MW und die STD der Fehler aus zehn wiederholten Versuchen(*).

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1539	0.0417	0.0279	0.0037	0.1712	0.0196	0.0329	0.0013	0.1154	0.0825	0.1790	0.0180	0.0358	0.0006
STD		0.0074	0.0112	0.0023	0.0023	0.0082	0.0034	0.0027	0.0006	0.0038	0.0046	0.0071	0.0042	0.0028	0.0003

Tabelle 7: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1850	0.0410	0.0406	0.0030	0.2236	0.0187	0.0550	0.0011	0.1326	0.0757	0.2322	0.0172	0.0599	0.0006
STD		0.0081	0.0104	0.0029	0.0012	0.0106	0.0052	0.0038	0.0004	0.0025	0.0060	0.0098	0.0050	0.0040	0.0003

Tabelle 8: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1912	0.0452	0.0483	0.0044	0.2612	0.0185	0.0710	0.0014	0.1473	0.0730	0.2646	0.0169	0.0766	0.0006
STD		0.0054	0.0170	0.0024	0.0033	0.0091	0.0067	0.0042	0.0011	0.0027	0.0063	0.0082	0.0038	0.0044	0.0003

Tabelle 9: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1972	0.0500	0.0515	0.0048	0.2685	0.0216	0.0756	0.0016	0.1590	0.0784	0.2719	0.0182	0.0825	0.0007
STD		0.0044	0.0109	0.0023	0.0023	0.0076	0.0038	0.0040	0.0006	0.0018	0.0044	0.0076	0.0057	0.0041	0.0005

Tabelle 10: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.2066	0.0410	0.0560	0.0035	0.2837	0.0189	0.0833	0.0013	0.1676	0.0797	0.2838	0.0175	0.0893	0.0006
STD		0.0082	0.0116	0.0042	0.0023	0.0171	0.0047	0.0089	0.0008	0.0019	0.0073	0.0125	0.0044	0.0065	0.0003

Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	:								MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0858	0.0473	0.0094	0.0044	0.0745	0.0186	0.0073	0.0015	0.0845	0.0785	0.0781	0.0218	0.0077	0.0008
STD		0.0117	0.0128	0.0023	0.0024	0.0084	0.0050	0.0017	0.0008	0.0056	0.0070	0.0091	0.0044	0.0019	0.0003

Tabelle 12: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.15). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0816	0.0438	0.0089	0.0037	0.0691	0.0189	0.0066	0.0013	0.0860	0.0805	0.0735	0.0214	0.0070	0.0008
STD		0.0063	0.0091	0.0013	0.0014	0.0047	0.0042	0.0008	0.0005	0.0038	0.0042	0.0054	0.0031	0.0008	0.0002

Tabelle 13: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.15). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0758	0.0601	0.0078	0.0068	0.0679	0.0239	0.0063	0.0021	0.0847	0.0800	0.0719	0.0251	0.0068	0.0012
STD		0.0155	0.0167	0.0027	0.0033	0.0121	0.0054	0.0020	0.0009	0.0064	0.0084	0.0135	0.0043	0.0021	0.0003

Tabelle 14: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.15). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0911	0.0500	0.0110	0.0052	0.0764	0.0195	0.0081	0.0015	0.0806	0.0736	0.0811	0.0214	0.0086	0.0009
STD		0.0094	0.0095	0.0021	0.0023	0.0050	0.0046	0.0010	0.0005	0.0036	0.0038	0.0058	0.0062	0.0012	0.0004

Tabelle 15: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.13). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0858	0.0473	0.0094	0.0044	0.0745	0.0186	0.0073	0.0015	0.0845	0.0785	0.0781	0.0218	0.0077	0.0008
STD		0.0123	0.0135	0.0024	0.0026	0.0088	0.0052	0.0018	0.0009	0.0059	0.0073	0.0096	0.0046	0.0020	0.0003

Tabelle 16: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0957	0.0589	0.0129	0.0068	0.0875	0.0250	0.0113	0.0021	0.0927	0.0823	0.0949	0.0210	0.0119	0.0008
STD		0.0141	0.0070	0.0031	0.0015	0.0126	0.0052	0.0027	0.0003	0.0039	0.0044	0.0134	0.0051	0.0030	0.0004

Tabelle 17: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1212	0.0635	0.0195	0.0070	0.1169	0.0220	0.0183	0.0017	0.0927	0.0753	0.1260	0.0155	0.0194	0.0004
STD		0.0096	0.0018	0.0025	0.0000	0.0107	0.0002	0.0026	0.0001	0.0044	0.0053	0.0106	0.0032	0.0027	0.0002

Tabelle 18: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.36). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.1312	0.0376	0.0216	0.0024	0.1365	0.0204	0.0226	0.0011	0.1128	0.0903	0.1444	0.0177	0.0243	0.0006
STD		0.0134	0.0107	0.0037	0.0014	0.0138	0.0078	0.0036	0.0005	0.0061	0.0080	0.0125	0.0039	0.0036	0.0003

Tabelle 19: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.1359	0.0718	0.0233	0.0109	0.1418	0.0290	0.0245	0.0033	0.1074	0.0837	0.1517	0.0204	0.0264	0.0008
STD		0.0078	0.0205	0.0029	0.0062	0.0095	0.0097	0.0034	0.0016	0.0051	0.0061	0.0090	0.0048	0.0032	0.0004

Tabelle 20: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.315). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1423	0.0536	0.0258	0.0061	0.1507	0.0231	0.0276	0.0020	0.1080	0.0810	0.1600	0.0152	0.0292	0.0004
STD		0.0052	0.0140	0.0025	0.0009	0.0040	0.0082	0.0019	0.0008	0.0057	0.0071	0.0045	0.0026	0.0019	0.0002

Tabelle 21: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0571	0.0628	0.0060	0.0069	0.0298	0.0265	0.0023	0.0023	0.0800	0.0799	0.0305	0.0260	0.0017	0.0012
STD		0.0114	0.0132	0.0024	0.0028	0.0059	0.0061	0.0008	0.0005	0.0073	0.0077	0.0087	0.0055	0.0010	0.0004

Tabelle 22: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0590	0.0687	0.0059	0.0083	0.0299	0.0242	0.0025	0.0021	0.0714	0.0710	0.0319	0.0232	0.0020	0.0010
STD		0.0196	0.0114	0.0035	0.0024	0.0104	0.0044	0.0015	0.0006	0.0052	0.0056	0.0095	0.0049	0.0013	0.0005

Tabelle 23: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0582	0.0801	0.0065	0.0114	0.0285	0.0293	0.0024	0.0030	0.0763	0.0771	0.0283	0.0304	0.0015	0.0018
STD		0.0119	0.0138	0.0030	0.0033	0.0064	0.0060	0.0009	0.0008	0.0088	0.0091	0.0059	0.0044	0.0007	0.0006

Tabelle 24: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0698	0.0726	0.0089	0.0100	0.0327	0.0255	0.0033	0.0022	0.0712	0.0703	0.0342	0.0252	0.0026	0.0012
STD		0.0191	0.0191	0.0044	0.0050	0.0073	0.0056	0.0014	0.0008	0.0057	0.0053	0.0089	0.0067	0.0017	0.0006

Tabelle 25: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0586	0.0691	0.0066	0.0085	0.0288	0.0249	0.0025	0.0023	0.0797	0.0801	0.0303	0.0291	0.0018	0.0015
STD		0.0170	0.0114	0.0034	0.0029	0.0078	0.0056	0.0011	0.0007	0.0079	0.0082	0.0076	0.0078	0.0008	0.0008

Tabelle 26: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1782	0.0494	0.0389	0.0051	0.2140	0.0214	0.0513	0.0017	0.1294	0.0768	0.2256	0.0203	0.0559	0.0008
STD		0.0054	0.0107	0.0027	0.0019	0.0064	0.0050	0.0034	0.0006	0.0066	0.0071	0.0076	0.0041	0.0036	0.0003

Tabelle 27: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1844	0.0458	0.0431	0.0040	0.2266	0.0204	0.0583	0.0013	0.1305	0.0712	0.2376	0.0178	0.0629	0.0006
STD		0.0068	0.0165	0.0037	0.0030	0.0149	0.0046	0.0069	0.0008	0.0062	0.0053	0.0156	0.0043	0.0074	0.0003

Tabelle 28: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.36). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1908	0.0493	0.0438	0.0049	0.2327	0.0190	0.0589	0.0014	0.1300	0.0705	0.2395	0.0168	0.0631	0.0005
STD		0.0039	0.0113	0.0021	0.0024	0.0079	0.0035	0.0038	0.0006	0.0058	0.0052	0.0068	0.0031	0.0039	0.0002

Tabelle 29: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.31). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1848	0.0476	0.0435	0.0043	0.2269	0.0215	0.0586	0.0015	0.1344	0.0743	0.2405	0.0191	0.0640	0.0007
STD		0.0092	0.0142	0.0036	0.0025	0.0158	0.0055	0.0060	0.0008	0.0050	0.0037	0.0132	0.0060	0.0067	0.0005

Tabelle 30: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1895	0.0552	0.0442	0.0055	0.2348	0.0221	0.0605	0.0019	0.1335	0.0717	0.2450	0.0186	0.0658	0.0007
STD		0.0081	0.0094	0.0037	0.0017	0.0131	0.0053	0.0062	0.0007	0.0041	0.0062	0.0132	0.0055	0.0064	0.0004

Tabelle 31: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.59). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0554	0.0744	0.0072	0.0104	0.0246	0.0231	0.0020	0.0023	0.0746	0.0762	0.0237	0.0322	0.0011	0.0018
STD		0.0132	0.0127	0.0039	0.0032	0.0044	0.0057	0.0007	0.0007	0.0059	0.0063	0.0044	0.0081	0.0005	0.0008

Tabelle 32: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.0570	0.0789	0.0073	0.0127	0.0220	0.0212	0.0018	0.0025	0.0629	0.0648	0.0204	0.0324	0.0009	0.0018
STD		0.0167	0.0208	0.0051	0.0091	0.0055	0.0046	0.0010	0.0013	0.0048	0.0050	0.0062	0.0043	0.0005	0.0005

Tabelle 33: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0481	0.0702	0.0040	0.0091	0.0255	0.0225	0.0015	0.0023	0.0765	0.0783	0.0239	0.0338	0.0011	0.0018
STD		0.0186	0.0225	0.0028	0.0049	0.0096	0.0074	0.0010	0.0011	0.0064	0.0070	0.0086	0.0059	0.0007	0.0008

Tabelle 34: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.0641	0.0983	0.0076	0.0179	0.0239	0.0235	0.0019	0.0029	0.0617	0.0643	0.0203	0.0373	0.0010	0.0023
STD		0.0184	0.0330	0.0039	0.0115	0.0045	0.0066	0.0008	0.0014	0.0032	0.0033	0.0051	0.0049	0.0005	0.0006

Tabelle 35: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0634	0.0874	0.0076	0.0160	0.0274	0.0201	0.0021	0.0028	0.0661	0.0684	0.0250	0.0360	0.0014	0.0022
STD		0.0229	0.0211	0.0041	0.0059	0.0107	0.0054	0.0010	0.0011	0.0075	0.0076	0.0103	0.0052	0.0010	0.0007

Tabelle 36: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.62). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.1538	0.0458	0.0287	0.0039	0.1812	0.0236	0.0368	0.0015	0.1268	0.0900	0.1889	0.0241	0.0397	0.0010
STD		0.0094	0.0124	0.0028	0.0019	0.0127	0.0062	0.0043	0.0007	0.0062	0.0078	0.0115	0.0057	0.0046	0.0005

Tabelle 37: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1816	0.0503	0.0405	0.0047	0.2193	0.0210	0.0538	0.0016	0.1252	0.0705	0.2291	0.0171	0.0582	0.0006
STD		0.0092	0.0086	0.0037	0.0019	0.0154	0.0044	0.0064	0.0006	0.0058	0.0051	0.0146	0.0049	0.0069	0.0003

Tabelle 38: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1835	0.0467	0.0409	0.0046	0.2224	0.0210	0.0546	0.0016	0.1328	0.0770	0.2312	0.0188	0.0594	0.0007
STD		0.0032	0.0114	0.0018	0.0025	0.0037	0.0056	0.0024	0.0008	0.0050	0.0053	0.0052	0.0041	0.0034	0.0003

Tabelle 39: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.1924	0.0498	0.0463	0.0056	0.2377	0.0174	0.0631	0.0014	0.1251	0.0599	0.2521	0.0156	0.0692	0.0005
STD		0.0095	0.0144	0.0039	0.0033	0.0152	0.0048	0.0069	0.0008	0.0050	0.0038	0.0119	0.0047	0.0067	0.0004

Tabelle 40: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.1947	0.0582	0.0469	0.0061	0.2440	0.0191	0.0649	0.0016	0.1268	0.0608	0.2541	0.0160	0.0700	0.0006
STD		0.0074	0.0105	0.0038	0.0021	0.0137	0.0044	0.0068	0.0006	0.0046	0.0060	0.0131	0.0037	0.0067	0.0003

Tabelle 41: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1421	0.0548	0.0255	0.0064	0.1692	0.0210	0.0325	0.0017	0.1083	0.0766	0.1777	0.0244	0.0350	0.0011
STD		0.0114	0.0184	0.0034	0.0041	0.0153	0.0071	0.0051	0.0010	0.0052	0.0069	0.0149	0.0066	0.0053	0.0008

Tabelle 42: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1401	0.0552	0.0242	0.0061	0.1659	0.0178	0.0308	0.0014	0.1035	0.0734	0.1721	0.0182	0.0325	0.0006
STD		0.0090	0.0131	0.0025	0.0027	0.0053	0.0036	0.0018	0.0006	0.0053	0.0056	0.0056	0.0028	0.0020	0.0002

Tabelle 43: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1372	0.0572	0.0238	0.0061	0.1648	0.0191	0.0309	0.0016	0.0984	0.0683	0.1704	0.0180	0.0326	0.0006
STD		0.0148	0.0103	0.0036	0.0025	0.0142	0.0033	0.0044	0.0004	0.0067	0.0071	0.0132	0.0047	0.0046	0.0003

Tabelle 44: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1434	0.0522	0.0250	0.0058	0.1670	0.0175	0.0308	0.0013	0.0984	0.0679	0.1727	0.0183	0.0329	0.0006
STD		0.0099	0.0172	0.0028	0.0031	0.0101	0.0050	0.0035	0.0007	0.0054	0.0042	0.0104	0.0041	0.0038	0.0003

Tabelle 45: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1303	0.0497	0.0218	0.0050	0.1533	0.0195	0.0272	0.0015	0.1019	0.0758	0.1598	0.0187	0.0288	0.0007
STD		0.0111	0.0161	0.0031	0.0030	0.0112	0.0068	0.0034	0.0008	0.0085	0.0085	0.0109	0.0073	0.0036	0.0005

Tabelle 46: Simulation Ridge mit 2000 Objekten (n1=1000, n2=1000), 20 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1240	0.0433	0.0208	0.0038	0.1063	0.0165	0.0175	0.0011	0.0832	0.0661	0.1165	0.0145	0.0188	0.0004
STD		0.0094	0.0073	0.0024	0.0013	0.0089	0.0035	0.0023	0.0004	0.0036	0.0041	0.0088	0.0030	0.0024	0.0002

Tabelle 47: Simulation Ridge mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.1327	0.0568	0.0233	0.0066	0.1099	0.0165	0.0188	0.0015	0.0757	0.0575	0.1192	0.0135	0.0203	0.0004
STD		0.0071	0.0097	0.0026	0.0025	0.0060	0.0024	0.0020	0.0004	0.0028	0.0038	0.0068	0.0032	0.0021	0.0002

Tabelle 48: Simulation Ridge mit 2000 Objekten (n1=1000, n2=1000), 60 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1171	0.0461	0.0186	0.0041	0.1030	0.0205	0.0161	0.0015	0.0946	0.0788	0.1139	0.0171	0.0177	0.0006
STD		0.0076	0.0110	0.0023	0.0017	0.0084	0.0046	0.0023	0.0006	0.0035	0.0052	0.0101	0.0042	0.0027	0.0003

Tabelle 49: Simulation Ridge mit 2000 Objekten (n1=1000, n2=1000), 80 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1312	0.0534	0.0230	0.0058	0.1113	0.0189	0.0189	0.0016	0.0828	0.0640	0.1236	0.0148	0.0209	0.0005
STD		0.0064	0.0104	0.0021	0.0025	0.0060	0.0031	0.0023	0.0007	0.0023	0.0035	0.0075	0.0028	0.0026	0.0002

Tabelle 50: Simulation Ridge mit 2000 Objekten (n1=1000, n2=1000), 100 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1287	0.0517	0.0225	0.0054	0.1120	0.0192	0.0191	0.0016	0.0861	0.0675	0.1233	0.0158	0.0208	0.0005
STD		0.0108	0.0134	0.0035	0.0025	0.0103	0.0052	0.0033	0.0007	0.0031	0.0055	0.0122	0.0031	0.0037	0.0003

9.1.3 Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels Simulationsstudien

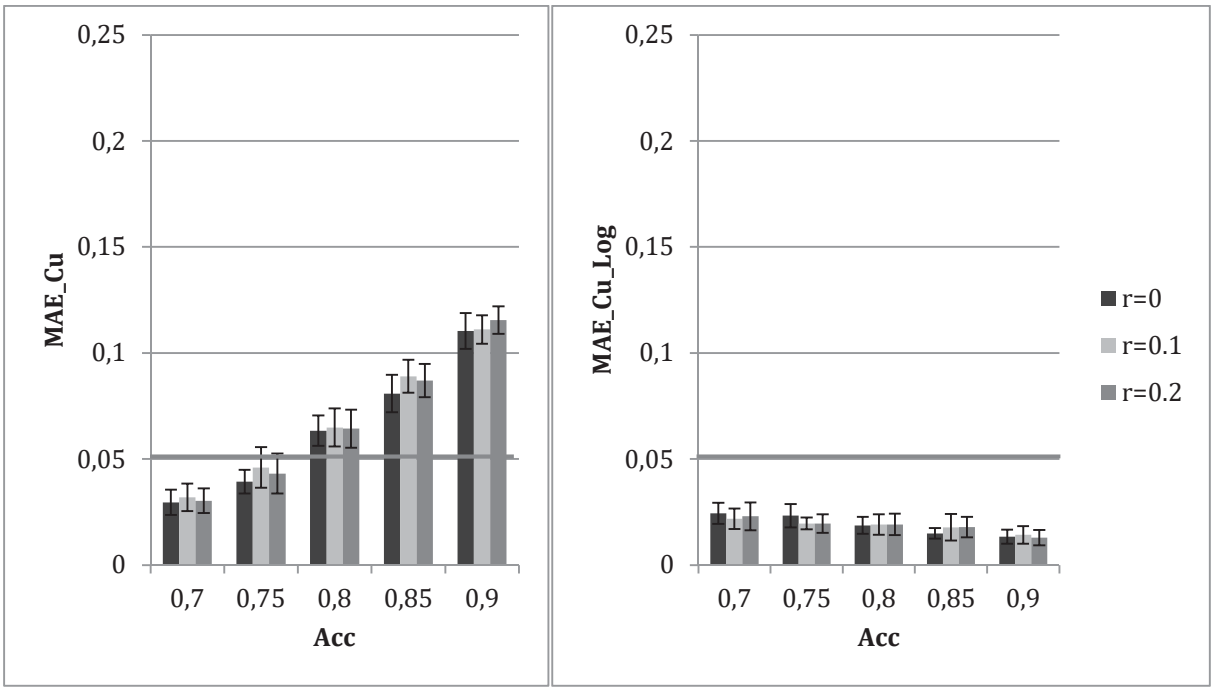


Abbildung 69: Ergebnisse für die **PLSDA** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) unabhängig von der Korrelation steigt der MAE_Cu an. Nur bei einer Korrektklassifizierungsrate Acc=0.7 und 0.75 gibt die PLSDA sich nah an der wahren Wahrscheinlichkeit befindende Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

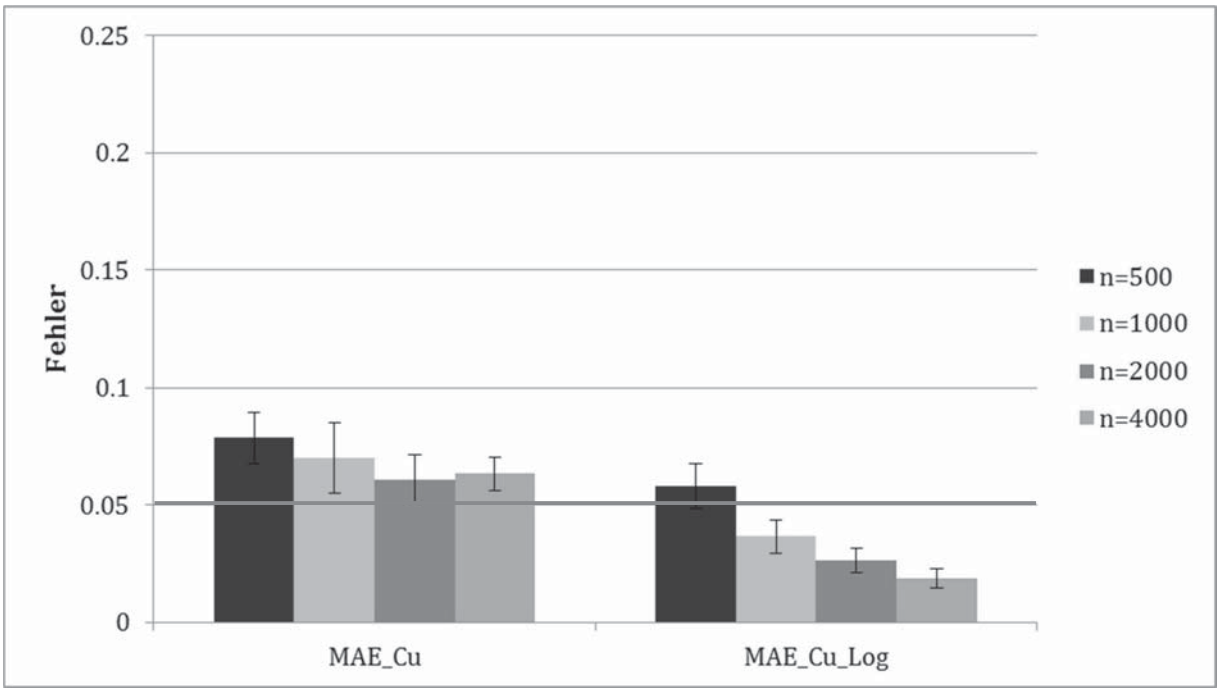


Abbildung 70: Ergebnisse für die **PLSDA** mit einer Korrektklassifizierungsrate Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Die Standardabweichung nimmt erwartungsgemäß mit zunehmender Datensatzgröße ab. Ab einer Datensatzgröße von 1000 Molekülen werden nach Kalibrierung gute Wahrscheinlichkeitsschätzer erhalten (Werte befinden sich unterhalb oder auf Höhe der roten Linie).

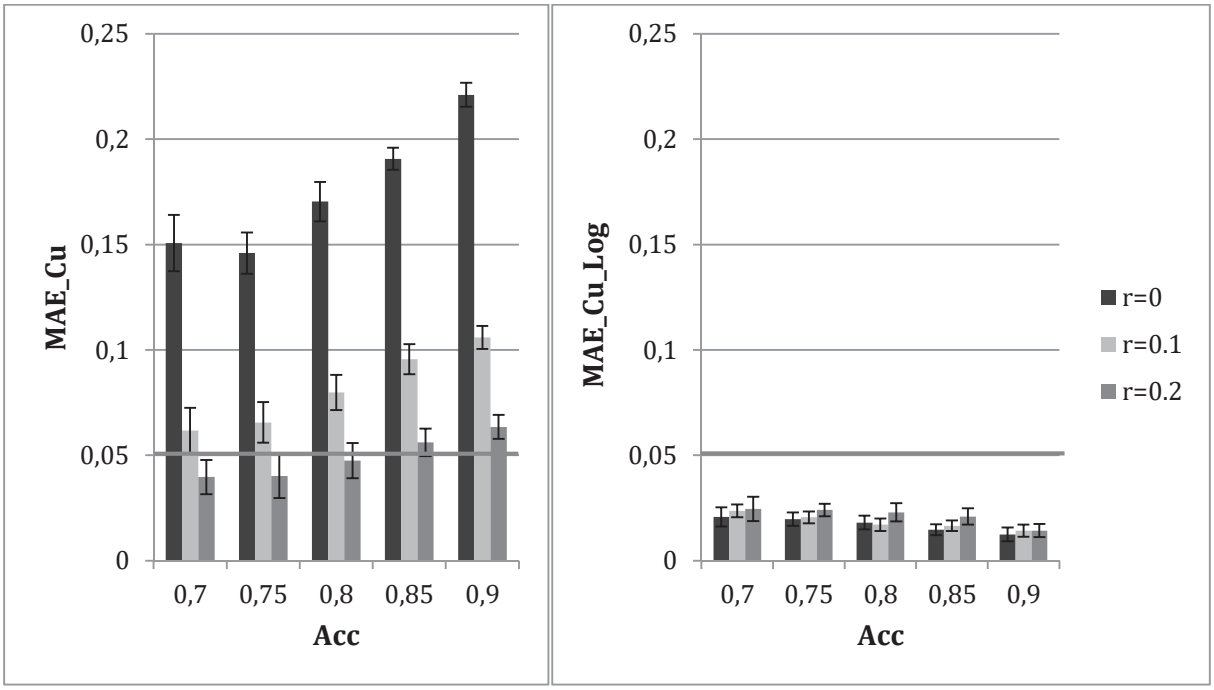


Abbildung 71: Ergebnisse für den **RFR** mit 4000 Objekten, 40 Variablen und r=0. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate

(Acc) und mit abnehmender Korrelation steigt der MAE_Cu an. Nur bei korrelierten Daten und mittlerer Korrektklassifizierungsrate (Acc) gibt der RFR gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

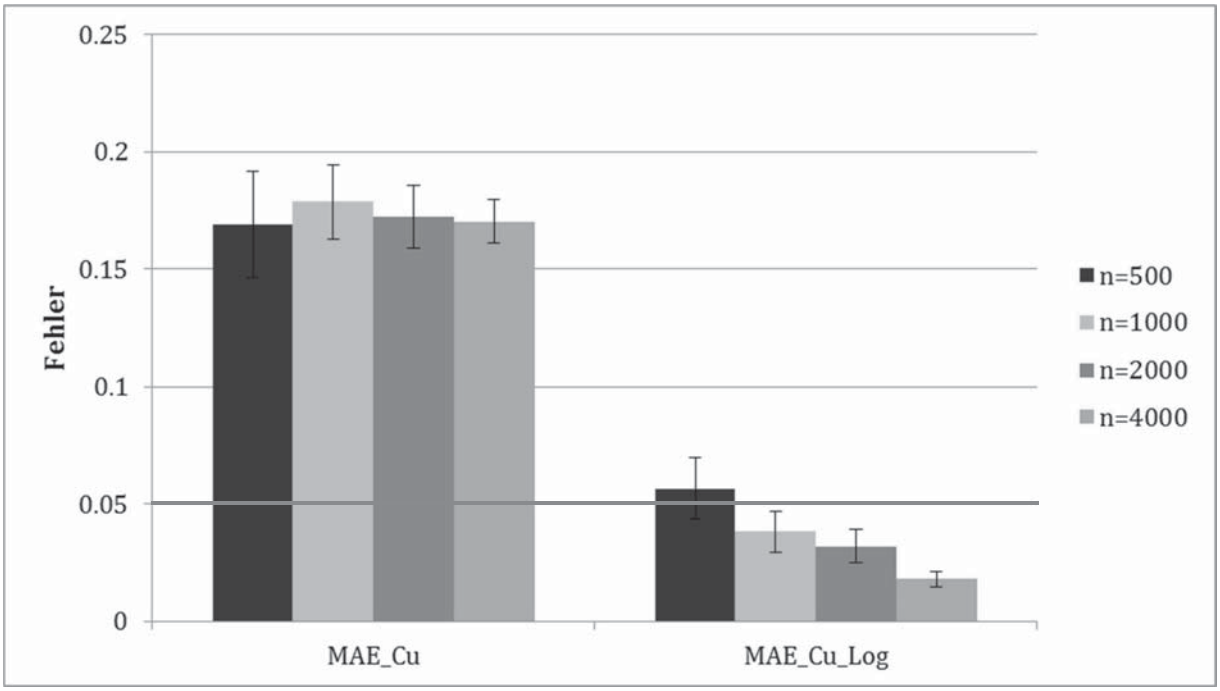


Abbildung 72: Ergebnisse für den **RFR** mit einer Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab, genauso wie die Standardabweichung. Nur nach Kalibrierung werden ab 1000 Molekülen gut kalibrierte Wahrscheinlichkeitsschätzer hervorgebracht (Werte befinden sich unterhalb der roten Linie).

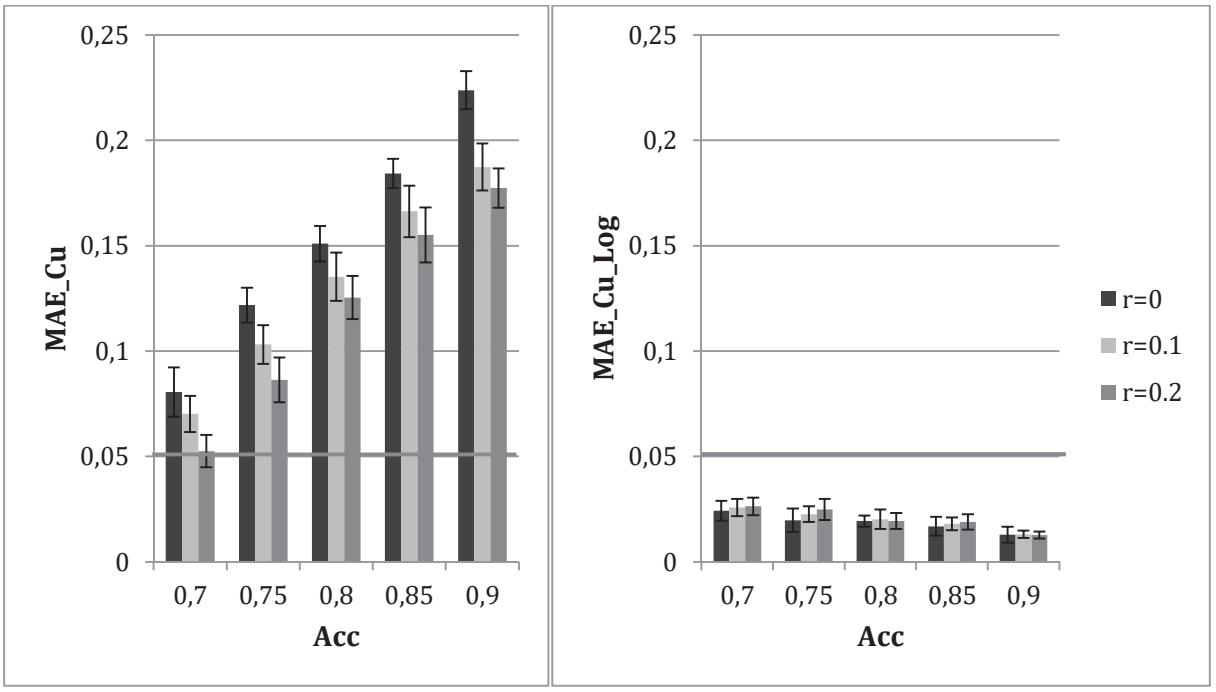


Abbildung 73: Ergebnisse für die **SVR** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Acc und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) und mit abnehmender Korrelation steigt der MAE_Cu an. Nur bei einer Korrektklassifizierungsrate Acc=0.7 und r=0.2 gibt die SVM gut kalibrierte Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

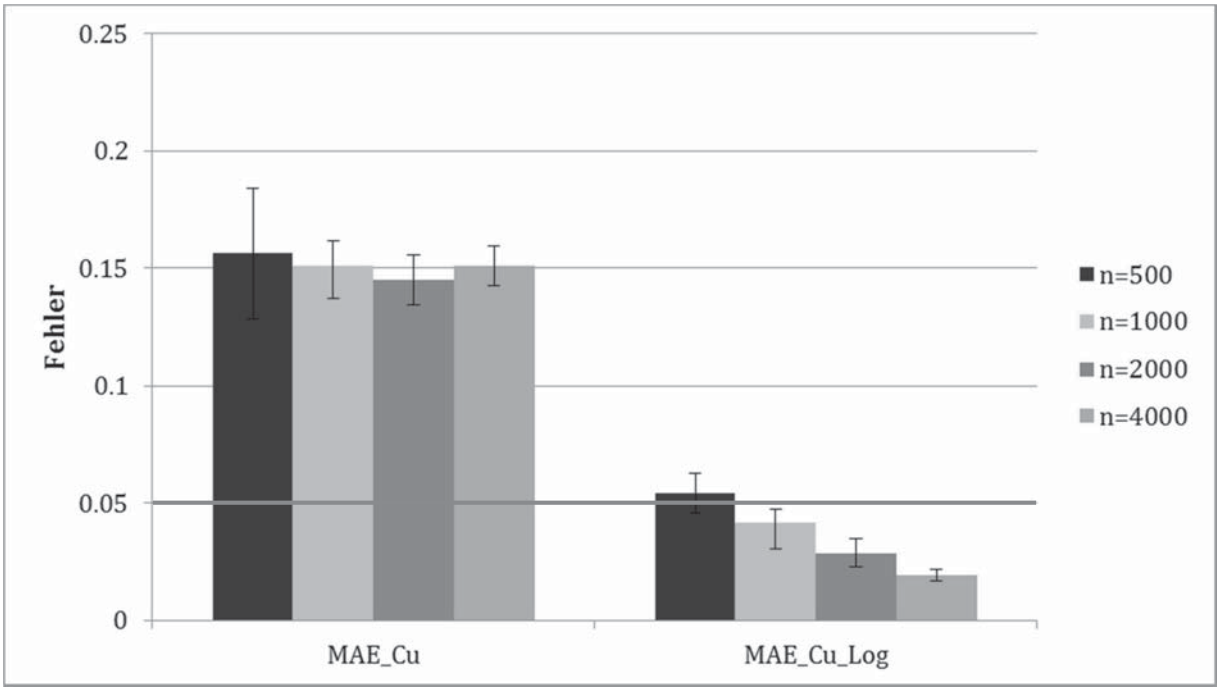


Abbildung 74: Ergebnisse für die **SVR** mit einer Korrektklassifizierungsrate Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwi-

schen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab, genauso wie die Standardabweichung. Nur nach Kalibrierung werden ab 500 Molekülen gut kalibrierte Wahrscheinlichkeitsschätzer hervorgebracht (Werte befinden sich unterhalb der roten Linie).

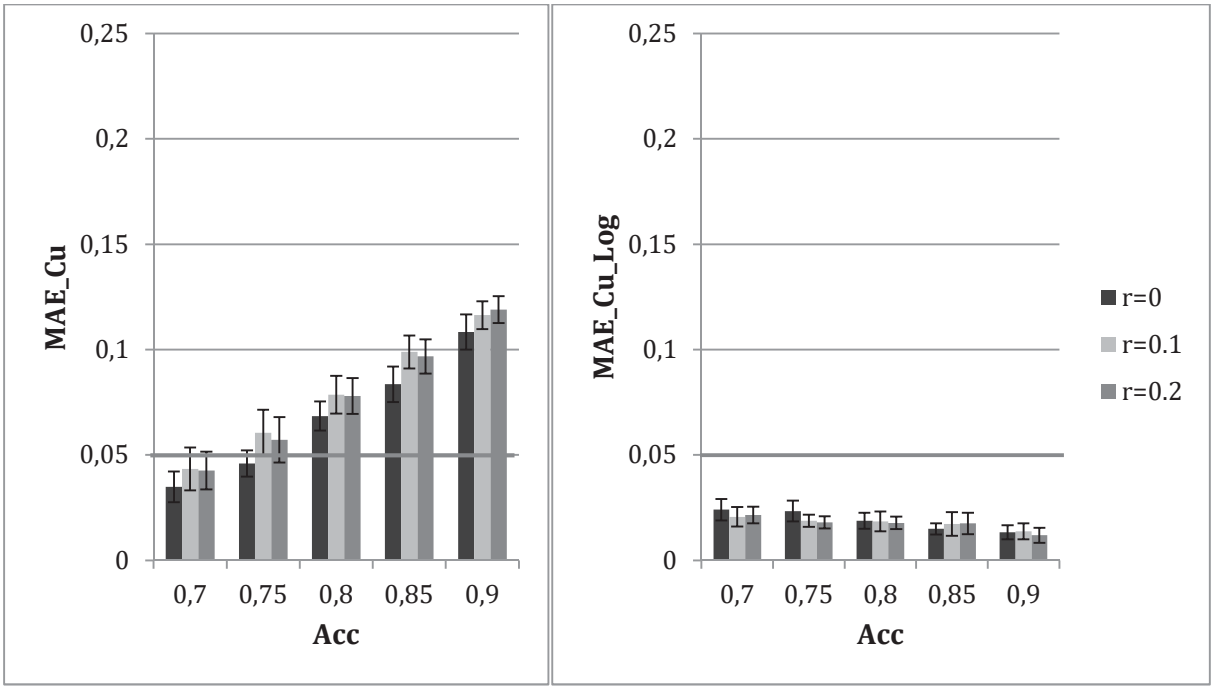


Abbildung 75: Ergebnisse für die **Ridge** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) (nahezu unabhängig von der Korrelation) steigt der MAE_Cu an. Nur bei einer Korrektklassifizierungsrate Acc=0.7 und 0.75 gibt die Ridge sich nah an der wahren Wahrscheinlichkeit befindende Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

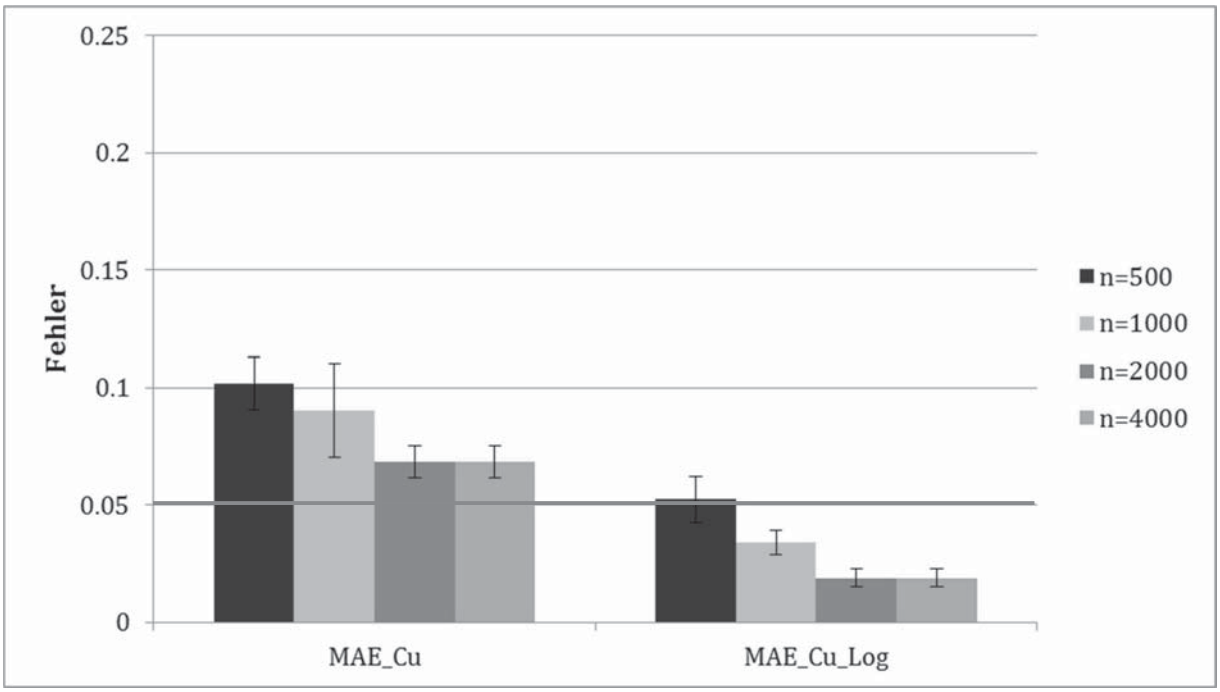


Abbildung 76: Ergebnisse für die **Ridge** mit einer Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Mit zunehmender Datensatzgröße nehmen, sowohl der MAE_Cu, als auch der MAE_Cu_Log ab. Nach Kalibrierung durch logistische Regression werden gute Wahrscheinlichkeitsschätzer erhalten (Werte befinden sich unterhalb oder auf Höhe der roten Linie).

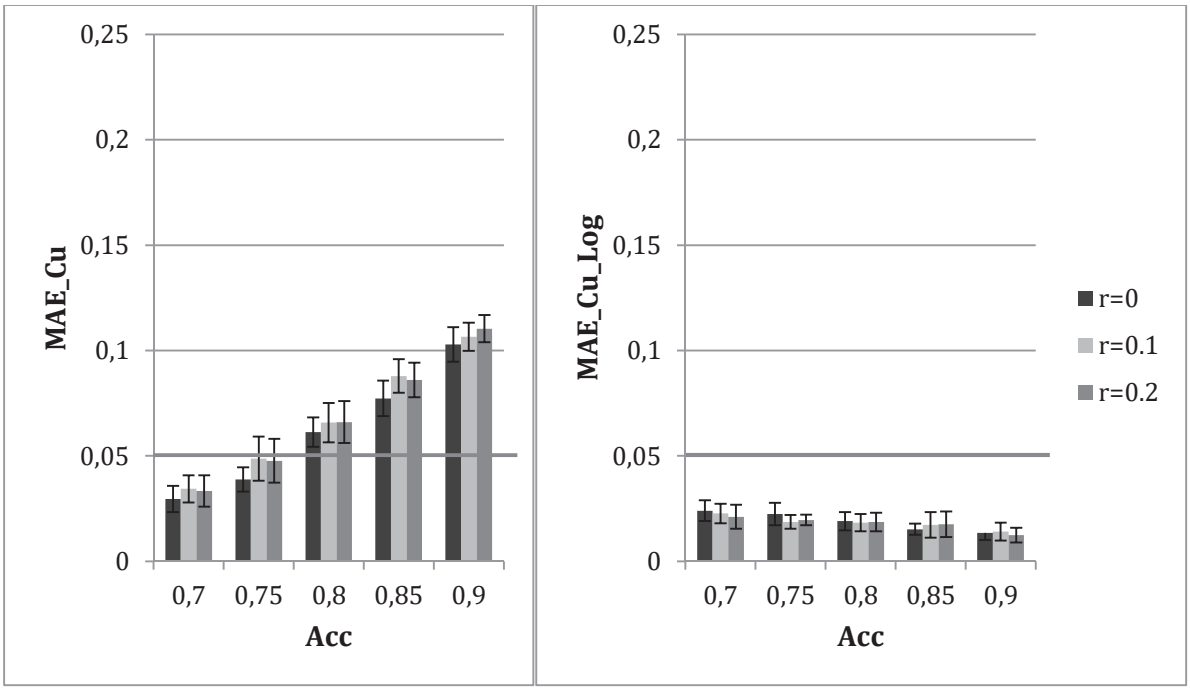


Abbildung 77: Ergebnisse für das **Elastic Net** mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) (nahezu unabhängig von der Korrelation) steigt der MAE_Cu an. Nur bei einer Acc=0.7 und 0.75 gibt das Elastic Net sich nah an der wahren Wahrscheinlichkeit befindende Klassenzugehörigkeits-

Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

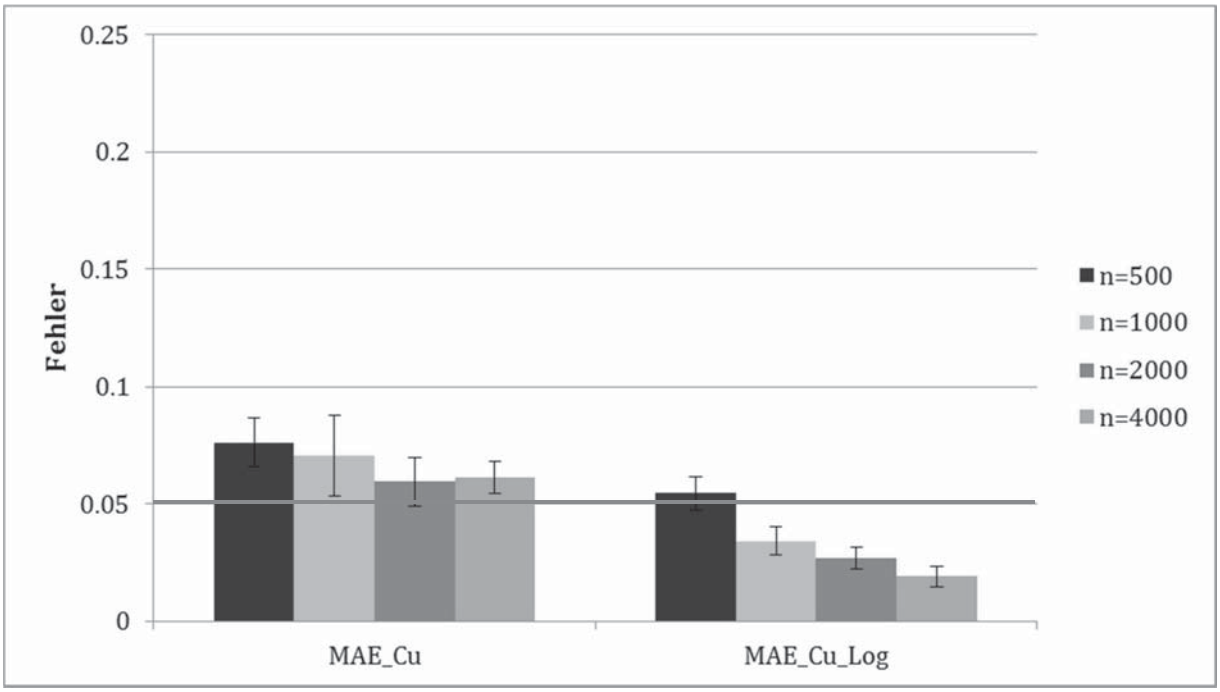


Abbildung 78: Ergebnisse für das **Elastic Net** mit einer Korrektklassifizierungsrate Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Mit zunehmender Datensatzgröße nehmen, sowohl der MAE_Cu (allerdings nur sehr leicht), als auch der MAE_Cu_Log ab. Nach Kalibrierung durch logistische Regression werden gute Wahrscheinlichkeitsschätzer erhalten (Werte befinden sich unterhalb oder auf Höhe der roten Linie).

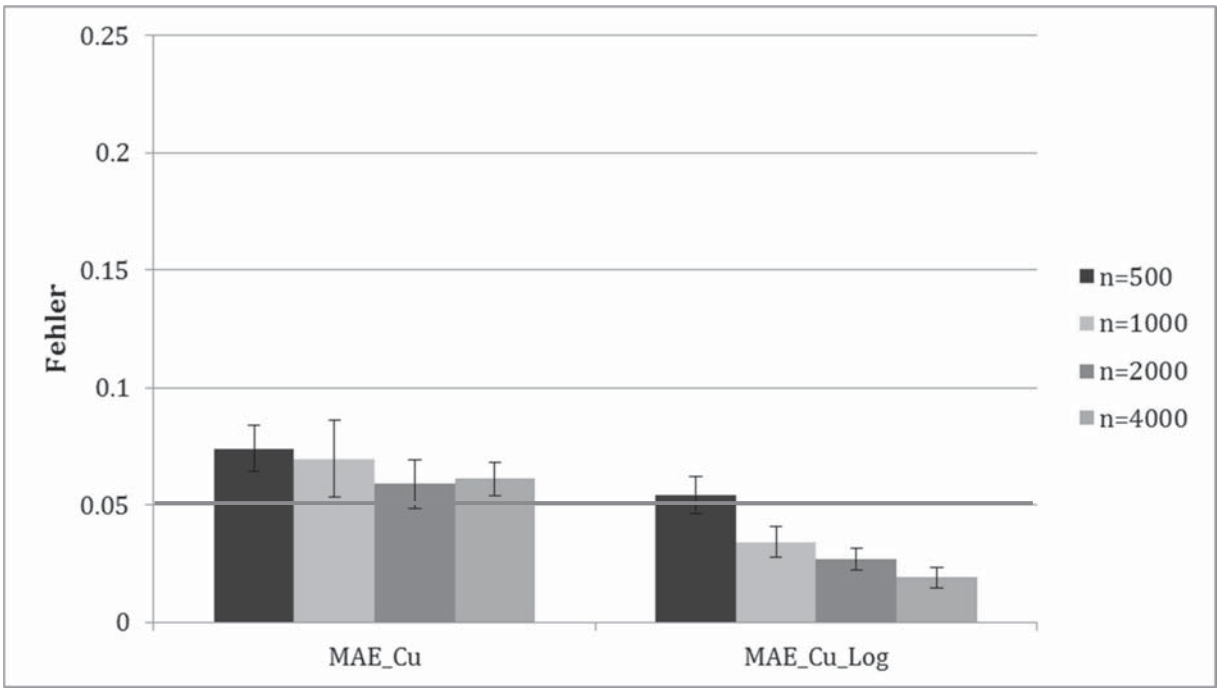


Abbildung 79: Ergebnisse für das **Lasso** (abgeschnitten) mit einer Korrektklassifizierungsrate Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Nur für den MAE_Cu_Log ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Der Fehler nimmt mit zunehmender Datensatzgröße ab. Nur nach Kalibrierung werden gut kalibrierte Wahrscheinlichkeitsschätzer hervorgebracht (Werte befinden sich unterhalb der roten Linie). Im Vergleich zu den skalierten Ergebnissen unterscheiden sich die abgeschnitten lediglich durch die absolute Höhe des MAE_Cu. Dieser ist deutlich geringer.

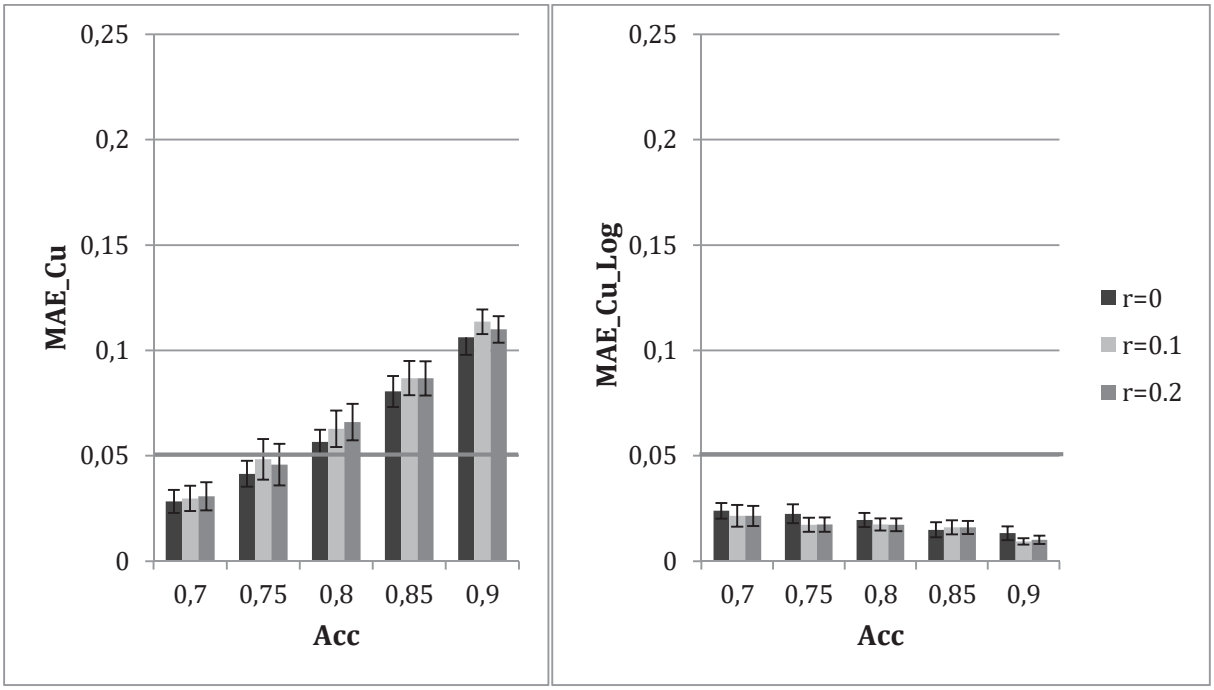


Abbildung 80: Ergebnisse für die **SPLS** (abgeschnitten) mit 4000 Objekten und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurden die Korrektklassifizierungsrate (Acc) und die Korrelation. Mit steigender Korrektklassifizierungsrate (Acc) (nahezu unabhängig von der Korrelation) steigt der MAE_Cu an. Nur bei einer Korrekt-

klassifizierungsrate Acc=0.7 und 0.75 gibt die SPLS sich nah an der wahren Wahrscheinlichkeit befindende Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer aus (unterhalb der roten Linie). Der MAE_Cu_Log hingegen bleibt immer niedrig, alle kalibrierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer sind nah an der wahren Wahrscheinlichkeit.

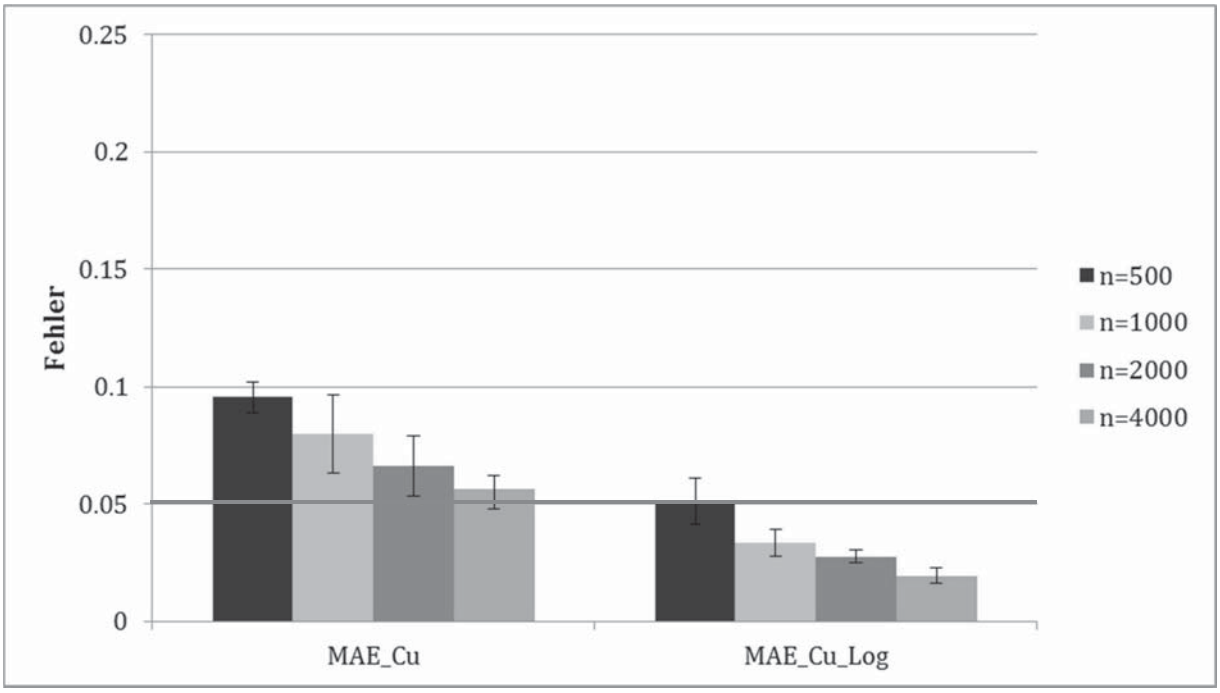


Abbildung 81: Ergebnisse für die **SPLS** (abgeschnitten) mit einer Korrektklassifizierungsrate Acc=0.8, r=0 und 40 Variablen. Aufgetragen sind die Mittelwerte der Fehler aus zehn wiederholten Versuchen sowie deren Standardabweichung. Variiert wurde die Datensatzgröße. Sowohl für den MAE_Cu, als auch für den MAE_Cu_Log, ist eine Abhängigkeit zwischen der Anzahl an Objekten und der Größe des Fehlers erkennbar. Alle kalibrierten Wahrscheinlichkeitsschätzer befinden sich nah an der wahren Wahrscheinlichkeit. (Werte befinden sich unterhalb oder auf Höhe der roten Linie)

Tabelle 51: Simulation RF mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.1314	0.0629	0.0354	0.0065	0.1158	0.0591	0.0205	0.0059	0.2050	0.1883	0.1357	0.0627	0.0261	0.0063
STD		0.0307	0.0147	0.0262	0.0027	0.0230	0.0147	0.0034	0.0025	0.0044	0.0086	0.0258	0.0136	0.0071	0.0026

Tabelle 52: Simulation RF mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	/	0.0613	/	0.006	/	0.0586	/	0.0056	0.1954	0.1715	0.1505	0.0621	0.0314	0.0060
STD		/	0.0158	/	0.003	/	0.0166	/	0.0026	0.0050	0.0116	0.0285	0.0135	0.0084	0.0025

Tabelle 53: Simulation RF mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1539	0.0756	0.0341	0.0111	0.1851	0.0583	0.0420	0.0070	0.1722	0.1334	0.1951	0.0580	0.0476	0.0059
STD		0.0111	0.0231	0.0043	0.0077	0.0202	0.0157	0.0072	0.0040	0.0045	0.0107	0.0209	0.0145	0.0079	0.0031

Tabelle 54: Simulation RF mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.1655	0.0874	0.0392	0.0141	0.2147	0.0508	0.0539	0.0065	0.1523	0.1000	0.2273	0.0428	0.0590	0.0033
STD		0.0104	0.0262	0.0045	0.0071	0.0197	0.0166	0.0084	0.0032	0.0042	0.0131	0.0233	0.0104	0.0096	0.0016

Tabelle 55: Simulation RF mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1744	0.0870	0.0435	0.0162	0.2318	0.0372	0.0615	0.0049	0.1347	0.0747	0.2426	0.0359	0.0666	0.0024
STD		0.0092	0.0226	0.0044	0.0078	0.0144	0.0065	0.0078	0.0016	0.0043	0.0099	0.0169	0.0079	0.0069	0.0009

Tabelle 56: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.1323	0.0449	0.0251	0.0033	0.1227	0.0428	0.0193	0.0031	0.2082	0.1917	0.1152	0.0449	0.0180	0.0034
STD		0.0121	0.0089	0.0035	0.0013	0.0191	0.0125	0.0044	0.0017	0.0045	0.0091	0.0173	0.0105	0.0050	0.0015

Tabelle 57: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.73	0.1363	0.0391	0.0267	0.0026	0.1288	0.0387	0.0208	0.0025	0.2053	0.1856	0.1272	0.0427	0.0210	0.0029
STD		0.0111	0.0081	0.0039	0.0011	0.0212	0.0080	0.0047	0.0009	0.0048	0.0098	0.0193	0.0102	0.0052	0.0013

Tabelle 58: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.78	0.1456	0.0392	0.0291	0.0027	0.1669	0.0348	0.0327	0.0022	0.1853	0.1512	0.1694	0.0332	0.0351	0.0018
STD		0.0127	0.0104	0.0043	0.0016	0.0232	0.0087	0.0068	0.0011	0.0044	0.0108	0.0185	0.0065	0.0064	0.0006

Tabelle 59 Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.82	0.1579	0.0433	0.0338	0.0030	0.1958	0.0340	0.0430	0.0021	0.1698	0.1261	0.1966	0.0341	0.0457	0.0019
STD		0.0120	0.0111	0.0043	0.0013	0.0218	0.0095	0.0076	0.0010	0.0045	0.0115	0.0199	0.0070	0.0073	0.0008

Tabelle 60: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.41). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1040	0.0965	0.0146	0.0221	0.0829	0.0278	0.0103	0.0042	0.0728	0.0656	0.0858	0.0265	0.0104	0.0014
STD		0.0185	0.0403	0.0050	0.0182	0.0118	0.0080	0.0030	0.0023	0.0065	0.0091	0.0133	0.0070	0.0034	0.0009

Tabelle 61: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.41). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0872	0.0689	0.0321	0.0177	0.0519	0.0470	0.0040	0.0043	0.1974	0.1959	0.0662	0.0509	0.0064	0.0039
STD		0.0499	0.0378	0.0482	0.0295	0.0055	0.0152	0.0016	0.0025	0.0077	0.0091	0.0114	0.0117	0.0022	0.0017

Tabelle 62: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.51). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0793	0.0436	0.0081	0.0038	0.0752	0.0419	0.0077	0.0035	0.1750	0.1706	0.0797	0.0425	0.0088	0.0032
STD		0.0147	0.0130	0.0030	0.0021	0.0164	0.0133	0.0033	0.0021	0.0081	0.0105	0.0161	0.0121	0.0033	0.0019

Tabelle 63: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.61). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0854	0.0427	0.0094	0.0036	0.0863	0.0358	0.0099	0.0026	0.1525	0.1449	0.0913	0.0352	0.0111	0.0021
STD		0.0140	0.0199	0.0027	0.0035	0.0162	0.0125	0.0032	0.0019	0.0081	0.0111	0.0156	0.0089	0.0037	0.0010

Tabelle 64 Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.75). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1047	0.0575	0.0141	0.0064	0.1083	0.0368	0.0148	0.0033	0.1236	0.1114	0.1148	0.0337	0.0160	0.0019
STD		0.0139	0.0153	0.0037	0.0029	0.0151	0.0076	0.0040	0.0013	0.0073	0.0106	0.0164	0.0057	0.0041	0.0006

Tabelle 65: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.92). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1244	0.0670	0.0192	0.0087	0.1218	0.0276	0.0184	0.0026	0.0928	0.0759	0.1301	0.0238	0.0203	0.0010
STD		0.0151	0.0179	0.0042	0.0043	0.0124	0.0063	0.0036	0.0009	0.0062	0.0094	0.0136	0.0057	0.0040	0.0005

Tabelle 66: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0654	0.0594	0.0137	0.0080	0.0509	0.0443	0.0044	0.0039	0.1884	0.1885	0.0577	0.0479	0.0050	0.0036
STD		0.0319	0.0225	0.0260	0.0075	0.0189	0.0155	0.0027	0.0023	0.0083	0.0087	0.0142	0.0118	0.0021	0.0018

Tabelle 67: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0534	0.0531	0.0043	0.0055	0.0512	0.0439	0.0038	0.0032	0.1750	0.1745	0.0594	0.0468	0.0052	0.0033
STD		0.0154	0.0224	0.0021	0.0056	0.0143	0.0158	0.0017	0.0016	0.0083	0.0090	0.0110	0.0081	0.0018	0.0011

Tabelle 68: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.85). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0677	0.0528	0.0064	0.0051	0.0682	0.0395	0.0064	0.0031	0.1388	0.1364	0.0681	0.0382	0.0064	0.0023
STD		0.0118	0.0090	0.0020	0.0019	0.0127	0.0063	0.0020	0.0009	0.0078	0.0090	0.0143	0.0048	0.0025	0.0006

Tabelle 69 Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0778	0.0647	0.0086	0.0083	0.0747	0.0378	0.0079	0.0035	0.1137	0.1096	0.0778	0.0354	0.0080	0.0021
STD		0.0136	0.0269	0.0030	0.0077	0.0139	0.0134	0.0026	0.0022	0.0074	0.0095	0.0147	0.0103	0.0028	0.0011

Tabelle 70: Simulation RF mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1040	0.0965	0.0146	0.0221	0.0829	0.0278	0.0103	0.0042	0.0728	0.0656	0.0858	0.0265	0.0104	0.0014
STD		0.0185	0.0403	0.0050	0.0182	0.0118	0.0080	0.0030	0.0023	0.0065	0.0091	0.0133	0.0070	0.0034	0.0009

Tabelle 71: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.1284	0.0334	0.0293	0.0018	0.1095	0.0315	0.0150	0.0015	0.2121	0.1977	0.1066	0.0335	0.0146	0.0018
STD		0.0302	0.0094	0.0262	0.0009	0.0130	0.0073	0.0033	0.0006	0.0030	0.0060	0.0092	0.0056	0.0027	0.0005

Tabelle 72: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.1346	0.0284	0.0251	0.0014	0.1469	0.0280	0.0251	0.0014	0.1954	0.1700	0.1440	0.0321	0.0256	0.0017
STD		0.0119	0.0067	0.0041	0.0007	0.0094	0.0067	0.0037	0.0006	0.0032	0.0066	0.0109	0.0058	0.0033	0.0006

Tabelle 73: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.1463	0.0305	0.0295	0.0016	0.1749	0.0271	0.0351	0.0013	0.1793	0.1442	0.1732	0.0295	0.0358	0.0013
STD		0.0077	0.0067	0.0030	0.0009	0.0059	0.0061	0.0031	0.0007	0.0028	0.0056	0.0078	0.0061	0.0032	0.0006

Tabelle 74: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.1786	0.0371	0.0377	0.0028	0.2111	0.0204	0.0497	0.0013	0.1474	0.0952	0.2191	0.0194	0.0545	0.0007
STD		0.0083	0.0072	0.0029	0.0014	0.0115	0.0048	0.0042	0.0006	0.0026	0.0063	0.0103	0.0042	0.0042	0.0003

Tabelle 75: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1850	0.0410	0.0406	0.0030	0.2236	0.0187	0.0550	0.0011	0.1326	0.0757	0.2322	0.0172	0.0599	0.0006
STD		0.0081	0.0104	0.0029	0.0012	0.0106	0.0052	0.0038	0.0004	0.0025	0.0060	0.0098	0.0050	0.0040	0.0003

Tabelle 76: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0718	0.0274	0.0078	0.0013	0.0587	0.0228	0.0048	0.0009	0.1939	0.1904	0.0600	0.0297	0.0052	0.0014
STD		0.0158	0.0066	0.0058	0.0006	0.0135	0.0056	0.0016	0.0005	0.0038	0.0051	0.0129	0.0051	0.0017	0.0004

Tabelle 77: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0762	0.0339	0.0072	0.0022	0.0762	0.0313	0.0074	0.0019	0.1704	0.1643	0.0776	0.0310	0.0078	0.0016
STD		0.0136	0.0083	0.0019	0.0010	0.0150	0.0075	0.0022	0.0009	0.0036	0.0054	0.0142	0.0039	0.0023	0.0004

Tabelle 78: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0854	0.0339	0.0087	0.0020	0.0900	0.0272	0.0096	0.0014	0.1478	0.1392	0.0920	0.0262	0.0103	0.0011
STD		0.0121	0.0072	0.0022	0.0008	0.0149	0.0055	0.0027	0.0005	0.0039	0.0059	0.0149	0.0032	0.0028	0.0003

Tabelle 79: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0987	0.0358	0.0118	0.0024	0.1026	0.0224	0.0125	0.0013	0.1218	0.1100	0.1084	0.0221	0.0137	0.0008
STD		0.0082	0.0112	0.0018	0.0014	0.0091	0.0070	0.0020	0.0007	0.0035	0.0053	0.0106	0.0027	0.0023	0.0002

Tabelle 80: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1116	0.0443	0.0152	0.0039	0.1096	0.0185	0.0146	0.0012	0.0963	0.0822	0.1163	0.0166	0.0158	0.0005
STD		0.0081	0.0081	0.0019	0.0020	0.0080	0.0025	0.0020	0.0005	0.0034	0.0050	0.0072	0.0029	0.0019	0.0002

Tabelle 81: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0509	0.0324	0.0035	0.0018	0.0453	0.0275	0.0029	0.0012	0.1871	0.1859	0.0464	0.0308	0.0031	0.0015
STD		0.0078	0.0074	0.0009	0.0007	0.0093	0.0073	0.0009	0.0005	0.0041	0.0047	0.0103	0.0045	0.0011	0.0004

Tabelle 82: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0494	0.0329	0.0033	0.0017	0.0503	0.0294	0.0034	0.0014	0.1684	0.1666	0.0514	0.0295	0.0037	0.0014
STD		0.0087	0.0083	0.0009	0.0008	0.0096	0.0067	0.0011	0.0006	0.0041	0.0051	0.0099	0.0058	0.0012	0.0005

Tabelle 83: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0572	0.0368	0.0043	0.0023	0.0577	0.0302	0.0044	0.0017	0.1410	0.1383	0.0605	0.0288	0.0047	0.0013
STD		0.0097	0.0074	0.0014	0.0009	0.0105	0.0048	0.0014	0.0005	0.0041	0.0051	0.0098	0.0044	0.0013	0.0005

Tabelle 84: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0664	0.0365	0.0057	0.0024	0.0643	0.0233	0.0053	0.0013	0.1153	0.1114	0.0673	0.0244	0.0056	0.0009
STD		0.0085	0.0072	0.0014	0.0010	0.0080	0.0041	0.0012	0.0004	0.0042	0.0050	0.0077	0.0040	0.0012	0.0003

Tabelle 85: Simulation RF mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0844	0.0484	0.0096	0.0050	0.0699	0.0181	0.0070	0.0014	0.0798	0.0740	0.0735	0.0205	0.0071	0.0007
STD		0.0077	0.0118	0.0019	0.0025	0.0047	0.0024	0.0010	0.0005	0.0036	0.0041	0.0056	0.0035	0.0010	0.0002

Tabelle 86: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.68	0.1177	0.0234	0.0200	0.0008	0.0979	0.0220	0.0124	0.0007	0.2154	0.2022	0.0925	0.0223	0.0119	0.0008
STD		0.0086	0.0040	0.0033	0.0004	0.0096	0.0045	0.0017	0.0003	0.0018	0.0034	0.0104	0.0036	0.0022	0.0003

Tabelle 87: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.1387	0.0213	0.0263	0.0007	0.1522	0.0209	0.0270	0.0007	0.1920	0.1644	0.1456	0.0199	0.0263	0.0006
STD		0.0073	0.0054	0.0026	0.0003	0.0059	0.0054	0.0021	0.0003	0.0018	0.0036	0.0072	0.0033	0.0026	0.0002

Tabelle 88: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1506	0.0197	0.0307	0.0007	0.1806	0.0168	0.0366	0.0005	0.1748	0.1371	0.1769	0.0175	0.0375	0.0005
STD		0.0038	0.0051	0.0016	0.0004	0.0049	0.0033	0.0016	0.0002	0.0013	0.0030	0.0064	0.0032	0.0022	0.0003

Tabelle 89: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1761	0.0218	0.0357	0.0009	0.2009	0.0150	0.0445	0.0005	0.1581	0.1117	0.2023	0.0149	0.0474	0.0004
STD		0.0042	0.0063	0.0015	0.0006	0.0052	0.0046	0.0019	0.0003	0.0016	0.0031	0.0049	0.0031	0.0020	0.0002

Tabelle 90: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.1827	0.0286	0.0388	0.0016	0.2180	0.0147	0.0510	0.0007	0.1421	0.0894	0.2215	0.0131	0.0549	0.0003
STD		0.0066	0.0081	0.0019	0.0008	0.0058	0.0036	0.0023	0.0003	0.0018	0.0029	0.0043	0.0029	0.0021	0.0001

Tabelle 91: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0648	0.0236	0.0068	0.0009	0.0503	0.0211	0.0035	0.0008	0.1925	0.1900	0.0515	0.0227	0.0036	0.0008
STD		0.0143	0.0058	0.0054	0.0005	0.0089	0.0055	0.0010	0.0003	0.0033	0.0046	0.0091	0.0040	0.0012	0.0003

Tabelle 92: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0642	0.0215	0.0051	0.0009	0.0672	0.0196	0.0056	0.0008	0.1692	0.1645	0.0677	0.0208	0.0059	0.0007
STD		0.0082	0.0055	0.0012	0.0004	0.0089	0.0050	0.0014	0.0003	0.0032	0.0049	0.0084	0.0039	0.0013	0.0003

Tabelle 93: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0769	0.0209	0.0073	0.0008	0.0812	0.0173	0.0080	0.0006	0.1461	0.1389	0.0846	0.0186	0.0085	0.0005
STD		0.0081	0.0038	0.0012	0.0003	0.0086	0.0026	0.0013	0.0002	0.0034	0.0050	0.0093	0.0026	0.0015	0.0001

Tabelle 94: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0945	0.0304	0.0106	0.0019	0.0967	0.0190	0.0109	0.0010	0.1203	0.1102	0.1014	0.0175	0.0117	0.0005
STD		0.0056	0.0071	0.0015	0.0011	0.0060	0.0035	0.0015	0.0005	0.0033	0.0048	0.0065	0.0030	0.0015	0.0001

Tabelle 95: Simulation RF mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1075	0.0298	0.0142	0.0018	0.1029	0.0132	0.0131	0.0006	0.0939	0.0813	0.1103	0.0140	0.0141	0.0003
STD		0.0075	0.0071	0.0019	0.0008	0.0060	0.0035	0.0016	0.0003	0.0029	0.0044	0.0058	0.0038	0.0016	0.0002

Tabelle 96: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0380	0.0386	0.0026	0.0043	0.0326	0.0219	0.0015	0.0008	0.1961	0.1957	0.0357	0.0250	0.0018	0.0010
STD		0.0166	0.0138	0.0025	0.0058	0.0106	0.0048	0.0010	0.0004	0.0037	0.0046	0.0093	0.0063	0.0009	0.0005

Tabelle 97: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0386	0.0208	0.0021	0.0007	0.0399	0.0191	0.0023	0.0006	0.1682	0.1672	0.0422	0.0233	0.0024	0.0008
STD		0.0095	0.0042	0.0010	0.0003	0.0111	0.0038	0.0011	0.0003	0.0040	0.0052	0.0095	0.0042	0.0010	0.0003

Tabelle 98: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0472	0.0285	0.0031	0.0015	0.0487	0.0231	0.0032	0.0010	0.1406	0.1387	0.0524	0.0231	0.0035	0.0008
STD		0.0070	0.0044	0.0010	0.0006	0.0076	0.0030	0.0011	0.0003	0.0039	0.0049	0.0082	0.0034	0.0011	0.0003

Tabelle 99: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0590	0.0320	0.0047	0.0022	0.0564	0.0193	0.0043	0.0011	0.1146	0.1114	0.0608	0.0212	0.0046	0.0007
STD		0.0081	0.0077	0.0011	0.0010	0.0066	0.0055	0.0009	0.0005	0.0037	0.0046	0.0070	0.0033	0.0010	0.0002

Tabelle 100: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.0841	0.0453	0.0092	0.0041	0.0624	0.0135	0.0058	0.0010	0.0667	0.0617	0.0661	0.0143	0.0061	0.0004
STD		0.0089	0.0106	0.0019	0.0017	0.0052	0.0028	0.0011	0.0004	0.0028	0.0040	0.0054	0.0033	0.0010	0.0002

Tabelle 101: Simulation KNN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.1140	0.0688	0.0326	0.0069	0.0888	0.0660	0.0124	0.0066	0.2005	0.1935	0.0944	0.0607	0.0130	0.0056
STD		0.0376	0.0138	0.0372	0.0024	0.0070	0.0173	0.0036	0.0028	0.0106	0.0119	0.0123	0.0100	0.0037	0.0017

Tabelle 102: Simulation KNN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.1164	0.0565	0.0200	0.0057	0.1130	0.0500	0.0184	0.0043	0.1817	0.1691	0.1127	0.0531	0.0182	0.0045
STD		0.0156	0.0105	0.0056	0.0028	0.0149	0.0102	0.0044	0.0013	0.0114	0.0132	0.0159	0.0148	0.0046	0.0023

Tabelle 103: Simulation KNN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.1296	0.0686	0.0230	0.0100	0.1366	0.0512	0.0252	0.0052	0.1533	0.1334	0.1377	0.0509	0.0256	0.0040
STD		0.0159	0.0173	0.0051	0.0071	0.0216	0.0044	0.0065	0.0020	0.0114	0.0137	0.0213	0.0053	0.0069	0.0009

Tabelle 104: Simulation KNN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1322	0.0755	0.0241	0.0106	0.1424	0.0493	0.0263	0.0052	0.1340	0.1106	0.1483	0.0408	0.0280	0.0026
STD		0.0134	0.0193	0.0053	0.0046	0.0160	0.0074	0.0056	0.0015	0.0116	0.0137	0.0176	0.0066	0.0066	0.0007

Tabelle 105: Simulation KNN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1443	0.1032	0.0286	0.0177	0.1424	0.0458	0.0275	0.0055	0.0990	0.0751	0.1529	0.0363	0.0292	0.0028
STD		0.0197	0.0033	0.0072	0.0041	0.0182	0.0081	0.0058	0.0018	0.0102	0.0121	0.0178	0.0111	0.0062	0.0022

Tabelle 106: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.24). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0677	0.0459	0.0072	0.0031	0.0607	0.0396	0.0059	0.0024	0.2027	0.1989	0.0586	0.0382	0.0055	0.0023
STD		0.0147	0.0123	0.0030	0.0012	0.0140	0.0128	0.0026	0.0012	0.0080	0.0106	0.0138	0.0093	0.0027	0.0012

Tabelle 107: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0860	0.0367	0.0101	0.0024	0.0845	0.0343	0.0097	0.0022	0.1835	0.1754	0.0836	0.0316	0.0097	0.0016
STD		0.0171	0.0083	0.0033	0.0012	0.0194	0.0100	0.0034	0.0013	0.0083	0.0116	0.0182	0.0044	0.0034	0.0005

Tabelle 108: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.1101	0.0429	0.0155	0.0040	0.1198	0.0323	0.0174	0.0024	0.1535	0.1378	0.1212	0.0315	0.0182	0.0017
STD		0.0116	0.0155	0.0023	0.0032	0.0166	0.0110	0.0035	0.0023	0.0082	0.0116	0.0158	0.0108	0.0039	0.0014

Tabelle 109 Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1215	0.0603	0.0185	0.0074	0.1315	0.0403	0.0206	0.0043	0.1342	0.1149	0.1380	0.0289	0.0222	0.0015
STD		0.0128	0.0145	0.0031	0.0036	0.0167	0.0134	0.0038	0.0029	0.0077	0.0115	0.0143	0.0094	0.0041	0.0011

Tabelle 110: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1361	0.0618	0.0244	0.0079	0.1377	0.0271	0.0240	0.0025	0.1013	0.0793	0.1453	0.0237	0.0252	0.0010
STD		0.0177	0.0119	0.0061	0.0036	0.0171	0.0009	0.0051	0.0009	0.0070	0.0112	0.0167	0.0047	0.0053	0.0004

Tabelle 111: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.48). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0482	0.0460	0.0037	0.0032	0.0480	0.0438	0.0035	0.0029	0.1857	0.1859	0.0492	0.0417	0.0036	0.0027
STD		0.0110	0.0081	0.0016	0.0010	0.0131	0.0094	0.0018	0.0011	0.0101	0.0116	0.0126	0.0117	0.0017	0.0014

Tabelle 112: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0538	0.0517	0.0045	0.0041	0.0529	0.0476	0.0044	0.0036	0.1673	0.1672	0.0504	0.0437	0.0040	0.0030
STD		0.0147	0.0146	0.0023	0.0022	0.0146	0.0160	0.0021	0.0021	0.0098	0.0112	0.0161	0.0115	0.0022	0.0014

Tabelle 113: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.68). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0573	0.0565	0.0059	0.0056	0.0485	0.0403	0.0043	0.0033	0.1331	0.1325	0.0488	0.0362	0.0040	0.0022
STD		0.0166	0.0122	0.0033	0.0023	0.0143	0.0106	0.0024	0.0013	0.0089	0.0106	0.0172	0.0120	0.0028	0.0014

Tabelle 114 Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.81). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.0648	0.0584	0.0078	0.0069	0.0445	0.0317	0.0046	0.0027	0.1017	0.1007	0.0474	0.0337	0.0039	0.0019
STD		0.0204	0.0137	0.0042	0.0033	0.0141	0.0084	0.0026	0.0009	0.0083	0.0098	0.0169	0.0046	0.0026	0.0005

Tabelle 115: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=0.92). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0821	0.1104	0.0118	0.0264	0.0435	0.0307	0.0050	0.0044	0.0700	0.0686	0.0469	0.0301	0.0042	0.0017
STD		0.0250	0.0293	0.0063	0.0119	0.0118	0.0093	0.0023	0.0018	0.0078	0.0090	0.0149	0.0070	0.0025	0.0009

Tabelle 116: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0498	0.0451	0.0039	0.0031	0.0476	0.0406	0.0036	0.0024	0.1900	0.1902	0.0453	0.0412	0.0032	0.0026
STD		0.0148	0.0127	0.0020	0.0015	0.0148	0.0105	0.0020	0.0010	0.0095	0.0097	0.0155	0.0093	0.0019	0.0011

Tabelle 117: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.7). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0589	0.0508	0.0057	0.0050	0.0545	0.0420	0.0046	0.0036	0.1703	0.1707	0.0480	0.0423	0.0035	0.0028
STD		0.0145	0.0121	0.0037	0.0030	0.0108	0.0071	0.0023	0.0012	0.0110	0.0117	0.0134	0.0111	0.0020	0.0015

Tabelle 118: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.89). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0573	0.0520	0.0056	0.0048	0.0458	0.0386	0.0039	0.0029	0.1328	0.1332	0.0433	0.0401	0.0031	0.0026
STD		0.0115	0.0129	0.0026	0.0024	0.0111	0.0099	0.0017	0.0015	0.0102	0.0111	0.0123	0.0094	0.0017	0.0012

Tabelle 119 Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.02). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0661	0.0705	0.0086	0.0094	0.0440	0.0398	0.0045	0.0043	0.1091	0.1092	0.0431	0.0388	0.0034	0.0026
STD		0.0170	0.0193	0.0041	0.0052	0.0102	0.0096	0.0019	0.0022	0.0097	0.0108	0.0119	0.0077	0.0018	0.0011

Tabelle 120: Simulation KNN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.0906	0.1233	0.0172	0.0306	0.0342	0.0271	0.0043	0.0040	0.0613	0.0615	0.0364	0.0308	0.0028	0.0019
STD		0.0329	0.0268	0.0154	0.0135	0.0080	0.0078	0.0020	0.0016	0.0082	0.0097	0.0086	0.0073	0.0014	0.0009

Tabelle 121: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0647	0.0351	0.0061	0.0030	0.0554	0.0264	0.0042	0.0013	0.2037	0.2005	0.0564	0.0257	0.0043	0.0011
STD		0.0099	0.0163	0.0023	0.0035	0.0085	0.0083	0.0013	0.0007	0.0040	0.0036	0.0088	0.0032	0.0013	0.0002

Tabelle 122: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0806	0.0281	0.0091	0.0015	0.0838	0.0262	0.0095	0.0013	0.1792	0.1710	0.0851	0.0244	0.0095	0.0010
STD		0.0104	0.0104	0.0019	0.0012	0.0121	0.0084	0.0020	0.0008	0.0053	0.0054	0.0104	0.0063	0.0020	0.0005

Tabelle 123: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0982	0.0318	0.0132	0.0022	0.1081	0.0247	0.0149	0.0014	0.1489	0.1350	0.1131	0.0228	0.0156	0.0008
STD		0.0095	0.0113	0.0024	0.0015	0.0103	0.0077	0.0025	0.0008	0.0050	0.0057	0.0104	0.0047	0.0027	0.0004

Tabelle 124: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.39). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1110	0.0351	0.0163	0.0024	0.1171	0.0203	0.0173	0.0011	0.1247	0.1082	0.1240	0.0169	0.0184	0.0005
STD		0.0109	0.0048	0.0030	0.0007	0.0115	0.0017	0.0029	0.0003	0.0049	0.0053	0.0115	0.0031	0.0030	0.0002

Tabelle 125: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1195	0.0374	0.0187	0.0033	0.1172	0.0198	0.0180	0.0012	0.0970	0.0797	0.1266	0.0168	0.0194	0.0005
STD		0.0087	/	0.0026	/	0.0109	/	0.0027	/	0.0045	0.0052	0.0107	0.0041	0.0028	0.0003

Tabelle 126: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0425	0.0293	0.0028	0.0014	0.0380	0.0238	0.0024	0.0009	0.1885	0.1877	0.0363	0.0258	0.0020	0.0010
STD		0.0123	0.0042	0.0013	0.0004	0.0135	0.0035	0.0013	0.0002	0.0049	0.0052	0.0120	0.0050	0.0011	0.0003

Tabelle 127: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.53). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0448	0.0296	0.0034	0.0015	0.0434	0.0279	0.0031	0.0013	0.1669	0.1658	0.0410	0.0276	0.0027	0.0013
STD		0.0122	0.0080	0.0018	0.0009	0.0127	0.0083	0.0016	0.0008	0.0052	0.0059	0.0140	0.0063	0.0017	0.0006

Tabelle 128: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.63). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0463	0.0419	0.0034	0.0036	0.0446	0.0301	0.0031	0.0020	0.1409	0.1394	0.0456	0.0265	0.0031	0.0011
STD		0.0145	0.0085	0.0022	0.0016	0.0143	0.0041	0.0019	0.0005	0.0059	0.0067	0.0135	0.0048	0.0018	0.0004

Tabelle 129: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.74). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0495	0.0438	0.0042	0.0038	0.0424	0.0264	0.0034	0.0017	0.1138	0.1117	0.0466	0.0240	0.0034	0.0010
STD		0.0176	0.0117	0.0026	0.0030	0.0153	0.0065	0.0021	0.0007	0.0063	0.0078	0.0160	0.0053	0.0021	0.0004

Tabelle 130: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.89). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0580	0.0487	0.0059	0.0045	0.0393	0.0197	0.0037	0.0014	0.0811	0.0789	0.0447	0.0197	0.0034	0.0007
STD		0.0141	0.0077	0.0030	0.0024	0.0112	0.0018	0.0019	0.0005	0.0039	0.0051	0.0135	0.0056	0.0020	0.0003

Tabelle 131: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0448	0.0348	0.0031	0.0020	0.0416	0.0282	0.0027	0.0014	0.1895	0.1887	0.0363	0.0318	0.0022	0.0016
STD		0.0135	0.0036	0.0015	0.0002	0.0141	0.0065	0.0016	0.0004	0.0052	0.0048	0.0130	0.0078	0.0016	0.0008

Tabelle 132: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.7). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0427	0.0322	0.0026	0.0018	0.0424	0.0276	0.0025	0.0014	0.1651	0.1645	0.0362	0.0271	0.0020	0.0012
STD		0.0079	0.0085	0.0011	0.0009	0.0079	0.0085	0.0010	0.0007	0.0056	0.0055	0.0103	0.0063	0.0010	0.0005

Tabelle 133: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.82). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0424	0.0355	0.0028	0.0026	0.0406	0.0256	0.0025	0.0015	0.1415	0.1409	0.0369	0.0275	0.0021	0.0014
STD		0.0122	0.0136	0.0020	0.0016	0.0097	0.0096	0.0013	0.0009	0.0065	0.0065	0.0117	0.0076	0.0012	0.0009

Tabelle 134: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0500	0.0428	0.0039	0.0037	0.0391	0.0259	0.0026	0.0017	0.1177	0.1167	0.0383	0.0250	0.0023	0.0010
STD		0.0092	0.0078	0.0014	0.0017	0.0073	0.0065	0.0009	0.0007	0.0069	0.0074	0.0116	0.0035	0.0013	0.0003

Tabelle 135: Simulation KNN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.05072	0.07021	0.00463	0.00789	0.02867	0.02671	0.00226	0.00221	0.07775	0.07722	0.03323	0.02310	0.00201	0.00098
STD		0.01393	0.01932	0.00256	0.00381	0.00900	0.00705	0.00131	0.00102	0.00529	0.00564	0.00890	0.00527	0.00117	0.00057

Tabelle 136: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0491	0.0263	0.0036	0.0014	0.0444	0.0209	0.0028	0.0007	0.2050	0.2031	0.0429	0.0203	0.0026	0.0006
STD		0.0099	0.0070	0.0014	0.0011	0.0058	0.0061	0.0006	0.0003	0.0040	0.0042	0.0065	0.0039	0.0007	0.0002

Tabelle 137: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.26). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.73	0.0643	0.0201	0.0054	0.0006	0.0680	0.0198	0.0057	0.0006	0.1853	0.1804	0.0677	0.0211	0.0056	0.0007
STD		0.0059	0.0050	0.0009	0.0003	0.0073	0.0048	0.0012	0.0003	0.0041	0.0045	0.0081	0.0038	0.0013	0.0002

Tabelle 138: Simulation RF mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.31). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.78	0.0836	0.0231	0.0091	0.0010	0.0887	0.0204	0.0096	0.0008	0.1596	0.1506	0.0914	0.0188	0.0100	0.0006
STD		0.0068	0.0061	0.0016	0.0005	0.0076	0.0051	0.0016	0.0004	0.0040	0.0044	0.0078	0.0041	0.0016	0.0003

Tabelle 139: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1022	0.0222	0.0135	0.0013	0.1064	0.0120	0.0141	0.0005	0.1236	0.1098	0.1137	0.0133	0.0151	0.0003
STD		0.0077	0.0070	0.0020	0.0010	0.0089	0.0023	0.0021	0.0003	0.0037	0.0042	0.0074	0.0036	0.0021	0.0002

Tabelle 140: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.48). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1179	0.0509	0.0176	0.0045	0.1050	0.0174	0.0149	0.0011	0.0795	0.0647	0.1136	0.0124	0.0161	0.0003
STD		0.0065	0.0087	0.0019	0.0013	0.0065	0.0025	0.0017	0.0003	0.0026	0.0029	0.0066	0.0028	0.0017	0.0002

Tabelle 141: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0437	0.0236	0.0031	0.0009	0.0497	0.0211	0.0034	0.0008	0.1925	0.1900	0.0515	0.0227	0.0036	0.0008
STD		0.0070	0.0058	0.0008	0.0005	0.0085	0.0055	0.0010	0.0003	0.0033	0.0046	0.0091	0.0040	0.0012	0.0003

Tabelle 142: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0322	0.0232	0.0018	0.0009	0.0310	0.0219	0.0016	0.0007	0.1755	0.1748	0.0303	0.0201	0.0014	0.0007
STD		0.0137	0.0042	0.0014	0.0005	0.0125	0.0043	0.0011	0.0003	0.0036	0.0035	0.0125	0.0052	0.0011	0.0003

Tabelle 143: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0408	0.0281	0.0029	0.0014	0.0373	0.0240	0.0023	0.0011	0.1481	0.1470	0.0377	0.0226	0.0022	0.0009
STD		0.0103	0.0078	0.0018	0.0008	0.0080	0.0066	0.0012	0.0006	0.0035	0.0036	0.0110	0.0045	0.0014	0.0004

Tabelle 144: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.71). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.83	0.0467	0.0274	0.0035	0.0015	0.0391	0.0194	0.0026	0.0008	0.1203	0.1188	0.0411	0.0204	0.0026	0.0007
STD		0.0108	0.0083	0.0021	0.0008	0.0081	0.0069	0.0013	0.0005	0.0029	0.0033	0.0079	0.0054	0.0012	0.0004

Tabelle 145: Simulation KNN mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0632	0.0507	0.0060	0.0055	0.0343	0.0136	0.0027	0.0010	0.0681	0.0666	0.0381	0.0173	0.0025	0.0005
STD		0.0155	0.0221	0.0032	0.0044	0.0075	0.0037	0.0012	0.0006	0.0028	0.0039	0.0077	0.0043	0.0010	0.0003

Tabelle 146: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0409	0.0250	0.0026	0.0010	0.0352	0.0192	0.0019	0.0005	0.2028	0.2019	0.0296	0.0220	0.0013	0.0007
STD		0.0126	0.0047	0.0013	0.0006	0.0125	0.0019	0.0011	0.0002	0.0040	0.0038	0.0109	0.0045	0.0009	0.0003

Tabelle 147: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.68). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0312	0.0241	0.0016	0.0011	0.0306	0.0215	0.0016	0.0008	0.1702	0.1697	0.0273	0.0223	0.0012	0.0008
STD		0.0100	0.0069	0.0010	0.0007	0.0107	0.0055	0.0010	0.0005	0.0035	0.0035	0.0106	0.0054	0.0009	0.0006

Tabelle 148: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.82). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0333	0.0284	0.0019	0.0014	0.0306	0.0244	0.0016	0.0010	0.1425	0.1421	0.0282	0.0227	0.0014	0.0008
STD		0.0118	0.0069	0.0011	0.0007	0.0114	0.0048	0.0009	0.0005	0.0030	0.0031	0.0119	0.0035	0.0010	0.0003

Tabelle 149: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.97). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0432	0.0374	0.0030	0.0026	0.0318	0.0239	0.0018	0.0013	0.1132	0.1127	0.0310	0.0234	0.0016	0.0009
STD		0.0142	0.0119	0.0018	0.0016	0.0097	0.0071	0.0010	0.0007	0.0025	0.0028	0.0098	0.0040	0.0010	0.0004

Tabelle 150: Simulation KNN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0486	0.0481	0.0042	0.0043	0.0266	0.0162	0.0018	0.0011	0.0734	0.0731	0.0286	0.0214	0.0014	0.0008
STD		0.0160	0.0144	0.0030	0.0027	0.0069	0.0035	0.0010	0.0005	0.0032	0.0040	0.0080	0.0030	0.0009	0.0003

Tabelle 151: Simulation SVM mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.24). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0896	0.0680	0.0127	0.0095	0.0891	0.0575	0.0112	0.0059	0.1993	0.1976	0.0984	0.0635	0.0137	0.0067
STD		0.0262	0.0231	0.0098	0.0082	0.0215	0.0158	0.0056	0.0035	0.0117	0.0107	0.0161	0.0152	0.0045	0.0032

Tabelle 152: Simulation SVM mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.1101	0.0567	0.0180	0.0055	0.1176	0.0537	0.0180	0.0052	0.1792	0.1685	0.1282	0.0596	0.0228	0.0058
STD		0.0351	0.0121	0.0119	0.0025	0.0320	0.0128	0.0083	0.0026	0.0104	0.0107	0.0240	0.0133	0.0066	0.0025

Tabelle 153: Simulation SVM mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1275	0.0707	0.0224	0.0085	0.1450	0.0571	0.0267	0.0058	0.1615	0.1392	0.1562	0.0533	0.0315	0.0046
STD		0.0165	0.0184	0.0056	0.0045	0.0239	0.0108	0.0074	0.0021	0.0091	0.0108	0.0271	0.0087	0.0091	0.0015

Tabelle 154: Simulation SVM mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.39). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1487	0.0838	0.0294	0.0139	0.1763	0.0528	0.0377	0.0069	0.1459	0.1116	0.1890	0.0480	0.0424	0.0041
STD		0.0166	0.0232	0.0065	0.0075	0.0258	0.0126	0.0099	0.0031	0.0079	0.0109	0.0254	0.0114	0.0101	0.0021

Tabelle 155: Simulation SVM mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.48). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1825	0.1100	0.0428	0.0246	0.2194	0.0414	0.0554	0.0062	0.1225	0.0701	0.2312	0.0335	0.0603	0.0023
STD		0.0168	0.0278	0.0072	0.0135	0.0229	0.0117	0.0102	0.0027	0.0059	0.0103	0.0239	0.0093	0.0110	0.0014

Tabelle 156: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0789	0.0414	0.0102	0.0030	0.0794	0.0396	0.0101	0.0026	0.1992	0.1924	0.0846	0.0439	0.0111	0.0031
STD		0.0176	0.0138	0.0052	0.0020	0.0188	0.0137	0.0051	0.0016	0.0048	0.0084	0.0161	0.0105	0.0051	0.0013

Tabelle 157: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0966	0.0459	0.0133	0.0034	0.1015	0.0462	0.0145	0.0034	0.1876	0.1746	0.1112	0.0481	0.0172	0.0037
STD		0.0152	0.0110	0.0047	0.0013	0.0158	0.0114	0.0051	0.0014	0.0050	0.0090	0.0144	0.0101	0.0053	0.0015

Tabelle 158: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1295	0.0408	0.0221	0.0031	0.1457	0.0360	0.0262	0.0026	0.1701	0.1446	0.1529	0.0413	0.0292	0.0028
STD		0.0110	0.0096	0.0043	0.0014	0.0139	0.0105	0.0053	0.0012	0.0055	0.0098	0.0147	0.0109	0.0063	0.0014

Tabelle 159 Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1535	0.0459	0.0302	0.0042	0.1792	0.0336	0.0379	0.0027	0.1546	0.1162	0.1902	0.0351	0.0422	0.0020
STD		0.0119	0.0165	0.0046	0.0033	0.0164	0.0122	0.0063	0.0020	0.0055	0.0100	0.0143	0.0103	0.0071	0.0011

Tabelle 160: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1756	0.0569	0.0393	0.0061	0.2134	0.0283	0.0523	0.0023	0.1355	0.0823	0.2250	0.0278	0.0573	0.0013
STD		0.0101	0.0175	0.0045	0.0034	0.0152	0.0066	0.0066	0.0011	0.0058	0.0097	0.0152	0.0050	0.0075	0.0005

Tabelle 161: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0718	0.0658	0.0159	0.0145	0.0632	0.0445	0.0062	0.0034	0.2043	0.2018	0.0672	0.0458	0.0069	0.0033
STD		0.0301	0.0313	0.0312	0.0261	0.0148	0.0174	0.0029	0.0023	0.0085	0.0101	0.0145	0.0104	0.0029	0.0014

Tabelle 162: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0829	0.0483	0.0114	0.0040	0.0835	0.0448	0.0105	0.0034	0.1850	0.1783	0.0892	0.0453	0.0116	0.0033
STD		0.0167	0.0133	0.0060	0.0021	0.0169	0.0129	0.0034	0.0014	0.0099	0.0117	0.0149	0.0102	0.0032	0.0015

Tabelle 163: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.1016	0.0443	0.0139	0.0034	0.1117	0.0396	0.0162	0.0027	0.1645	0.1509	0.1171	0.0421	0.0180	0.0027
STD		0.0154	0.0090	0.0032	0.0017	0.0159	0.0083	0.0036	0.0013	0.0105	0.0127	0.0144	0.0063	0.0036	0.0009

Tabelle 164 Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1328	0.0585	0.0227	0.0063	0.1526	0.0370	0.0278	0.0033	0.1380	0.1118	0.1596	0.0344	0.0305	0.0020
STD		0.0143	0.0131	0.0042	0.0030	0.0138	0.0069	0.0040	0.0015	0.0101	0.0125	0.0136	0.0055	0.0042	0.0007

Tabelle 165: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1628	0.0790	0.0334	0.0137	0.1889	0.0268	0.0411	0.0032	0.1087	0.0689	0.1971	0.0245	0.0439	0.0011
STD		0.0110	0.0293	0.0046	0.0098	0.0105	0.0071	0.0049	0.0016	0.0097	0.0113	0.0118	0.0050	0.0051	0.0004

Tabelle 166: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0781	0.0592	0.0179	0.0066	0.0633	0.0431	0.0063	0.0033	0.2076	0.2057	0.0634	0.0481	0.0063	0.0037
STD		0.0341	0.0147	0.0312	0.0029	0.0137	0.0180	0.0026	0.0019	0.0093	0.0104	0.0149	0.0116	0.0025	0.0018

Tabelle 167: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0764	0.0532	0.0097	0.0060	0.0763	0.0472	0.0090	0.0041	0.1875	0.1826	0.0801	0.0484	0.0099	0.0039
STD		0.0164	0.0158	0.0049	0.0048	0.0153	0.0171	0.0040	0.0022	0.0108	0.0126	0.0194	0.0168	0.0044	0.0025

Tabelle 168: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.9). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1032	0.0488	0.0146	0.0049	0.1119	0.0389	0.0168	0.0031	0.1597	0.1463	0.1171	0.0402	0.0179	0.0026
STD		0.0127	0.0171	0.0034	0.0032	0.0153	0.0116	0.0042	0.0016	0.0111	0.0141	0.0143	0.0083	0.0042	0.0011

Tabelle 169 Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.1). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1305	0.0566	0.0224	0.0067	0.1498	0.0312	0.0273	0.0028	0.1325	0.1080	0.1549	0.0338	0.0288	0.0020
STD		0.0146	0.0221	0.0049	0.0047	0.0156	0.0093	0.0058	0.0017	0.0096	0.0135	0.0162	0.0065	0.0060	0.0008

Tabelle 170: Simulation SVM mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1465	0.0778	0.0281	0.0143	0.1714	0.0286	0.0345	0.0033	0.1083	0.0754	0.1800	0.0265	0.0369	0.0013
STD		0.0157	0.0270	0.0063	0.0109	0.0165	0.0056	0.0069	0.0017	0.0093	0.0123	0.0166	0.0046	0.0067	0.0005

Tabelle 171: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0792	0.0310	0.0092	0.0017	0.0718	0.0286	0.0080	0.0013	0.2012	0.1942	0.0745	0.0312	0.0086	0.0015
STD		0.0171	0.0096	0.0035	0.0011	0.0151	0.0080	0.0031	0.0007	0.0029	0.0052	0.0136	0.0049	0.0029	0.0004

Tabelle 172: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.1129	0.0322	0.0171	0.0018	0.1151	0.0315	0.0178	0.0017	0.1821	0.1647	0.1209	0.0325	0.0193	0.0017
STD		0.0112	0.0071	0.0027	0.0009	0.0120	0.0070	0.0028	0.0008	0.0035	0.0059	0.0125	0.0072	0.0032	0.0008

Tabelle 173: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1423	0.0300	0.0259	0.0017	0.1570	0.0265	0.0302	0.0013	0.1654	0.1353	0.1603	0.0291	0.0321	0.0014
STD		0.0098	0.0079	0.0031	0.0010	0.0115	0.0066	0.0042	0.0007	0.0041	0.0061	0.0133	0.0053	0.0044	0.0005

Tabelle 174: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1604	0.0343	0.0328	0.0022	0.1862	0.0238	0.0414	0.0012	0.1509	0.1087	0.1949	0.0235	0.0447	0.0009
STD		0.0087	0.0047	0.0032	0.0007	0.0159	0.0038	0.0054	0.0003	0.0052	0.0059	0.0145	0.0041	0.0060	0.0004

Tabelle 175: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1768	0.0399	0.0397	0.0032	0.2132	0.0206	0.0527	0.0013	0.1375	0.0839	0.2244	0.0196	0.0570	0.0007
STD		0.0072	0.0137	0.0037	0.0020	0.0140	0.0056	0.0064	0.0007	0.0060	0.0056	0.0157	0.0053	0.0071	0.0004

Tabelle 176: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0589	0.0359	0.0055	0.0026	0.0585	0.0289	0.0052	0.0014	0.2046	0.2012	0.0632	0.0353	0.0061	0.0020
STD		0.0158	0.0119	0.0029	0.0021	0.0176	0.0047	0.0028	0.0005	0.0049	0.0045	0.0147	0.0048	0.0027	0.0005

Tabelle 177: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.57). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0911	0.0320	0.0108	0.0017	0.0961	0.0305	0.0119	0.0015	0.1795	0.1688	0.1013	0.0321	0.0131	0.0017
STD		0.0113	0.0082	0.0029	0.0009	0.0122	0.0074	0.0033	0.0008	0.0054	0.0054	0.0125	0.0057	0.0033	0.0007

Tabelle 178: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1135	0.0317	0.0155	0.0019	0.1258	0.0276	0.0183	0.0015	0.1639	0.1468	0.1301	0.0288	0.0196	0.0015
STD		0.0096	0.0098	0.0026	0.0011	0.0088	0.0092	0.0028	0.0010	0.0055	0.0056	0.0122	0.0052	0.0033	0.0006

Tabelle 179: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.77). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1337	0.0341	0.0218	0.0023	0.1547	0.0227	0.0273	0.0012	0.1426	0.1155	0.1615	0.0235	0.0296	0.0009
STD		0.0067	0.0122	0.0021	0.0016	0.0073	0.0069	0.0026	0.0008	0.0060	0.0057	0.0081	0.0057	0.0029	0.0006

Tabelle 180: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1544	0.0473	0.0292	0.0044	0.1826	0.0203	0.0374	0.0014	0.1195	0.0823	0.1894	0.0199	0.0400	0.0007
STD		0.0082	0.0117	0.0027	0.0023	0.0089	0.0042	0.0033	0.0007	0.0068	0.0055	0.0089	0.0043	0.0037	0.0004

Tabelle 181: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.62). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0593	0.0407	0.0052	0.0034	0.0563	0.0319	0.0048	0.0017	0.2013	0.1983	0.0590	0.0341	0.0053	0.0018
STD		0.0090	0.0127	0.0017	0.0027	0.0091	0.0071	0.0016	0.0006	0.0032	0.0038	0.0086	0.0062	0.0015	0.0007

Tabelle 182: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0730	0.0273	0.0074	0.0013	0.0780	0.0248	0.0084	0.0011	0.1861	0.1791	0.0828	0.0320	0.0093	0.0016
STD		0.0094	0.0110	0.0018	0.0010	0.0112	0.0094	0.0022	0.0008	0.0040	0.0048	0.0087	0.0050	0.0019	0.0006

Tabelle 183: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.85). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.1019	0.0358	0.0127	0.0021	0.1143	0.0321	0.0152	0.0017	0.1663	0.1522	0.1174	0.0307	0.0165	0.0016
STD		0.0083	0.0053	0.0015	0.0006	0.0096	0.0057	0.0019	0.0005	0.0045	0.0055	0.0102	0.0054	0.0022	0.0006

Tabelle 184: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=1.05). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1292	0.0296	0.0208	0.0018	0.1499	0.0199	0.0263	0.0009	0.1393	0.1134	0.1579	0.0227	0.0285	0.0009
STD		0.0056	0.0087	0.0017	0.0011	0.0082	0.0058	0.0026	0.0005	0.0063	0.0058	0.0092	0.0046	0.0030	0.0004

Tabelle 185: Simulation SVM mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1509	0.0511	0.0287	0.0047	0.1783	0.0189	0.0362	0.0013	0.1092	0.0737	0.1857	0.0189	0.0383	0.0006
STD		0.0105	0.0119	0.0039	0.0026	0.0092	0.0032	0.0037	0.0005	0.0068	0.0056	0.0080	0.0033	0.0035	0.0002

Tabelle 186: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0884	0.0211	0.0103	0.0007	0.0833	0.0209	0.0097	0.0007	0.1996	0.1904	0.0838	0.0242	0.0101	0.0009
STD		0.0120	0.0045	0.0028	0.0002	0.0130	0.0056	0.0026	0.0003	0.0028	0.0021	0.0107	0.0062	0.0025	0.0004

Tabelle 187: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.28). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.1124	0.0240	0.0193	0.0009	0.1169	0.0230	0.0213	0.0008	0.1827	0.1615	0.1204	0.0266	0.0220	0.0011
STD		0.0084	0.0047	0.0028	0.0005	0.0083	0.0045	0.0031	0.0004	0.0032	0.0016	0.0074	0.0054	0.0030	0.0004

Tabelle 188: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.1511	0.0213	0.0278	0.0007	0.1658	0.0178	0.0325	0.0006	0.1652	0.1318	0.1698	0.0180	0.0345	0.0005
STD		0.0037	0.0045	0.0021	0.0003	0.0055	0.0034	0.0030	0.0002	0.0032	0.0015	0.0064	0.0030	0.0030	0.0002

Tabelle 189: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1670	0.0249	0.0340	0.0012	0.1936	0.0167	0.0429	0.0007	0.1537	0.1093	0.1999	0.0159	0.0462	0.0004
STD		0.0041	0.0083	0.0023	0.0008	0.0073	0.0047	0.0037	0.0004	0.0032	0.0024	0.0088	0.0038	0.0040	0.0002

Tabelle 190: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1881	0.0373	0.0430	0.0026	0.2314	0.0152	0.0588	0.0008	0.1326	0.0713	0.2424	0.0120	0.0644	0.0003
STD		0.0072	0.0083	0.0026	0.0012	0.0114	0.0035	0.0045	0.0004	0.0035	0.0030	0.0114	0.0031	0.0051	0.0002

Tabelle 191: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0757	0.0263	0.0087	0.0012	0.0669	0.0232	0.0059	0.0009	0.1976	0.1926	0.0697	0.0245	0.0065	0.0010
STD		0.0156	0.0051	0.0048	0.0005	0.0081	0.0042	0.0010	0.0003	0.0030	0.0042	0.0079	0.0050	0.0014	0.0004

Tabelle 192: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0955	0.0212	0.0115	0.0008	0.1019	0.0208	0.0128	0.0007	0.1781	0.1656	0.1052	0.0213	0.0138	0.0007
STD		0.0091	0.0058	0.0022	0.0004	0.0074	0.0055	0.0023	0.0004	0.0035	0.0041	0.0088	0.0044	0.0023	0.0003

Tabelle 193: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.7). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.1199	0.0248	0.0178	0.0011	0.1369	0.0209	0.0217	0.0008	0.1598	0.1382	0.1413	0.0189	0.0232	0.0006
STD		0.0095	0.0047	0.0026	0.0004	0.0089	0.0051	0.0026	0.0003	0.0041	0.0041	0.0096	0.0038	0.0031	0.0002

Tabelle 194: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.1424	0.0307	0.0247	0.0018	0.1696	0.0193	0.0318	0.0009	0.1390	0.1072	0.1747	0.0177	0.0340	0.0005
STD		0.0061	0.0087	0.0028	0.0008	0.0106	0.0054	0.0040	0.0004	0.0047	0.0039	0.0095	0.0031	0.0038	0.0002

Tabelle 195: Simulation SVM mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1591	0.0327	0.0315	0.0022	0.1931	0.0134	0.0415	0.0007	0.1180	0.0762	0.2006	0.0126	0.0443	0.0003
STD		0.0046	0.0114	0.0014	0.0015	0.0064	0.0032	0.0027	0.0003	0.0045	0.0036	0.0070	0.0020	0.0029	0.0001

Tabelle 196: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.62). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0635	0.0327	0.0068	0.0032	0.0495	0.0223	0.0036	0.0009	0.2060	0.2035	0.0521	0.0249	0.0040	0.0010
STD		0.0130	0.0129	0.0033	0.0043	0.0096	0.0063	0.0012	0.0005	0.0031	0.0041	0.0084	0.0045	0.0013	0.0003

Tabelle 197: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.75). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0886	0.0264	0.0101	0.0011	0.0857	0.0240	0.0093	0.0009	0.1867	0.1782	0.0885	0.0243	0.0100	0.0009
STD		0.0117	0.0052	0.0030	0.0005	0.0096	0.0044	0.0017	0.0003	0.0034	0.0040	0.0096	0.0043	0.0019	0.0003

Tabelle 198: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.92). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1099	0.0216	0.0153	0.0008	0.1255	0.0192	0.0188	0.0007	0.1616	0.1431	0.1299	0.0179	0.0201	0.0005
STD		0.0082	0.0051	0.0022	0.0003	0.0082	0.0056	0.0025	0.0003	0.0037	0.0039	0.0085	0.0032	0.0028	0.0002

Tabelle 199: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=1.08). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1314	0.0260	0.0215	0.0013	0.1575	0.0170	0.0278	0.0007	0.1394	0.1115	0.1635	0.0178	0.0299	0.0005
STD		0.0104	0.0080	0.0030	0.0008	0.0097	0.0047	0.0036	0.0004	0.0036	0.0037	0.0095	0.0031	0.0037	0.0002

Tabelle 200: Simulation SVM mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1529	0.0352	0.0290	0.0024	0.1850	0.0144	0.0380	0.0007	0.1126	0.0746	0.1914	0.0127	0.0403	0.0003
STD		0.0063	0.0094	0.0017	0.0011	0.0073	0.0023	0.0024	0.0003	0.0045	0.0036	0.0063	0.0014	0.0023	0.0001

Tabelle 201: Simulation NN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.1209	/	0.0218	/	0.1244	/	0.0220	/	0.2114	0.2030	0.1223	0.0728	0.0221	0.0084
STD		0.0318	/	0.0097	/	0.0272	/	0.0073	/	0.0172	0.0120	0.0283	0.0154	0.0082	0.0032

Tabelle 202: Simulation NN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.1065	/	0.0168	/	0.0997	/	0.0142	/	0.1814	0.1791	0.0994	0.0689	0.0144	0.0075
STD		0.0272	/	0.0078	/	0.0202	/	0.0053	/	0.0139	0.0113	0.0169	0.0138	0.0043	0.0030

Tabelle 203: Simulation NN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1093	0.0773	0.0178	0.0109	0.0850	0.0530	0.0114	0.0057	0.1482	0.1484	0.0831	0.0619	0.0111	0.0063
STD		0.0242	0.0147	0.0089	0.0037	0.0136	0.0129	0.0038	0.0020	0.0142	0.0118	0.0183	0.0074	0.0049	0.0015

Tabelle 204: Simulation NN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1016	0.0997	0.0199	0.0196	0.0673	0.0519	0.0086	0.0069	0.1194	0.1212	0.0665	0.0594	0.0075	0.0059
STD		0.0263	0.0248	0.0127	0.0109	0.0151	0.0075	0.0040	0.0022	0.0125	0.0106	0.0148	0.0095	0.0031	0.0020

Tabelle 205: Simulation NN mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.48). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1563	0.1526	0.0486	0.0480	0.0528	0.0441	0.0095	0.0080	0.0779	0.0815	0.0499	0.0584	0.0058	0.0059
STD		0.0467	0.0364	0.0293	0.0230	0.0147	0.0123	0.0037	0.0030	0.0111	0.0093	0.0145	0.0146	0.0046	0.0029

Tabelle 206: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.19). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0893	/	0.0116	/	0.0916	/	0.0119	/	0.2127	0.2053	0.0874	0.0451	0.0112	0.0032
STD		0.0224	/	0.0050	/	0.0232	/	0.0051	/	0.0117	0.0073	0.0219	0.0112	0.0050	0.0013

Tabelle 207: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.754	0.0687	0.0425	0.0070	0.0030	0.0735	0.0323	0.0076	0.0018	0.1865	0.1825	0.0718	0.0450	0.0078	0.0032
STD		0.0162	0.0015	0.0031	0.0005	0.0198	0.0006	0.0038	0.0000	0.0128	0.0100	0.0208	0.0073	0.0039	0.0009

Tabelle 208: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0580	0.0539	0.0050	0.0052	0.0567	0.0378	0.0046	0.0030	0.1451	0.1441	0.0544	0.0375	0.0046	0.0023
STD		0.0156	0.0064	0.0024	0.0020	0.0163	0.0061	0.0024	0.0008	0.0132	0.0115	0.0144	0.0075	0.0022	0.0008

Tabelle 209 Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0616	0.0615	0.0062	0.0083	0.0450	0.0383	0.0036	0.0036	0.1174	0.1180	0.0442	0.0440	0.0031	0.0031
STD		0.0186	0.0187	0.0033	0.0068	0.0142	0.0075	0.0018	0.0017	0.0123	0.0110	0.0131	0.0069	0.0016	0.0010

Tabelle 210: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0760	0.1019	0.0115	0.0197	0.0361	0.0326	0.0038	0.0043	0.0812	0.0829	0.0336	0.0385	0.0022	0.0025
STD		0.0192	0.0245	0.0053	0.0089	0.0110	0.0084	0.0016	0.0018	0.0105	0.0098	0.0094	0.0055	0.0010	0.0008

Tabelle 211: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0938	/	0.0118	/	0.0964	/	0.0122	/	0.2089	0.2026	0.0912	0.0499	0.0119	0.0040
STD		0.0215	/	0.0046	/	0.0205	/	0.0044	/	0.0132	0.0112	0.0206	0.0141	0.0045	0.0021

Tabelle 212: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0782	0.0426	0.0090	0.0032	0.0809	0.0257	0.0090	0.0017	0.1803	0.1770	0.0777	0.0489	0.0086	0.0039
STD		0.0107	/	0.0024	/	0.0089	/	0.0026	/	0.0118	0.0103	0.0059	0.0119	0.0020	0.0018

Tabelle 213: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0670	0.0528	0.0068	0.0052	0.0641	0.0395	0.0058	0.0034	0.1499	0.1491	0.0629	0.0457	0.0057	0.0032
STD		0.0119	0.0155	0.0023	0.0031	0.0129	0.0121	0.0020	0.0015	0.0151	0.0134	0.0117	0.0095	0.0018	0.0013

Tabelle 214 Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.78). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.069	0.070	0.009	0.009	0.050	0.037	0.005	0.003	0.116	0.117	0.051	0.045	0.004	0.003
STD		0.026	0.019	0.008	0.006	0.016	0.010	0.003	0.002	0.012	0.011	0.012	0.007	0.002	0.001

Tabelle 215: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0916	0.1084	0.0158	0.0211	0.0420	0.0366	0.0046	0.0046	0.0812	0.0827	0.0417	0.0425	0.0037	0.0031
STD		0.0235	0.0287	0.0074	0.0110	0.0144	0.0115	0.0023	0.0023	0.0126	0.0119	0.0128	0.0103	0.0021	0.0014

Tabelle 216: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0983	/	0.0121	/	0.1011	/	0.0128	/	0.2138	0.2067	0.0937	0.0496	0.0118	0.0038
STD		0.0186	/	0.0041	/	0.0200	/	0.0046	/	0.0132	0.0100	0.0200	0.0090	0.0042	0.0013

Tabelle 217: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.68). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.73	0.0853	/	0.0102	/	0.0890	/	0.0107	/	0.1930	0.1880	0.0859	0.0552	0.0109	0.0048
STD		0.0168	/	0.0034	/	0.0193	/	0.0042	/	0.0146	0.0119	0.0220	0.0132	0.0053	0.0022

Tabelle 218: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.87). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0668	0.0643	0.0065	0.0164	0.0620	0.0367	0.0055	0.0032	0.1486	0.1476	0.0608	0.0446	0.0056	0.0032
STD		0.0131	0.0345	0.0030	0.0305	0.0139	0.0081	0.0026	0.0015	0.0143	0.0121	0.0137	0.0072	0.0024	0.0010

Tabelle 219 Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.83	0.0872	0.0702	0.0130	0.0097	0.0658	0.0409	0.0074	0.0040	0.1276	0.1272	0.0601	0.0481	0.0057	0.0037
STD		0.0201	0.0106	0.0054	0.0033	0.0163	0.0097	0.0028	0.0013	0.0151	0.0130	0.0165	0.0092	0.0023	0.0013

Tabelle 220: Simulation NN mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1119	0.1382	0.0233	0.0361	0.0407	0.0324	0.0052	0.0050	0.0718	0.0735	0.0400	0.0446	0.0035	0.0034
STD		0.0277	0.0512	0.0118	0.0255	0.0126	0.0116	0.0024	0.0028	0.0105	0.0105	0.0095	0.0069	0.0016	0.0011

Tabelle 221: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0785	/	0.0084	/	0.0835	/	0.0091	/	0.1995	0.1934	0.0808	0.0349	0.0089	0.0020
STD		0.0085	/	0.0020	/	0.0084	/	0.0021	/	0.0079	0.0060	0.0099	0.0079	0.0025	0.0010

Tabelle 222: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0421	0.0205	0.0028	0.0007	0.0433	0.0204	0.0029	0.0007	0.1700	0.1681	0.0460	0.0234	0.0031	0.0009
STD		0.0080	0.0061	0.0010	0.0003	0.0073	0.0065	0.0010	0.0004	0.0038	0.0045	0.0062	0.0049	0.0009	0.0003

Tabelle 223: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0582	0.0420	0.0046	0.0034	0.0575	0.0325	0.0042	0.0021	0.1487	0.1476	0.0552	0.0362	0.0043	0.0020
STD		0.0138	0.0176	0.0021	0.0030	0.0107	0.0131	0.0014	0.0017	0.0082	0.0073	0.0120	0.0075	0.0016	0.0009

Tabelle 224: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0570	0.0583	0.0055	0.0060	0.0451	0.0374	0.0032	0.0029	0.1215	0.1220	0.0433	0.0386	0.0028	0.0024
STD		0.0160	0.0175	0.0029	0.0034	0.0064	0.0107	0.0011	0.0016	0.0057	0.0060	0.0073	0.0097	0.0011	0.0013

Tabelle 225: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0515	0.0743	0.0046	0.0096	0.0290	0.0321	0.0019	0.0033	0.0871	0.0894	0.0285	0.0386	0.0016	0.0026
STD		0.0075	0.0189	0.0021	0.0037	0.0060	0.0119	0.0006	0.0017	0.0062	0.0058	0.0062	0.0065	0.0007	0.0010

Tabelle 226: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0808	/	0.0094	/	0.0856	/	0.0102	/	0.2092	0.2017	0.0810	0.0332	0.0092	0.0018
STD		0.0170	/	0.0033	/	0.0176	/	0.0035	/	0.0084	0.0053	0.0185	0.0078	0.0037	0.0008

Tabelle 227: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0732	0.0329	0.0071	0.0020	0.0767	0.0256	0.0074	0.0016	0.1813	0.1770	0.0741	0.0380	0.0072	0.0022
STD		0.0131	/	0.0021	/	0.0112	/	0.0019	/	0.0073	0.0055	0.0109	0.0079	0.0019	0.0010

Tabelle 228: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0630	0.0469	0.0057	0.0038	0.0621	0.0344	0.0052	0.0024	0.1528	0.1512	0.0596	0.0430	0.0050	0.0028
STD		0.0100	0.0103	0.0019	0.0021	0.0084	0.0055	0.0013	0.0011	0.0061	0.0052	0.0108	0.0075	0.0015	0.0009

Tabelle 229: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0521	0.0508	0.0040	0.0040	0.0436	0.0347	0.0027	0.0021	0.1194	0.1196	0.0417	0.0361	0.0026	0.0020
STD		0.0141	0.0111	0.0024	0.0016	0.0104	0.0079	0.0012	0.0008	0.0043	0.0041	0.0086	0.0075	0.0010	0.0007

Tabelle 230: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.0582	0.0651	0.0058	0.0089	0.0365	0.0298	0.0025	0.0027	0.0903	0.0914	0.0350	0.0344	0.0020	0.0020
STD		0.0182	0.0260	0.0042	0.0063	0.0078	0.0129	0.0011	0.0015	0.0055	0.0054	0.0082	0.0102	0.0009	0.0012

Tabelle 231: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.57). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0819	/	0.0091	/	0.0875	/	0.0100	/	0.2035	0.1963	0.0842	0.0332	0.0095	0.0018
STD		0.0130	/	0.0030	/	0.0129	/	0.0032	/	0.0068	0.0051	0.0128	0.0073	0.0031	0.0008

Tabelle 232: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.66). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0779	0.0474	0.0087	0.0040	0.0821	0.0349	0.0089	0.0024	0.1831	0.1782	0.0782	0.0397	0.0085	0.0025
STD		0.0176	/	0.0039	/	0.0181	/	0.0038	/	0.0094	0.0067	0.0145	0.0081	0.0033	0.0011

Tabelle 233: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.81). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0609	0.0565	0.0056	0.0060	0.0604	0.0351	0.0050	0.0030	0.1506	0.1491	0.0574	0.0406	0.0048	0.0025
STD		0.0084	0.0093	0.0015	0.0024	0.0091	0.0105	0.0013	0.0008	0.0073	0.0067	0.0093	0.0116	0.0013	0.0011

Tabelle 234: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0607	0.0524	0.0055	0.0048	0.0513	0.0361	0.0037	0.0024	0.1240	0.1235	0.0499	0.0375	0.0036	0.0021
STD		0.0106	0.0151	0.0016	0.0029	0.0107	0.0080	0.0012	0.0011	0.0083	0.0075	0.0098	0.0074	0.0013	0.0007

Tabelle 235: Simulation NN mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0625	0.0796	0.0074	0.0119	0.0334	0.0278	0.0025	0.0029	0.0816	0.0830	0.0325	0.0375	0.0019	0.0023
STD		0.0131	0.0156	0.0037	0.0051	0.0078	0.0076	0.0009	0.0012	0.0070	0.0065	0.0090	0.0088	0.0009	0.0010

Tabelle 236: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0685	/	0.0063	/	0.0718	/	0.0068	/	0.2051	0.2000	0.0700	0.0262	0.0066	0.0011
STD		0.0084	/	0.0015	/	0.0091	#/	0.0016	/	0.0053	0.0040	0.0092	0.0066	0.0018	0.0005

Tabelle 237: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0609	/	0.0049	/	0.0658	/	0.0054	/	0.1816	0.1782	0.0656	0.0307	0.0055	0.0014
STD		0.0072	/	0.0010	/	0.0081	/	0.0013	/	0.0058	0.0048	0.0082	0.0042	0.0013	0.0003

Tabelle 238: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0499	0.0468	0.0034	0.0050	0.0548	0.0285	0.0036	0.0016	0.1453	0.1438	0.0539	0.0333	0.0037	0.0016
STD		0.0072	0.0170	0.0009	0.0061	0.0051	0.0049	0.0006	0.0005	0.0050	0.0045	0.0063	0.0046	0.0009	0.0004

Tabelle 239: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0490	0.0503	0.0034	0.0040	0.0452	0.0357	0.0026	0.0022	0.1215	0.1217	0.0435	0.0365	0.0026	0.0020
STD		0.0091	0.0086	0.0011	0.0011	0.0062	0.0088	0.0007	0.0006	0.0053	0.0052	0.0061	0.0069	0.0007	0.0006

Tabelle 240: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0397	0.0605	0.0027	0.0058	0.0278	0.0236	0.0013	0.0018	0.0835	0.0852	0.0258	0.0341	0.0011	0.0018
STD		0.0090	0.0139	0.0013	0.0028	0.0051	0.0078	0.0006	0.0008	0.0054	0.0055	0.0060	0.0074	0.0005	0.0008

Tabelle 241: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0728	/	0.0067	/	0.0748	/	0.0070	/	0.2074	0.2020	0.0712	0.0259	0.0066	0.0011
STD		0.0125	/	0.0023	/	0.0125	/	0.0024	/	0.0058	0.0042	0.0125	0.0066	0.0023	0.0006

Tabelle 242: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.77	0.0625	/	0.0050	/	0.0647	/	0.0050	/	0.1697	0.1671	0.0642	0.0341	0.0051	0.0017
STD		0.0085	/	0.0013	/	0.0067	/	0.0010	/	0.0055	0.0050	0.0073	0.0054	0.0010	0.0005

Tabelle 243: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0554	0.0430	0.0043	0.0029	0.0564	0.0318	0.0040	0.0018	0.1459	0.1446	0.0553	0.0361	0.0040	0.0019
STD		0.0116	0.0071	0.0016	0.0008	0.0105	0.0047	0.0013	0.0005	0.0064	0.0059	0.0111	0.0031	0.0015	0.0003

Tabelle 244: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.73). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0547	0.0524	0.0041	0.0048	0.0456	0.0341	0.0026	0.0023	0.1147	0.1148	0.0445	0.0393	0.0027	0.0022
STD		0.0121	0.0110	0.0016	0.0019	0.0080	0.0089	0.0009	0.0008	0.0044	0.0042	0.0104	0.0036	0.0011	0.0005

Tabelle 245: Simulation NN mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0480	0.0762	0.0039	0.0103	0.0266	0.0191	0.0012	0.0020	0.0684	0.0703	0.0245	0.0335	0.0011	0.0018
STD		0.0103	0.0229	0.0017	0.0061	0.0043	0.0050	0.0004	0.0008	0.0026	0.0028	0.0035	0.0039	0.0004	0.0003

Tabelle 246: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0711	/	0.0065	/	0.0740	/	0.0068	/	0.2053	0.2000	0.0714	0.0260	0.0067	0.0010
STD		0.0109	/	0.0016	/	0.0110	/	0.0017	/	0.0052	0.0041	0.0123	0.0055	0.0018	0.0004

Tabelle 247: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.67). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0641	/	0.0055	/	0.0680	/	0.0057	/	0.1750	0.1719	0.0683	0.0330	0.0059	0.0016
STD		0.0114	/	0.0015	/	0.0101	/	0.0013	/	0.0060	0.0054	0.0105	0.0077	0.0016	0.0006

Tabelle 248: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.82). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0600	0.0473	0.0048	0.0034	0.0579	0.0327	0.0041	0.0021	0.1445	0.1434	0.0597	0.0389	0.0044	0.0023
STD		0.0076	0.0035	0.0014	0.0009	0.0052	0.0076	0.0009	0.0008	0.0038	0.0035	0.0055	0.0039	0.0007	0.0005

Tabelle 249: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.97). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0555	0.0619	0.0045	0.0060	0.0472	0.0398	0.0029	0.0027	0.1159	0.1163	0.0478	0.0400	0.0030	0.0023
STD		0.0147	0.0140	0.0025	0.0026	0.0065	0.0082	0.0010	0.0010	0.0045	0.0042	0.0073	0.0056	0.0010	0.0007

Tabelle 250: Simulation NN mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0497	0.0636	0.0042	0.0069	0.0311	0.0217	0.0016	0.0018	0.0775	0.0790	0.0300	0.0348	0.0015	0.0019
STD		0.0097	0.0112	0.0014	0.0019	0.0056	0.0035	0.0005	0.0004	0.0043	0.0046	0.0075	0.0038	0.0007	0.0004

Tabelle 251: Simulation LDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.1078	/	0.0176	/	0.1079	/	0.0169	/	0.2030	0.1977	0.1091	0.0770	0.0183	0.0096
STD		0.0127	/	0.0053	/	0.0099	/	0.0044	/	0.0121	0.0092	0.0177	0.0229	0.0053	0.0050

Tabelle 252: Simulation LDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0977	0.0720	0.0151	0.0085	0.0920	0.0476	0.0125	0.0049	0.1712	0.1693	0.0944	0.0708	0.0140	0.0079
STD		0.0262	/	0.0075	/	0.0148	/	0.0038	/	0.0121	0.0101	0.0153	0.0183	0.0036	0.0038

Tabelle 253: Simulation LDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.1049	0.0779	0.0178	0.0111	0.0854	0.0511	0.0116	0.0055	0.1412	0.1414	0.0797	0.0662	0.0111	0.0071
STD		0.0280	0.0228	0.0085	0.0053	0.0191	0.0178	0.0051	0.0027	0.0119	0.0107	0.0142	0.0157	0.0037	0.0028

Tabelle 254: Simulation LDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1162	0.1027	0.0267	0.0206	0.0717	0.0495	0.0102	0.0067	0.1145	0.1162	0.0663	0.0619	0.0087	0.0061
STD		0.0297	0.0265	0.0141	0.0116	0.0133	0.0123	0.0040	0.0028	0.0114	0.0100	0.0134	0.0115	0.0038	0.0024

Tabelle 255: Simulation LDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1563	0.1724	0.0438	0.0569	0.0577	0.0442	0.0102	0.0093	0.0754	0.0786	0.0505	0.0582	0.0066	0.0056
STD		0.0215	0.0479	0.0153	0.0230	0.0129	0.0192	0.0038	0.0054	0.0105	0.0095	0.0128	0.0065	0.0036	0.0019

Tabelle 256: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0684	0.0504	0.0075	0.0044	0.0683	0.0429	0.0075	0.0033	0.1908	0.1891	0.0647	0.0525	0.0068	0.0045
STD		0.0265	0.0079	0.0048	0.0020	0.0264	0.0064	0.0049	0.0012	0.0132	0.0103	0.0234	0.0146	0.0047	0.0021

Tabelle 257: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0605	0.0588	0.0063	0.0061	0.0595	0.0524	0.0060	0.0051	0.1784	0.1774	0.0625	0.0517	0.0063	0.0043
STD		0.0239	0.0168	0.0046	0.0025	0.0236	0.0183	0.0046	0.0021	0.0134	0.0108	0.0204	0.0146	0.0042	0.0021

Tabelle 258: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0633	0.0598	0.0068	0.0066	0.0550	0.0435	0.0051	0.0039	0.1419	0.1424	0.0527	0.0440	0.0045	0.0033
STD		0.0111	0.0218	0.0030	0.0045	0.0139	0.0132	0.0021	0.0021	0.0135	0.0118	0.0144	0.0115	0.0025	0.0014

Tabelle 259 Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0667	0.0787	0.0097	0.0136	0.0464	0.0413	0.0049	0.0049	0.1139	0.1150	0.0458	0.0427	0.0036	0.0032
STD		0.0215	0.0401	0.0085	0.0150	0.0148	0.0152	0.0033	0.0040	0.0132	0.0120	0.0149	0.0064	0.0026	0.0011

Tabelle 260: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0771	0.1095	0.0108	0.0247	0.0388	0.0352	0.0036	0.0051	0.0804	0.0815	0.0363	0.0375	0.0026	0.0024
STD		0.0310	0.0404	0.0077	0.0192	0.0132	0.0111	0.0022	0.0032	0.0125	0.0122	0.0148	0.0088	0.0022	0.0011

Tabelle 261: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0609	0.0451	0.0057	0.0030	0.0617	0.0386	0.0058	0.0022	0.1903	0.1904	0.0612	0.0465	0.0060	0.0034
STD		0.0132	0.0040	0.0026	0.0004	0.0134	0.0063	0.0026	0.0005	0.0112	0.0098	0.0137	0.0069	0.0027	0.0008

Tabelle 262: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0541	0.0527	0.0053	0.0047	0.0511	0.0438	0.0047	0.0035	0.1638	0.1649	0.0568	0.0495	0.0050	0.0037
STD		0.0106	0.0129	0.0015	0.0019	0.0123	0.0093	0.0015	0.0012	0.0115	0.0104	0.0123	0.0048	0.0019	0.0006

Tabelle 263: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0593	0.0537	0.0059	0.0049	0.0488	0.0416	0.0043	0.0032	0.1378	0.1395	0.0473	0.0460	0.0037	0.0033
STD		0.0159	0.0080	0.0026	0.0019	0.0121	0.0068	0.0014	0.0007	0.0114	0.0104	0.0120	0.0065	0.0017	0.0007

Tabelle 264 Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.0644	0.0712	0.0074	0.0099	0.0402	0.0397	0.0036	0.0040	0.1021	0.1042	0.0398	0.0433	0.0029	0.0030
STD		0.0104	0.0162	0.0021	0.0049	0.0108	0.0074	0.0011	0.0015	0.0107	0.0102	0.0084	0.0051	0.0013	0.0007

Tabelle 265: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0845	0.1207	0.0150	0.0303	0.0303	0.0310	0.0036	0.0050	0.0638	0.0662	0.0296	0.0386	0.0022	0.0026
STD		0.0209	0.0260	0.0080	0.0105	0.0079	0.0057	0.0016	0.0014	0.0094	0.0099	0.0060	0.0073	0.0010	0.0009

Tabelle 266: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0600	0.0515	0.0056	0.0038	0.0610	0.0458	0.0057	0.0031	0.1909	0.1910	0.0613	0.0462	0.0060	0.0033
STD		0.0129	0.0035	0.0025	0.0002	0.0134	0.0056	0.0025	0.0004	0.0113	0.0098	0.0141	0.0070	0.0027	0.0008

Tabelle 267: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0562	0.0576	0.0056	0.0053	0.0533	0.0478	0.0051	0.0040	0.1672	0.1681	0.0582	0.0500	0.0053	0.0038
STD		0.0150	0.0152	0.0024	0.0024	0.0149	0.0117	0.0021	0.0017	0.0116	0.0104	0.0122	0.0061	0.0019	0.0008

Tabelle 268: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.9). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0520	0.0538	0.0050	0.0046	0.0435	0.0415	0.0037	0.0030	0.1323	0.1342	0.0455	0.0453	0.0035	0.0032
STD		0.0115	0.0110	0.0019	0.0019	0.0114	0.0091	0.0013	0.0011	0.0113	0.0104	0.0116	0.0071	0.0017	0.0008

Tabelle 269 Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.1). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.0651	0.0792	0.0082	0.0130	0.0388	0.0398	0.0036	0.0044	0.0976	0.0997	0.0391	0.0427	0.0029	0.0030
STD		0.0179	0.0162	0.0041	0.0056	0.0106	0.0057	0.0013	0.0012	0.0105	0.0102	0.0085	0.0054	0.0013	0.0008

Tabelle 270: Simulation LDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0814	0.1201	0.0138	0.0288	0.0323	0.0327	0.0038	0.0049	0.0688	0.0711	0.0313	0.0403	0.0023	0.0028
STD		0.0199	0.0329	0.0061	0.0160	0.0074	0.0067	0.0013	0.0023	0.0096	0.0100	0.0067	0.0074	0.0011	0.0010

Tabelle 271: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0366	0.0398	0.0022	0.0022	0.0367	0.0336	0.0022	0.0016	0.1938	0.1938	0.0385	0.0314	0.0023	0.0016
STD		0.0093	0.0036	0.0010	0.0006	0.0098	0.0035	0.0011	0.0006	0.0061	0.0057	0.0059	0.0060	0.0007	0.0006

Tabelle 272: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0368	0.0421	0.0020	0.0029	0.0365	0.0342	0.0020	0.0020	0.1704	0.1709	0.0378	0.0352	0.0022	0.0019
STD		0.0069	0.0127	0.0008	0.0014	0.0060	0.0110	0.0007	0.0010	0.0059	0.0058	0.0069	0.0078	0.0008	0.0008

Tabelle 273: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0407	0.0469	0.0027	0.0039	0.0335	0.0387	0.0020	0.0027	0.1363	0.1377	0.0335	0.0396	0.0019	0.0025
STD		0.0063	0.0134	0.0011	0.0031	0.0034	0.0094	0.0006	0.0016	0.0055	0.0057	0.0059	0.0086	0.0007	0.0010

Tabelle 274: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0437	0.0646	0.0037	0.0064	0.0289	0.0399	0.0019	0.0031	0.1054	0.1078	0.0277	0.0424	0.0014	0.0028
STD		0.0119	0.0150	0.0018	0.0028	0.0060	0.0066	0.0008	0.0009	0.0049	0.0051	0.0059	0.0066	0.0006	0.0008

Tabelle 275: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0494	0.0861	0.0053	0.0134	0.0240	0.0303	0.0018	0.0035	0.0789	0.0819	0.0225	0.0410	0.0010	0.0028
STD		0.0130	0.0168	0.0035	0.0053	0.0040	0.0070	0.0010	0.0011	0.0043	0.0044	0.0044	0.0050	0.0005	0.0007

Tabelle 276: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0371	0.0310	0.0023	0.0017	0.0371	0.0235	0.0022	0.0009	0.1921	0.1921	0.0386	0.0357	0.0023	0.0019
STD		0.0097	0.0203	0.0010	0.0017	0.0093	0.0120	0.0009	0.0008	0.0062	0.0054	0.0070	0.0074	0.0008	0.0007

Tabelle 277: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0343	0.0381	0.0021	0.0026	0.0340	0.0320	0.0020	0.0020	0.1659	0.1665	0.0340	0.0356	0.0019	0.0020
STD		0.0090	0.0067	0.0010	0.0010	0.0087	0.0060	0.0009	0.0008	0.0064	0.0058	0.0055	0.0060	0.0006	0.0007

Tabelle 278: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0375	0.0456	0.0027	0.0034	0.0326	0.0379	0.0021	0.0024	0.1401	0.1412	0.0312	0.0375	0.0017	0.0022
STD		0.0093	0.0118	0.0014	0.0021	0.0073	0.0104	0.0010	0.0013	0.0064	0.0061	0.0062	0.0088	0.0006	0.0009

Tabelle 279: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.7). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0329	0.0484	0.0019	0.0038	0.0249	0.0361	0.0013	0.0023	0.1160	0.1174	0.0258	0.0358	0.0011	0.0020
STD		0.0049	0.0107	0.0008	0.0015	0.0055	0.0067	0.0004	0.0007	0.0063	0.0061	0.0065	0.0074	0.0005	0.0007

Tabelle 280: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0449	0.0622	0.0039	0.0067	0.0226	0.0263	0.0016	0.0022	0.0824	0.0841	0.0223	0.0356	0.0009	0.0020
STD		0.0063	0.0162	0.0012	0.0028	0.0043	0.0072	0.0004	0.0009	0.0059	0.0057	0.0069	0.0058	0.0005	0.0007

Tabelle 281: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.53).

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0367	0.0335	0.0022	0.0018	0.0366	0.0259	0.0022	0.0010	0.1933	0.1933	0.0388	0.0359	0.0024	0.0020
STD		0.0088	0.0148	0.0010	0.0014	0.0087	0.0071	0.0009	0.0004	0.0062	0.0054	0.0074	0.0075	0.0009	0.0007

Tabelle 282: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.62).

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0338	0.0392	0.0018	0.0025	0.0344	0.0306	0.0019	0.0016	0.1757	0.1761	0.0355	0.0354	0.0020	0.0020
STD		0.0089	0.0071	0.0007	0.0007	0.0092	0.0053	0.0007	0.0005	0.0063	0.0057	0.0055	0.0057	0.0005	0.0006

Tabelle 283: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.78). **

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0378	0.0469	0.0026	0.0041	0.0334	0.0376	0.0021	0.0027	0.1447	0.1457	0.0319	0.0374	0.0018	0.0022
STD		0.0099	0.0116	0.0011	0.0023	0.0080	0.0103	0.0008	0.0015	0.0064	0.0060	0.0059	0.0085	0.0006	0.0009

Tabelle 285: Simulation LDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=1.16). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0335	0.0463	0.0019	0.0036	0.0236	0.0311	0.0011	0.0019	0.1055	0.1071	0.0246	0.0359	0.0010	0.0020
STD		0.0062	0.0071	0.0006	0.0011	0.0064	0.0074	0.0004	0.0007	0.0062	0.0060	0.0071	0.0070	0.0005	0.0007

Tabelle 286: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0242	0.0328	0.0010	0.0018	0.0255	0.0261	0.0011	0.0011	0.1909	0.1916	0.0287	0.0278	0.0013	0.0012
STD		0.0047	0.0056	0.0003	0.0005	0.0046	0.0074	0.0003	0.0005	0.0047	0.0046	0.0044	0.0060	0.0004	0.0005

Tabelle 287: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0261	0.0345	0.0011	0.0018	0.0259	0.0287	0.0011	0.0014	0.1680	0.1691	0.0270	0.0301	0.0011	0.0014
STD		0.0039	0.0060	0.0003	0.0007	0.0038	0.0045	0.0003	0.0004	0.0048	0.0048	0.0042	0.0051	0.0003	0.0004

Tabelle 288: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0257	0.0410	0.0012	0.0024	0.0216	0.0359	0.0009	0.0018	0.1342	0.1357	0.0216	0.0332	0.0008	0.0016
STD		0.0074	0.0079	0.0007	0.0009	0.0056	0.0077	0.0005	0.0007	0.0048	0.0051	0.0046	0.0053	0.0003	0.0005

Tabelle 289: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0281	0.0463	0.0014	0.0031	0.0201	0.0333	0.0009	0.0019	0.1131	0.1149	0.0190	0.0350	0.0006	0.0018
STD		0.0087	0.0103	0.0007	0.0012	0.0051	0.0109	0.0004	0.0008	0.0048	0.0052	0.0042	0.0073	0.0003	0.0007

Tabelle 290: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0394	0.0680	0.0032	0.0076	0.0178	0.0248	0.0011	0.0023	0.0767	0.0790	0.0165	0.0352	0.0006	0.0020
STD		0.0089	0.0171	0.0015	0.0032	0.0034	0.0077	0.0004	0.0009	0.0043	0.0051	0.0032	0.0078	0.0002	0.0009

Tabelle 291: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0249	0.0305	0.0010	0.0017	0.0250	0.0221	0.0010	0.0009	0.1879	0.1885	0.0258	0.0258	0.0011	0.0011
STD		0.0055	0.0080	0.0004	0.0008	0.0053	0.0053	0.0004	0.0004	0.0048	0.0051	0.0039	0.0049	0.0004	0.0004

Tabelle 292: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0214	0.0305	0.0008	0.0015	0.0207	0.0260	0.0007	0.0012	0.1620	0.1630	0.0230	0.0271	0.0009	0.0012
STD		0.0055	0.0071	0.0004	0.0006	0.0054	0.0061	0.0003	0.0005	0.0052	0.0057	0.0038	0.0054	0.0003	0.0005

Tabelle 293: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0271	0.0400	0.0014	0.0027	0.0225	0.0328	0.0011	0.0019	0.1365	0.1378	0.0229	0.0311	0.0009	0.0016
STD		0.0078	0.0110	0.0008	0.0017	0.0069	0.0078	0.0006	0.0010	0.0052	0.0057	0.0059	0.0051	0.0004	0.0006

Tabelle 294: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0267	0.0460	0.0014	0.0032	0.0181	0.0289	0.0008	0.0016	0.1080	0.1096	0.0195	0.0319	0.0007	0.0016
STD		0.0087	0.0096	0.0008	0.0015	0.0052	0.0030	0.0004	0.0005	0.0049	0.0053	0.0053	0.0035	0.0004	0.0004

Tabelle 295: Simulation LDA mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.0377	0.0494	0.0025	0.0043	0.0178	0.0197	0.0010	0.0013	0.0792	0.0810	0.0156	0.0298	0.0005	0.0015
STD		0.0108	0.0146	0.0015	0.0022	0.0034	0.0039	0.0004	0.0005	0.0043	0.0048	0.0039	0.0033	0.0003	0.0003

Tabelle 296: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0260	0.0308	0.0011	0.0017	0.0258	0.0235	0.0011	0.0009	0.1947	0.1952	0.0267	0.0265	0.0012	0.0011
STD		0.0061	0.0045	0.0006	0.0006	0.0052	0.0043	0.0005	0.0003	0.0047	0.0049	0.0038	0.0043	0.0004	0.0004

Tabelle 297: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0218	0.0286	0.0008	0.0013	0.0214	0.0242	0.0007	0.0010	0.1659	0.1668	0.0232	0.0267	0.0009	0.0011
STD		0.0055	0.0059	0.0004	0.0005	0.0055	0.0051	0.0004	0.0004	0.0052	0.0057	0.0042	0.0054	0.0003	0.0005

Tabelle 298: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0270	0.0394	0.0014	0.0026	0.0225	0.0324	0.0011	0.0018	0.1372	0.1385	0.0230	0.0309	0.0009	0.0016
STD		0.0076	0.0114	0.0007	0.0017	0.0066	0.0078	0.0005	0.0010	0.0052	0.0057	0.0058	0.0052	0.0004	0.0006

Tabelle 299: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0269	0.0449	0.0014	0.0031	0.0184	0.0297	0.0008	0.0016	0.1105	0.1121	0.0199	0.0320	0.0007	0.0016
STD		0.0087	0.0109	0.0009	0.0015	0.0055	0.0038	0.0004	0.0006	0.0049	0.0054	0.0054	0.0035	0.0004	0.0004

Tabelle 300: Simulation LDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0344	0.0536	0.0021	0.0049	0.0157	0.0188	0.0008	0.0013	0.0728	0.0746	0.0148	0.0290	0.0005	0.0014
STD		0.0086	0.0122	0.0010	0.0022	0.0043	0.0028	0.0003	0.0004	0.0042	0.0047	0.0037	0.0036	0.0003	0.0003

Tabelle 301: Simulation PLSDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0898	0.0734	0.0131	0.0093	0.0846	0.0641	0.0116	0.0063	0.2081	0.1972	0.0869	0.0632	0.0124	0.0063
STD		0.0200	0.0128	0.0057	0.0062	0.0184	0.0050	0.0044	0.0014	0.0142	0.0119	0.0176	0.0107	0.0044	0.0020

Tabelle 302: Simulation PLSDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0862	0.0673	0.0119	0.0074	0.0841	0.0641	0.0115	0.0068	0.1807	0.1721	0.0830	0.0669	0.0103	0.0068
STD		0.0206	0.0163	0.0042	0.0041	0.0226	0.0146	0.0045	0.0033	0.0097	0.0091	0.0149	0.0095	0.0034	0.0016

Tabelle 303: Simulation PLSDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0813	0.0663	0.0101	0.0075	0.0809	0.0557	0.0100	0.0055	0.1539	0.1442	0.0788	0.0583	0.0095	0.0055
STD		0.0096	0.0164	0.0026	0.0033	0.0100	0.0121	0.0027	0.0020	0.0084	0.0088	0.0110	0.0096	0.0024	0.0018

Tabelle 304: Simulation PLSDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0872	0.0692	0.0112	0.0094	0.0898	0.0460	0.0119	0.0048	0.1311	0.1181	0.0861	0.0468	0.0110	0.0038
STD		0.0154	0.0169	0.0037	0.0048	0.0198	0.0106	0.0046	0.0016	0.0073	0.0087	0.0150	0.0111	0.0030	0.0018

Tabelle 305: Simulation PLSDA mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1271	0.0866	0.0224	0.0166	0.1296	0.0371	0.0227	0.0051	0.0993	0.0780	0.1152	0.0345	0.0177	0.0025
STD		0.0203	0.0296	0.0067	0.0123	0.0229	0.0094	0.0067	0.0030	0.0056	0.0086	0.0169	0.0091	0.0044	0.0012

Tabelle 306: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0528	0.0427	0.0047	0.0032	0.0472	0.0386	0.0037	0.0025	0.1945	0.1884	0.0525	0.0409	0.0045	0.0026
STD		0.0147	0.0110	0.0023	0.0021	0.0151	0.0081	0.0022	0.0012	0.0129	0.0109	0.0145	0.0074	0.0024	0.0010

Tabelle 307: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0545	0.0377	0.0048	0.0024	0.0515	0.0365	0.0044	0.0023	0.1833	0.1767	0.0532	0.0386	0.0045	0.0024
STD		0.0140	0.0104	0.0025	0.0011	0.0167	0.0098	0.0026	0.0010	0.0126	0.0115	0.0149	0.0065	0.0021	0.0008

Tabelle 308: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0641	0.0467	0.0063	0.0040	0.0656	0.0402	0.0065	0.0031	0.1520	0.1417	0.0702	0.0366	0.0072	0.0023
STD		0.0183	0.0110	0.0033	0.0022	0.0197	0.0096	0.0035	0.0018	0.0111	0.0122	0.0151	0.0072	0.0027	0.0009

Tabelle 309 Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0865	0.0477	0.0111	0.0040	0.0904	0.0341	0.0117	0.0025	0.1295	0.1145	0.0898	0.0332	0.0108	0.0018
STD		0.0134	0.0093	0.0035	0.0014	0.0139	0.0088	0.0035	0.0008	0.0097	0.0120	0.0120	0.0062	0.0028	0.0006

Tabelle 310: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1113	0.0552	0.0172	0.0060	0.1158	0.0278	0.0180	0.0023	0.1038	0.0820	0.1085	0.0273	0.0163	0.0013
STD		0.0121	0.0188	0.0032	0.0041	0.0124	0.0093	0.0034	0.0013	0.0078	0.0109	0.0100	0.0065	0.0029	0.0006

Tabelle 311: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0514	0.0412	0.0050	0.0030	0.0473	0.0370	0.0044	0.0024	0.1957	0.1929	0.0547	0.0437	0.0050	0.0030
STD		0.0117	0.0151	0.0027	0.0019	0.0121	0.0131	0.0025	0.0014	0.0104	0.0099	0.0115	0.0107	0.0023	0.0014

Tabelle 312: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0544	0.0428	0.0047	0.0029	0.0546	0.0414	0.0049	0.0027	0.1712	0.1667	0.0532	0.0410	0.0048	0.0027
STD		0.0126	0.0108	0.0021	0.0016	0.0148	0.0095	0.0025	0.0013	0.0098	0.0104	0.0115	0.0078	0.0025	0.0011

Tabelle 313: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0632	0.0472	0.0061	0.0043	0.0644	0.0392	0.0064	0.0031	0.1488	0.1409	0.0644	0.0361	0.0065	0.0022
STD		0.0158	0.0137	0.0034	0.0031	0.0180	0.0076	0.0039	0.0016	0.0089	0.0104	0.0125	0.0068	0.0027	0.0009

Tabelle 314 Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0903	0.0539	0.0118	0.0055	0.0936	0.0340	0.0123	0.0029	0.1202	0.1053	0.0916	0.0296	0.0114	0.0016
STD		0.0155	0.0183	0.0033	0.0037	0.0168	0.0107	0.0036	0.0017	0.0075	0.0098	0.0139	0.0068	0.0032	0.0008

Tabelle 315: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1312	0.0755	0.0222	0.0114	0.1335	0.0283	0.0227	0.0030	0.0911	0.0668	0.1207	0.0224	0.0194	0.0010
STD		0.0179	0.0160	0.0059	0.0048	0.0173	0.0059	0.0057	0.0010	0.0059	0.0086	0.0130	0.0058	0.0040	0.0005

Tabelle 316: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0518	0.0417	0.0051	0.0031	0.0477	0.0377	0.0045	0.0026	0.1963	0.1935	0.0549	0.0437	0.0050	0.0030
STD		0.0127	0.0157	0.0028	0.0021	0.0123	0.0136	0.0025	0.0015	0.0104	0.0099	0.0116	0.0106	0.0023	0.0014

Tabelle 317: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0529	0.0426	0.0044	0.0029	0.0524	0.0414	0.0045	0.0027	0.1742	0.1700	0.0520	0.0412	0.0047	0.0028
STD		0.0123	0.0110	0.0023	0.0016	0.0149	0.0095	0.0027	0.0013	0.0099	0.0104	0.0113	0.0079	0.0025	0.0011

Tabelle 318: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.9). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0648	0.0455	0.0065	0.0037	0.0654	0.0370	0.0066	0.0027	0.1443	0.1355	0.0682	0.0351	0.0071	0.0020
STD		0.0150	0.0091	0.0029	0.0015	0.0156	0.0048	0.0031	0.0008	0.0087	0.0103	0.0125	0.0057	0.0027	0.0008

Tabelle 319 Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.1). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.0942	0.0565	0.0124	0.0064	0.0978	0.0335	0.0131	0.0030	0.1167	0.1008	0.0948	0.0295	0.0122	0.0016
STD		0.0171	0.0222	0.0037	0.0050	0.0177	0.0108	0.0038	0.0020	0.0073	0.0097	0.0140	0.0069	0.0033	0.0008

Tabelle 320: Simulation PLSDA mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1264	0.0611	0.0210	0.0077	0.1289	0.0260	0.0216	0.0022	0.0949	0.0718	0.1171	0.0241	0.0183	0.0011
STD		0.0191	0.0132	0.0052	0.0036	0.0191	0.0050	0.0054	0.0009	0.0061	0.0088	0.0136	0.0055	0.0040	0.0005

Tabelle 321: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0444	0.0334	0.0031	0.0018	0.0407	0.0314	0.0026	0.0016	0.1952	0.1929	0.0395	0.0322	0.0025	0.0016
STD		0.0117	0.0105	0.0014	0.0011	0.0118	0.0092	0.0013	0.0009	0.0060	0.0053	0.0102	0.0066	0.0011	0.0007

Tabelle 322: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0431	0.0279	0.0028	0.0014	0.0443	0.0276	0.0029	0.0013	0.1738	0.1700	0.0466	0.0315	0.0032	0.0016
STD		0.0110	0.0072	0.0012	0.0006	0.0119	0.0068	0.0013	0.0006	0.0057	0.0055	0.0124	0.0044	0.0014	0.0005

Tabelle 323: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0586	0.0323	0.0052	0.0021	0.0619	0.0266	0.0056	0.0015	0.1444	0.1363	0.0610	0.0264	0.0053	0.0012
STD		0.0110	0.0089	0.0017	0.0013	0.0123	0.0067	0.0018	0.0008	0.0051	0.0053	0.0105	0.0051	0.0016	0.0004

Tabelle 324: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.0908	0.0390	0.0105	0.0031	0.0929	0.0240	0.0109	0.0015	0.1157	0.1012	0.0886	0.0217	0.0101	0.0009
STD		0.0097	0.0110	0.0021	0.0016	0.0098	0.0063	0.0021	0.0008	0.0043	0.0048	0.0079	0.0048	0.0018	0.0005

Tabelle 325: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1085	0.0446	0.0151	0.0035	0.1110	0.0208	0.0156	0.0013	0.0991	0.0797	0.1052	0.0178	0.0141	0.0006
STD		0.0093	0.0116	0.0022	0.0019	0.0091	0.0051	0.0021	0.0006	0.0038	0.0044	0.0058	0.0036	0.0017	0.0003

Tabelle 326: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0420	0.0321	0.0029	0.0015	0.0391	0.0308	0.0024	0.0014	0.1953	0.1932	0.0379	0.0344	0.0022	0.0017
STD		0.0117	0.0055	0.0014	0.0005	0.0112	0.0062	0.0011	0.0006	0.0063	0.0055	0.0085	0.0050	0.0009	0.0005

Tabelle 327: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0444	0.0314	0.0030	0.0017	0.0451	0.0306	0.0031	0.0017	0.1709	0.1670	0.0451	0.0322	0.0032	0.0017
STD		0.0117	0.0098	0.0013	0.0011	0.0112	0.0098	0.0013	0.0010	0.0061	0.0059	0.0112	0.0055	0.0015	0.0006

Tabelle 328: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0565	0.0341	0.0046	0.0020	0.0579	0.0300	0.0049	0.0016	0.1484	0.1412	0.0596	0.0277	0.0051	0.0013
STD		0.0106	0.0092	0.0017	0.0013	0.0127	0.0068	0.0020	0.0009	0.0057	0.0061	0.0099	0.0046	0.0016	0.0005

Tabelle 329: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.7). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0769	0.0341	0.0079	0.0021	0.0796	0.0245	0.0083	0.0013	0.1284	0.1170	0.0782	0.0233	0.0081	0.0009
STD		0.0087	0.0082	0.0017	0.0009	0.0105	0.0061	0.0019	0.0006	0.0053	0.0061	0.0086	0.0032	0.0018	0.0002

Tabelle 330: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1095	0.0424	0.0154	0.0033	0.1121	0.0216	0.0159	0.0013	0.1021	0.0833	0.1039	0.0196	0.0141	0.0007
STD		0.0086	0.0122	0.0023	0.0018	0.0088	0.0053	0.0023	0.0007	0.0046	0.0057	0.0062	0.0043	0.0017	0.0003

Tabelle 331: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.53).

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0417	0.0312	0.0029	0.0015	0.0387	0.0300	0.0024	0.0014	0.1964	0.1944	0.0381	0.0343	0.0022	0.0017
STD		0.0106	0.0065	0.0015	0.0006	0.0101	0.0063	0.0011	0.0006	0.0063	0.0055	0.0086	0.0053	0.0009	0.0005

Tabelle 332: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.64). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0426	0.0302	0.0028	0.0015	0.0426	0.0298	0.0028	0.0015	0.1763	0.1729	0.0426	0.0328	0.0029	0.0017
STD		0.0119	0.0091	0.0013	0.0007	0.0118	0.0093	0.0013	0.0007	0.0061	0.0058	0.0106	0.0048	0.0014	0.0005

Tabelle 333: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0558	0.0346	0.0045	0.0020	0.0574	0.0305	0.0047	0.0016	0.1490	0.1420	0.0591	0.0279	0.0050	0.0013
STD		0.0110	0.0098	0.0018	0.0012	0.0129	0.0074	0.0021	0.0009	0.0057	0.0061	0.0101	0.0049	0.0016	0.0005

Tabelle 334: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0872	0.0345	0.0100	0.0024	0.0901	0.0224	0.0104	0.0013	0.1201	0.1066	0.0862	0.0218	0.0097	0.0008
STD		0.0085	0.0091	0.0019	0.0012	0.0091	0.0059	0.0020	0.0006	0.0051	0.0060	0.0074	0.0039	0.0017	0.0003

Tabelle 335: Simulation PLSDA mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.15). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1094	0.0417	0.0154	0.0033	0.1121	0.0215	0.0159	0.0013	0.1023	0.0836	0.1036	0.0197	0.0140	0.0007
STD		0.0086	0.0133	0.0022	0.0018	0.0088	0.0060	0.0023	0.0007	0.0046	0.0057	0.0061	0.0043	0.0017	0.0003

Tabelle 336: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0316	0.0199	0.0016	0.0006	0.0296	0.0200	0.0014	0.0006	0.1924	0.1910	0.0296	0.0244	0.0014	0.0010
STD		0.0079	0.0048	0.0007	0.0003	0.0073	0.0050	0.0006	0.0003	0.0039	0.0042	0.0060	0.0050	0.0005	0.0003

Tabelle 337: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0373	0.0210	0.0022	0.0007	0.0387	0.0210	0.0023	0.0007	0.1718	0.1683	0.0393	0.0232	0.0024	0.0009
STD		0.0084	0.0066	0.0008	0.0004	0.0076	0.0072	0.0009	0.0004	0.0038	0.0045	0.0056	0.0055	0.0007	0.0004

Tabelle 338: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0584	0.0237	0.0048	0.0009	0.0613	0.0191	0.0051	0.0007	0.1431	0.1346	0.0634	0.0187	0.0054	0.0006
STD		0.0073	0.0060	0.0012	0.0005	0.0072	0.0049	0.0012	0.0003	0.0033	0.0047	0.0072	0.0040	0.0012	0.0003

Tabelle 339: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0795	0.0199	0.0083	0.0007	0.0829	0.0141	0.0088	0.0004	0.1262	0.1136	0.0809	0.0149	0.0084	0.0004
STD		0.0093	0.0047	0.0018	0.0003	0.0099	0.0035	0.0019	0.0002	0.0030	0.0046	0.0088	0.0025	0.0017	0.0001

Tabelle 340: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1167	0.0365	0.0169	0.0024	0.1201	0.0160	0.0177	0.0009	0.0984	0.0772	0.1103	0.0134	0.0157	0.0004
STD		0.0112	0.0088	0.0025	0.0010	0.0109	0.0045	0.0025	0.0004	0.0024	0.0042	0.0084	0.0033	0.0020	0.0002

Tabelle 341: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0353	0.0210	0.0019	0.0007	0.0318	0.0203	0.0016	0.0007	0.1905	0.1887	0.0319	0.0218	0.0016	0.0008
STD		0.0043	0.0046	0.0005	0.0003	0.0074	0.0057	0.0006	0.0003	0.0040	0.0045	0.0065	0.0048	0.0007	0.0003

Tabelle 342: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0429	0.0193	0.0027	0.0006	0.0444	0.0189	0.0028	0.0006	0.1671	0.1628	0.0460	0.0196	0.0029	0.0006
STD		0.0079	0.0055	0.0010	0.0003	0.0100	0.0055	0.0013	0.0003	0.0041	0.0052	0.0096	0.0029	0.0012	0.0002

Tabelle 343: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0615	0.0223	0.0053	0.0008	0.0648	0.0196	0.0057	0.0007	0.1453	0.1373	0.0649	0.0191	0.0057	0.0006
STD		0.0082	0.0061	0.0014	0.0004	0.0097	0.0050	0.0017	0.0003	0.0040	0.0055	0.0090	0.0048	0.0015	0.0003

Tabelle 344: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0861	0.0230	0.0099	0.0010	0.0903	0.0161	0.0106	0.0006	0.1224	0.1088	0.0890	0.0177	0.0101	0.0006
STD		0.0067	0.0068	0.0015	0.0005	0.0076	0.0052	0.0017	0.0003	0.0036	0.0053	0.0078	0.0063	0.0017	0.0004

Tabelle 345: Simulation PLSDA mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1155	0.0285	0.0166	0.0016	0.1194	0.0147	0.0174	0.0006	0.1003	0.0801	0.1111	0.0142	0.0156	0.0004
STD		0.0077	0.0094	0.0021	0.0009	0.0083	0.0054	0.0023	0.0004	0.0031	0.0049	0.0068	0.0042	0.0019	0.0002

Tabelle 346: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0350	0.0208	0.0021	0.0007	0.0287	0.0202	0.0014	0.0007	0.1969	0.1955	0.0303	0.0230	0.0014	0.0009
STD		0.0055	0.0053	0.0007	0.0004	0.0060	0.0072	0.0005	0.0004	0.0039	0.0042	0.0058	0.0043	0.0005	0.0003

Tabelle 347: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0404	0.0187	0.0023	0.0006	0.0414	0.0185	0.0024	0.0006	0.1705	0.1667	0.0432	0.0196	0.0026	0.0006
STD		0.0061	0.0044	0.0008	0.0003	0.0079	0.0043	0.0010	0.0002	0.0041	0.0051	0.0094	0.0032	0.0011	0.0002

Tabelle 348: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0615	0.0218	0.0052	0.0008	0.0647	0.0193	0.0057	0.0007	0.1460	0.1381	0.0643	0.0191	0.0056	0.0006
STD		0.0078	0.0069	0.0014	0.0005	0.0094	0.0052	0.0017	0.0003	0.0040	0.0055	0.0090	0.0047	0.0015	0.0003

Tabelle 349: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0846	0.0221	0.0095	0.0010	0.0888	0.0160	0.0101	0.0006	0.1244	0.1113	0.0870	0.0179	0.0096	0.0006
STD		0.0067	0.0074	0.0015	0.0005	0.0080	0.0061	0.0017	0.0003	0.0037	0.0054	0.0078	0.0062	0.0017	0.0004

Tabelle 350: Simulation PLSDA mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1211	0.0295	0.0183	0.0017	0.1251	0.0136	0.0191	0.0006	0.0955	0.0737	0.1155	0.0129	0.0169	0.0003
STD		0.0084	0.0060	0.0021	0.0007	0.0081	0.0043	0.0022	0.0003	0.0030	0.0047	0.0065	0.0038	0.0019	0.0002

Tabelle 351: Simulation RFR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	/	0.0751	/	0.0092	/	0.0730	/	0.0084	0.2135	0.1940	0.1359	0.0748	0.0295	0.0090
STD		/	0.0205	/	0.0048	/	0.0168	/	0.0039	0.0066	0.0106	0.0205	0.0149	0.0092	0.0033

Tabelle 352: Simulation RFR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	/	0.0669	/	0.0077	/	0.0610	/	0.0063	0.1920	0.1700	0.1421	0.0648	0.0306	0.0066
STD		/	0.0125	/	0.0039	/	0.0116	/	0.0020	0.0068	0.0107	0.0199	0.0079	0.0079	0.0015

Tabelle 353: Simulation RFR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	/	0.078	/	0.0113	/	0.0607	/	0.0074	0.1658	0.1357	0.1690	0.0567	0.0391	0.0057
STD		/	0.023	/	0.0068	/	0.0130	/	0.0033	0.0059	0.0116	0.0226	0.0132	0.0083	0.0025

Tabelle 354: Simulation RFR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.47). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1643	0.0831	0.0343	0.0138	0.1850	0.0534	0.0420	0.0070	0.1464	0.1077	0.1936	0.0536	0.0470	0.0055
STD		#DIV/0!	0.0242	#DIV/0!	0.0072	#DIV/0!	0.0147	#DIV/0!	0.0031	0.0054	0.0131	0.0207	0.0124	0.0084	0.0026

Tabelle 355: Simulation RFR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.53). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1767	0.0940	0.0399	0.0165	0.2054	0.0434	0.0506	0.0058	0.1260	0.0806	0.2108	0.0371	0.0522	0.0030
STD		0.0167	0.0226	0.0066	0.0069	0.0268	0.0146	0.0100	0.0030	0.0049	0.0150	0.0257	0.0106	0.0107	0.0019

Tabelle 356: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.19). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.68	#DIV/0!	0.0503	#DIV/0!	0.0041	#DIV/0!	0.0440	#DIV/0!	0.0032	0.2513	0.2069	0.1910	0.0502	0.0475	0.0042
STD		#DIV/0!	0.0112	#DIV/0!	0.0015	#DIV/0!	0.0140	#DIV/0!	0.0018	0.0094	0.0090	0.0157	0.0114	0.0083	0.0019

Tabelle 357: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	#DIV/0!	0.0361	#DIV/0!	0.0021	#DIV/0!	0.0351	#DIV/0!	0.0020	0.2187	0.1761	0.1687	0.0428	0.0443	0.0029
STD		#DIV/0!	0.0104	#DIV/0!	0.0010	#DIV/0!	0.0103	#DIV/0!	0.0010	0.0060	0.0116	0.0145	0.0066	0.0094	0.0009

Tabelle 358: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.78	/	0.0389	/	0.0027	/	0.0351	/	0.0022	0.2014	0.1557	0.1787	0.0382	0.0473	0.0023
STD		/	0.0131	/	0.0016	/	0.0112	/	0.0013	0.0052	0.0112	0.0157	0.0087	0.0081	0.0009

Tabelle 359 Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	/	0.0553	/	0.0061	/	0.0356	#DIV/0!	0.0031	0.1658	0.1127	0.2057	0.0312	0.0560	0.0016
STD		/	0.0158	/	0.0033	/	0.0120	#DIV/0!	0.0014	0.0047	0.0117	0.0212	0.0063	0.0099	0.0007

Tabelle 360: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.1705	0.0666	0.0400	0.0086	0.1973	0.0328	0.0514	0.0033	0.1437	0.0844	0.2274	0.0321	0.0636	0.0020
STD		/	0.0197	/	0.0051	/	0.0096	/	0.0017	0.0046	0.0114	0.0184	0.0055	0.0097	0.0008

Tabelle 361: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.41). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	/	0.0593	/	0.0080	/	0.0499	/	0.0045	0.2046	0.1985	0.0918	0.0516	0.0123	0.0043
STD		/	0.0259	/	0.0082	/	0.0207	/	0.0034	0.0091	0.0078	0.0173	0.0109	0.0058	0.0016

Tabelle 362: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.51). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0726	0.0406	0.0068	0.0028	0.0717	0.0379	0.0067	0.0024	0.1771	0.1728	0.0760	0.0419	0.0088	0.0029
STD		0.0139	0.0111	0.0026	0.0014	0.0140	0.0101	0.0027	0.0012	0.0094	0.0104	0.0143	0.0105	0.0025	0.0013

Tabelle 363: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.61). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0844	0.0401	0.0101	0.0026	0.0835	0.0365	0.0100	0.0022	0.1528	0.1465	0.0845	0.0389	0.0104	0.0024
STD		0.0158	0.0057	0.0036	0.0009	0.0180	0.0042	0.0035	0.0005	0.0088	0.0108	0.0177	0.0069	0.0033	0.0011

Tabelle 364 Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.75). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1000	0.0507	0.0125	0.0052	0.1016	0.0333	0.0128	0.0027	0.1236	0.1135	0.1061	0.0309	0.0136	0.0017
STD		0.0124	0.0167	0.0032	0.0047	0.0135	0.0103	0.0034	0.0022	0.0075	0.0101	0.0146	0.0071	0.0037	0.0007

Tabelle 365: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=0.92). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1172	0.0722	0.0177	0.0101	0.1147	0.0280	0.0168	0.0030	0.0941	0.0787	0.1230	0.0243	0.0185	0.0010
STD		0.0142	0.0167	0.0034	0.0038	0.0128	0.0082	0.0031	0.0013	0.0066	0.0097	0.0140	0.0033	0.0036	0.0003

Tabelle 366: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0518	0.0417	0.0051	0.0031	0.0477	0.0377	0.0045	0.0026	0.1963	0.1935	0.0549	0.0437	0.0050	0.0030
STD		0.0127	0.0157	0.0028	0.0021	0.0123	0.0136	0.0025	0.0015	0.0104	0.0099	0.0116	0.0106	0.0023	0.0014

Tabelle 367: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0604	0.0498	0.0062	0.0042	0.0569	0.0446	0.0050	0.0034	0.1765	0.1763	0.0575	0.0477	0.0051	0.0036
STD		0.0145	0.0148	0.0042	0.0019	0.0095	0.0147	0.0014	0.0017	0.0099	0.0100	0.0067	0.0114	0.0011	0.0015

Tabelle 368: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.85). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0583	0.0553	0.0051	0.0054	0.0571	0.0402	0.0050	0.0033	0.1369	0.1353	0.0618	0.0394	0.0054	0.0026
STD		0.0132	0.0178	0.0020	0.0035	0.0146	0.0108	0.0021	0.0018	0.0082	0.0093	0.0157	0.0061	0.0023	0.0009

Tabelle 369 Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0723	0.0626	0.0073	0.0073	0.0680	0.0369	0.0065	0.0034	0.1131	0.1095	0.0728	0.0343	0.0072	0.0020
STD		0.0147	0.0203	0.0030	0.0039	0.0134	0.0124	0.0025	0.0018	0.0076	0.0096	0.0138	0.0085	0.0027	0.0010

Tabelle 370: Simulation RFR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.0979	0.0712	0.0134	0.0108	0.0778	0.0242	0.0094	0.0028	0.0736	0.0667	0.0825	0.0240	0.0098	0.0011
STD		0.0154	0.0227	0.0045	0.0060	0.0110	0.0075	0.0029	0.0016	0.0064	0.0087	0.0128	0.0061	0.0031	0.0006

Tabelle 371: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.19). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.68	/	0.0431	/	0.0040	/	0.0326	/	0.0017	0.2412	0.2046	0.1668	0.0365	0.0380	0.0021
STD		/	0.0177	/	0.0043	/	0.0068	/	0.0008	0.0068	0.0067	0.0169	0.0075	0.0069	0.0008

Tabelle 372: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.24). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	/	0.0282	/	0.0015	/	0.0281	/	0.0014	0.2157	0.1790	0.1574	0.0320	0.0373	0.0018
STD		/	0.0123	/	0.0011	/	0.0121	/	0.0011	0.0064	0.0064	0.0165	0.0081	0.0067	0.0010

Tabelle 373: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.78	/	0.0327	/	0.0018	/	0.0306	/	0.0016	0.1959	0.1535	0.1721	0.0321	0.0433	0.0017
STD		/	0.0092	/	0.0008	/	0.0093	/	0.0008	0.0045	0.0061	0.0134	0.0071	0.0049	0.0007

Tabelle 374: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.36). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.83	/	0.0346	/	0.0021	/	0.0249	/	0.0013	0.1669	0.1185	0.1971	0.0242	0.0503	0.0010
STD		/	0.0105	/	0.0010	/	0.0090	/	0.0008	0.0041	0.0053	0.0116	0.0072	0.0048	0.0006

Tabelle 375: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1825	0.0439	0.0415	0.0039	0.2133	0.0213	0.0533	0.0015	0.1357	0.0802	0.2236	0.0197	0.0587	0.0007
STD		0.0092	0.0092	0.0034	0.0015	0.0102	0.0054	0.0041	0.0007	0.0042	0.0056	0.0091	0.0051	0.0043	0.0004

Tabelle 376: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	/	0.0359	/	0.0024	/	0.0277	/	0.0013	0.2005	0.1914	0.0886	0.0313	0.0112	0.0015
STD		/	0.0088	/	0.0014	/	0.0068	/	0.0005	0.0041	0.0049	0.0082	0.0069	0.0019	0.0006

Tabelle 377: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0759	0.0326	0.0078	0.0018	0.0768	0.0300	0.0081	0.0016	0.1726	0.1655	0.0804	0.0302	0.0091	0.0014
STD		0.0117	0.0068	0.0022	0.0009	0.0133	0.0054	0.0024	0.0007	0.0042	0.0060	0.0130	0.0034	0.0026	0.0003

Tabelle 378: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0828	0.0330	0.0088	0.0022	0.0860	0.0265	0.0094	0.0016	0.1482	0.1400	0.0885	0.0267	0.0100	0.0012
STD		0.0124	0.0080	0.0025	0.0012	0.0141	0.0055	0.0028	0.0007	0.0042	0.0062	0.0126	0.0025	0.0026	0.0003

Tabelle 379: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0935	0.0383	0.0112	0.0029	0.0956	0.0247	0.0116	0.0014	0.1224	0.1118	0.1017	0.0233	0.0124	0.0009
STD		0.0091	0.0104	0.0019	0.0017	0.0101	0.0058	0.0021	0.0007	0.0036	0.0053	0.0098	0.0052	0.0022	0.0004

Tabelle 380: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1076	0.0441	0.0146	0.0038	0.1050	0.0197	0.0139	0.0013	0.0964	0.0831	0.1121	0.0175	0.0151	0.0006
STD		0.0079	0.0079	0.0018	0.0015	0.0075	0.0035	0.0017	0.0005	0.0033	0.0045	0.0074	0.0033	0.0017	0.0002

Tabelle 381: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.57). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0626	0.0344	0.0050	0.0022	0.0598	0.0271	0.0047	0.0013	0.1874	0.1832	0.0644	0.0307	0.0062	0.0015
STD		0.0056	0.0092	0.0008	0.0012	0.0043	0.0058	0.0007	0.0006	0.0036	0.0044	0.0089	0.0046	0.0017	0.0004

Tabelle 382: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0570	0.0341	0.0051	0.0020	0.0561	0.0295	0.0046	0.0016	0.1701	0.1674	0.0581	0.0312	0.0050	0.0016
STD		0.0124	0.0081	0.0025	0.0009	0.0096	0.0072	0.0012	0.0007	0.0041	0.0050	0.0077	0.0029	0.0013	0.0003

Tabelle 383: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0533	0.0336	0.0040	0.0019	0.0539	0.0277	0.0041	0.0014	0.1415	0.1389	0.0581	0.0268	0.0045	0.0012
STD		0.0086	0.0081	0.0012	0.0009	0.0090	0.0064	0.0012	0.0006	0.0042	0.0054	0.0090	0.0054	0.0013	0.0006

Tabelle 384: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0642	0.0400	0.0055	0.0033	0.0604	0.0243	0.0049	0.0016	0.1140	0.1106	0.0631	0.0244	0.0051	0.0009
STD		0.0079	0.0111	0.0011	0.0018	0.0086	0.0069	0.0012	0.0007	0.0041	0.0048	0.0089	0.0043	0.0012	0.0003

Tabelle 385: Simulation RFR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0807	0.0526	0.0092	0.0056	0.0655	0.0192	0.0065	0.0016	0.0787	0.0734	0.0691	0.0205	0.0066	0.0007
STD		0.0089	0.0149	0.0021	0.0030	0.0060	0.0057	0.0011	0.0008	0.0037	0.0042	0.0063	0.0039	0.0010	0.0003

Tabelle 386: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.68	/	0.0184	/	0.0006	/	0.0164	/	0.0004	0.2363	0.2035	0.1507	0.0207	0.0318	0.0007
STD		/	0.0058	/	0.0003	/	0.0068	/	0.0003	0.0053	0.0032	0.0134	0.0046	0.0050	0.0003

Tabelle 387: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.23). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	/	0.0191	/	0.0006	/	0.0192	/	0.0006	0.2100	0.1768	0.1459	0.0197	0.0319	0.0006
STD		/	0.0041	/	0.0003	/	0.0040	/	0.0003	0.0055	0.0037	0.0097	0.0032	0.0047	0.0002

Tabelle 388: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	/	0.0206	/	0.0008	/	0.0180	/	0.0007	0.1839	0.1449	0.1703	0.0181	0.0386	0.0005
STD		/	0.0044	/	0.0003	/	0.0040	/	0.0003	0.0029	0.0038	0.0093	0.0033	0.0045	0.0002

Tabelle 389: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.83	/	0.0231	/	0.0009	/	0.0163	/	0.0006	0.1643	0.1200	0.1906	0.0147	0.0450	0.0004
STD		/	0.0047	/	0.0003	/	0.0041	/	0.0002	0.0028	0.0034	0.0052	0.0026	0.0031	0.0001

Tabelle 390: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.88	0.1764	0.0271	0.0389	0.0013	0.2056	0.0140	0.0499	0.0005	0.1373	0.0843	0.2210	0.0124	0.0556	0.0003
STD		0.0031	0.0064	0.0016	0.0006	0.0028	0.0035	0.0015	0.0002	0.0027	0.0031	0.0057	0.0033	0.0029	0.0002

Tabelle 391: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0623	0.0207	0.0057	0.0008	0.0571	0.0192	0.0053	0.0006	0.1954	0.1908	0.0617	0.0236	0.0058	0.0009
STD		0.0150	0.0071	0.0025	0.0004	0.0144	0.0057	0.0023	0.0003	0.0038	0.0047	0.0109	0.0031	0.0019	0.0002

Tabelle 392: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.51). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0625	0.0206	0.0053	0.0008	0.0631	0.0192	0.0055	0.0007	0.1671	0.1624	0.0656	0.0205	0.0060	0.0007
STD		0.0096	0.0049	0.0016	0.0003	0.0089	0.0045	0.0016	0.0003	0.0032	0.0049	0.0096	0.0028	0.0017	0.0002

Tabelle 393: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.61). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0725	0.0214	0.0066	0.0008	0.0756	0.0171	0.0071	0.0006	0.1433	0.1369	0.0798	0.0171	0.0077	0.0005
STD		0.0072	0.0062	0.0014	0.0005	0.0077	0.0047	0.0015	0.0003	0.0033	0.0049	0.0084	0.0030	0.0016	0.0002

Tabelle 394: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0889	0.0230	0.0095	0.0010	0.0900	0.0149	0.0096	0.0006	0.1201	0.1110	0.0956	0.0166	0.0105	0.0004
STD		0.0057	0.0081	0.0012	0.0008	0.0056	0.0038	0.0012	0.0004	0.0033	0.0048	0.0071	0.0025	0.0015	0.0001

Tabelle 395: Simulation RFR mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.86). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1045	0.0342	0.0134	0.0022	0.0990	0.0156	0.0122	0.0008	0.0938	0.0822	0.1059	0.0142	0.0131	0.0004
STD		0.0061	0.0122	0.0017	0.0016	0.0054	0.0041	0.0014	0.0006	0.0027	0.0042	0.0054	0.0029	0.0015	0.0001

Tabelle 396: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0450	0.0405	0.0046	0.0107	0.0368	0.0214	0.0021	0.0009	0.1938	0.1926	0.0397	0.0245	0.0026	0.0010
STD		0.0118	0.0286	0.0052	0.0255	0.0038	0.0067	0.0004	0.0006	0.0037	0.0047	0.0081	0.0057	0.0012	0.0005

Tabelle 397: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.66). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0379	0.0238	0.0021	0.0010	0.0392	0.0212	0.0022	0.0008	0.1668	0.1659	0.0403	0.0241	0.0024	0.0009
STD		0.0096	0.0076	0.0010	0.0008	0.0104	0.0061	0.0011	0.0005	0.0038	0.0050	0.0105	0.0030	0.0012	0.0003

Tabelle 398: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0450	0.0282	0.0028	0.0014	0.0454	0.0230	0.0028	0.0009	0.1408	0.1392	0.0475	0.0229	0.0030	0.0008
STD		0.0081	0.0054	0.0009	0.0005	0.0081	0.0036	0.0010	0.0003	0.0039	0.0050	0.0084	0.0043	0.0010	0.0004

Tabelle 399: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0572	0.0307	0.0045	0.0019	0.0534	0.0197	0.0040	0.0010	0.1148	0.1121	0.0562	0.0210	0.0041	0.0007
STD		0.0081	0.0066	0.0012	0.0009	0.0067	0.0041	0.0010	0.0004	0.0037	0.0046	0.0065	0.0038	0.0009	0.0002

Tabelle 400: Simulation RFR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.0772	0.0373	0.0076	0.0025	0.0599	0.0142	0.0052	0.0007	0.0788	0.0743	0.0634	0.0143	0.0054	0.0004
STD		0.0091	0.0092	0.0014	0.0012	0.0058	0.0030	0.0009	0.0003	0.0032	0.0043	0.0057	0.0031	0.0010	0.0002

Tabelle 401: Simulation SVR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.24). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0897	0.0694	0.0128	0.0099	0.0908	0.0592	0.0116	0.0064	0.1992	0.1976	0.0975	0.0638	0.0137	0.0068
STD		0.0248	0.0238	0.0098	0.0082	0.0199	0.0188	0.0055	0.0040	0.0120	0.0108	0.0160	0.0162	0.0045	0.0035

Tabelle 402: Simulation SVR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.1074	0.0566	0.0179	0.0053	0.1164	0.0538	0.0183	0.0050	0.1793	0.1685	0.1274	0.0592	0.0227	0.0058
STD		0.0354	0.0144	0.0119	0.0027	0.0329	0.0138	0.0084	0.0027	0.0107	0.0109	0.0245	0.0129	0.0069	0.0025

Tabelle 403: Simulation SVR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1279	0.0730	0.0227	0.0114	0.1453	0.0567	0.0270	0.0060	0.1618	0.1393	0.1563	0.0541	0.0316	0.0048
STD		0.0163	0.0172	0.0059	0.0115	0.0233	0.0083	0.0075	0.0013	0.0094	0.0112	0.0277	0.0084	0.0094	0.0015

Tabelle 404: Simulation SVR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.39). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1496	0.0772	0.0298	0.0125	0.1777	0.0477	0.0384	0.0059	0.1463	0.1118	0.1894	0.0494	0.0428	0.0042
STD		0.0173	0.0236	0.0067	0.0069	0.0278	0.0086	0.0100	0.0027	0.0082	0.0113	0.0251	0.0101	0.0100	0.0020

Tabelle 405: Simulation SVR mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.48). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1793	0.1053	0.0419	0.0233	0.2182	0.0401	0.0554	0.0059	0.1227	0.0701	0.2322	0.0338	0.0609	0.0023
STD		0.0129	0.0265	0.0062	0.0116	0.0216	0.0087	0.0098	0.0018	0.0064	0.0104	0.0239	0.0074	0.0111	0.0011

Tabelle 406: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0787	0.0418	0.0105	0.0033	0.0788	0.0381	0.0098	0.0026	0.1997	0.1928	0.0852	0.0433	0.0111	0.0031
STD		0.0164	0.0145	0.0049	0.0026	0.0195	0.0141	0.0048	0.0016	0.0046	0.0083	0.0162	0.0103	0.0052	0.0013

Tabelle 407: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.25). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0970	0.0462	0.0137	0.0033	0.0997	0.0462	0.0146	0.0033	0.1881	0.1753	0.1108	0.0479	0.0173	0.0037
STD		0.0140	0.0109	0.0048	0.0013	0.0151	0.0109	0.0052	0.0012	0.0048	0.0089	0.0137	0.0084	0.0055	0.0012

Tabelle 408: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.1302	0.0427	0.0221	0.0032	0.1445	0.0379	0.0256	0.0027	0.1707	0.1457	0.1510	0.0416	0.0287	0.0028
STD		0.0109	0.0095	0.0046	0.0013	0.0136	0.0103	0.0054	0.0012	0.0052	0.0094	0.0138	0.0108	0.0061	0.0016

Tabelle 409 Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1542	0.0501	0.0300	0.0048	0.1782	0.0348	0.0372	0.0029	0.1552	0.1178	0.1870	0.0363	0.0412	0.0023
STD		0.0074	0.0171	0.0038	0.0031	0.0124	0.0115	0.0052	0.0017	0.0056	0.0096	0.0141	0.0125	0.0066	0.0015

Tabelle 410: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1739	0.0589	0.0386	0.0066	0.2097	0.0288	0.0508	0.0024	0.1359	0.0840	0.2217	0.0277	0.0558	0.0013
STD		0.0089	0.0209	0.0048	0.0043	0.0162	0.0098	0.0075	0.0014	0.0060	0.0097	0.0157	0.0048	0.0075	0.0004

Tabelle 411: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0718	0.0578	0.0160	0.0060	0.0624	0.0438	0.0062	0.0032	0.2045	0.2020	0.0670	0.0461	0.0069	0.0033
STD		0.0318	0.0189	0.0312	0.0033	0.0173	0.0167	0.0032	0.0023	0.0087	0.0102	0.0147	0.0105	0.0028	0.0014

Tabelle 412: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0849	0.0468	0.0122	0.0040	0.0841	0.0439	0.0106	0.0034	0.1851	0.1784	0.0894	0.0454	0.0116	0.0034
STD		0.0167	0.0119	0.0063	0.0020	0.0150	0.0113	0.0031	0.0015	0.0102	0.0117	0.0148	0.0105	0.0034	0.0016

Tabelle 413: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0983	0.0459	0.0132	0.0035	0.1075	0.0404	0.0153	0.0028	0.1643	0.1511	0.1149	0.0420	0.0176	0.0027
STD		0.0132	0.0103	0.0029	0.0014	0.0132	0.0102	0.0032	0.0013	0.0110	0.0126	0.0144	0.0073	0.0036	0.0010

Tabelle 414 Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1292	0.0566	0.0215	0.0061	0.1479	0.0343	0.0264	0.0030	0.1366	0.1120	0.1547	0.0341	0.0289	0.0019
STD		0.0134	0.0088	0.0042	0.0020	0.0138	0.0048	0.0041	0.0009	0.0102	0.0123	0.0124	0.0051	0.0040	0.0006

Tabelle 415: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1602	0.0878	0.0322	0.0163	0.1847	0.0281	0.0392	0.0035	0.1068	0.0691	0.1923	0.0250	0.0419	0.0011
STD		0.0131	0.0253	0.0056	0.0087	0.0144	0.0062	0.0058	0.0014	0.0083	0.0109	0.0151	0.0051	0.0058	0.0004

Tabelle 416: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0786	0.0578	0.0181	0.0059	0.0620	0.0431	0.0064	0.0031	0.2077	0.2058	0.0637	0.0479	0.0064	0.0037
STD		0.0328	0.0158	0.0310	0.0029	0.0156	0.0169	0.0024	0.0019	0.0094	0.0104	0.0153	0.0112	0.0026	0.0018

Tabelle 417: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0758	0.0541	0.0095	0.0061	0.0741	0.0470	0.0086	0.0041	0.1876	0.1827	0.0806	0.0487	0.0100	0.0039
STD		0.0165	0.0156	0.0047	0.0047	0.0180	0.0166	0.0039	0.0020	0.0110	0.0125	0.0199	0.0168	0.0046	0.0026

Tabelle 418: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.9). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1003	0.0494	0.0140	0.0046	0.1069	0.0398	0.0157	0.0030	0.1593	0.1464	0.1140	0.0402	0.0172	0.0026
STD		0.0123	0.0114	0.0032	0.0020	0.0159	0.0078	0.0040	0.0011	0.0115	0.0139	0.0150	0.0076	0.0043	0.0011

Tabelle 419 Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.1). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1253	0.0586	0.0204	0.0072	0.1433	0.0329	0.0249	0.0031	0.1304	0.1080	0.1481	0.0345	0.0266	0.0020
STD		0.0146	0.0150	0.0054	0.0042	0.0157	0.0049	0.0060	0.0014	0.0101	0.0133	0.0160	0.0049	0.0058	0.0006

Tabelle 420: Simulation SVR mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1420	0.0655	0.0265	0.0098	0.1652	0.0254	0.0323	0.0024	0.1063	0.0754	0.1746	0.0264	0.0350	0.0012
STD		0.0131	0.0192	0.0057	0.0061	0.0137	0.0049	0.0061	0.0012	0.0092	0.0120	0.0177	0.0050	0.0070	0.0005

Tabelle 421: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0740	0.0309	0.0080	0.0016	0.0663	0.0274	0.0070	0.0012	0.2015	0.1953	0.0708	0.0310	0.0077	0.0015
STD		0.0131	0.0092	0.0026	0.0010	0.0145	0.0069	0.0026	0.0006	0.0034	0.0052	0.0128	0.0039	0.0024	0.0003

Tabelle 422: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.1063	0.0321	0.0153	0.0018	0.1092	0.0315	0.0159	0.0017	0.1824	0.1668	0.1149	0.0309	0.0174	0.0016
STD		0.0120	0.0091	0.0030	0.0009	0.0133	0.0092	0.0033	0.0009	0.0041	0.0060	0.0114	0.0059	0.0027	0.0005

Tabelle 423: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.31). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.1309	0.0333	0.0219	0.0018	0.1405	0.0295	0.0245	0.0015	0.1682	0.1438	0.1450	0.0288	0.0263	0.0013
STD		0.0078	0.0094	0.0023	0.0009	0.0111	0.0090	0.0028	0.0007	0.0048	0.0063	0.0107	0.0059	0.0033	0.0005

Tabelle 424: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1561	0.0348	0.0301	0.0024	0.1792	0.0253	0.0373	0.0015	0.1493	0.1113	0.1857	0.0239	0.0405	0.0010
STD		0.0076	0.0103	0.0029	0.0012	0.0121	0.0079	0.0040	0.0007	0.0054	0.0060	0.0121	0.0043	0.0049	0.0003

Tabelle 425: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.43). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1740	0.0369	0.0372	0.0030	0.2082	0.0183	0.0487	0.0011	0.1327	0.0829	0.2179	0.0190	0.0530	0.0007
STD		0.0085	0.0089	0.0036	0.0014	0.0151	0.0044	0.0062	0.0005	0.0058	0.0054	0.0144	0.0055	0.0065	0.0003

Tabelle 426: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0596	0.0396	0.0057	0.0035	0.0567	0.0290	0.0048	0.0015	0.2046	0.2015	0.0619	0.0355	0.0057	0.0019
STD		0.0130	0.0141	0.0023	0.0034	0.0147	0.0049	0.0021	0.0005	0.0043	0.0045	0.0112	0.0056	0.0020	0.0006

Tabelle 427: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.57). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0894	0.0317	0.0109	0.0015	0.0936	0.0303	0.0113	0.0014	0.1789	0.1691	0.0975	0.0309	0.0121	0.0016
STD		0.0127	0.0077	0.0038	0.0007	0.0093	0.0070	0.0023	0.0006	0.0044	0.0053	0.0096	0.0042	0.0024	0.0005

Tabelle 428: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1070	0.0331	0.0141	0.0023	0.1178	0.0279	0.0165	0.0017	0.1618	0.1465	0.1232	0.0279	0.0177	0.0014
STD		0.0105	0.0083	0.0024	0.0012	0.0101	0.0074	0.0027	0.0008	0.0039	0.0049	0.0119	0.0055	0.0030	0.0006

Tabelle 429: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.77). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1314	0.0351	0.0207	0.0025	0.1513	0.0230	0.0257	0.0013	0.1410	0.1158	0.1559	0.0234	0.0276	0.0010
STD		0.0065	0.0145	0.0017	0.0021	0.0057	0.0081	0.0014	0.0010	0.0052	0.0056	0.0058	0.0063	0.0020	0.0006

Tabelle 430: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.9). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1437	0.0437	0.0256	0.0039	0.1712	0.0183	0.0330	0.0013	0.1192	0.0864	0.1782	0.0197	0.0353	0.0007
STD		0.0081	0.0129	0.0022	0.0019	0.0069	0.0052	0.0023	0.0006	0.0059	0.0054	0.0069	0.0037	0.0025	0.0004

Tabelle 431: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.62). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0569	0.0396	0.0047	0.0036	0.0542	0.0291	0.0046	0.0015	0.2020	0.1992	0.0587	0.0351	0.0053	0.0019
STD		0.0125	0.0114	0.0018	0.0038	0.0122	0.0056	0.0017	0.0004	0.0036	0.0042	0.0078	0.0055	0.0012	0.0006

Tabelle 432: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0680	0.0285	0.0068	0.0014	0.0725	0.0260	0.0076	0.0012	0.1857	0.1793	0.0790	0.0311	0.0087	0.0016
STD		0.0121	0.0115	0.0020	0.0011	0.0106	0.0103	0.0018	0.0009	0.0039	0.0048	0.0100	0.0056	0.0019	0.0006

Tabelle 433: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.85). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0960	0.0334	0.0114	0.0021	0.1092	0.0296	0.0139	0.0017	0.1650	0.1525	0.1118	0.0299	0.0149	0.0015
STD		0.0095	0.0097	0.0018	0.0012	0.0106	0.0082	0.0024	0.0009	0.0045	0.0055	0.0111	0.0060	0.0026	0.0006

Tabelle 434: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=1.05). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1228	0.0299	0.0186	0.0020	0.1436	0.0201	0.0234	0.0010	0.1362	0.1135	0.1490	0.0230	0.0252	0.0009
STD		0.0074	0.0115	0.0016	0.0013	0.0071	0.0073	0.0022	0.0007	0.0057	0.0057	0.0087	0.0049	0.0029	0.0004

Tabelle 435: Simulation SVR mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1344	0.0440	0.0221	0.0044	0.1587	0.0188	0.0283	0.0014	0.1158	0.0882	0.1646	0.0198	0.0302	0.0007
STD		0.0088	0.0147	0.0020	0.0027	0.0070	0.0051	0.0023	0.0007	0.0058	0.0055	0.0069	0.0036	0.0026	0.0003

Tabelle 436: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0855	0.0223	0.0101	0.0008	0.0812	0.0218	0.0088	0.0008	0.2003	0.1921	0.0806	0.0242	0.0091	0.0009
STD		0.0163	0.0053	0.0045	0.0003	0.0147	0.0057	0.0031	0.0003	0.0028	0.0021	0.0117	0.0048	0.0027	0.0003

Tabelle 437: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.28). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.1163	0.0201	0.0168	0.0007	0.1212	0.0195	0.0180	0.0006	0.1816	0.1638	0.1218	0.0197	0.0186	0.0007
STD		0.0110	0.0071	0.0034	0.0004	0.0120	0.0068	0.0036	0.0004	0.0035	0.0015	0.0083	0.0056	0.0032	0.0004

Tabelle 438: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1351	0.0237	0.0227	0.0010	0.1478	0.0203	0.0261	0.0008	0.1674	0.1411	0.1510	0.0193	0.0276	0.0006
STD		0.0080	0.0045	0.0030	0.0004	0.0084	0.0038	0.0037	0.0003	0.0037	0.0013	0.0084	0.0026	0.0037	0.0002

Tabelle 439: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.37). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.1545	0.0267	0.0296	0.0013	0.1772	0.0179	0.0367	0.0008	0.1514	0.1139	0.1842	0.0169	0.0394	0.0005
STD		0.0044	0.0057	0.0023	0.0005	0.0049	0.0032	0.0034	0.0003	0.0035	0.0020	0.0070	0.0044	0.0036	0.0003

Tabelle 440: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1792	0.0325	0.0383	0.0019	0.2154	0.0150	0.0506	0.0007	0.1283	0.0761	0.2238	0.0128	0.0551	0.0003
STD		0.0056	0.0069	0.0023	0.0009	0.0088	0.0033	0.0038	0.0004	0.0032	0.0028	0.0091	0.0038	0.0044	0.0002

Tabelle 441: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0777	0.0293	0.0103	0.0016	0.0679	0.0236	0.0060	0.0009	0.1982	0.1932	0.0702	0.0258	0.0066	0.0010
STD		0.0146	0.0068	0.0068	0.0009	0.0078	0.0044	0.0012	0.0003	0.0029	0.0042	0.0086	0.0041	0.0016	0.0004

Tabelle 442: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0944	0.0230	0.0117	0.0008	0.0996	0.0222	0.0124	0.0008	0.1781	0.1664	0.1031	0.0226	0.0133	0.0008
STD		0.0107	0.0045	0.0030	0.0003	0.0095	0.0038	0.0021	0.0003	0.0032	0.0042	0.0092	0.0037	0.0024	0.0002

Tabelle 443: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.7). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.1097	0.0252	0.0156	0.0011	0.1297	0.0218	0.0198	0.0008	0.1584	0.1389	0.1352	0.0202	0.0213	0.0006
STD		0.0099	0.0066	0.0027	0.0005	0.0119	0.0067	0.0033	0.0004	0.0032	0.0042	0.0115	0.0046	0.0036	0.0003

Tabelle 444: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.82). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.1306	0.0293	0.0219	0.0016	0.1588	0.0182	0.0287	0.0008	0.1365	0.1078	0.1663	0.0180	0.0309	0.0005
STD		0.0093	0.0056	0.0035	0.0007	0.0132	0.0031	0.0046	0.0003	0.0038	0.0040	0.0122	0.0030	0.0049	0.0002

Tabelle 445: Simulation SVR mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1466	0.0364	0.0272	0.0029	0.1786	0.0143	0.0361	0.0008	0.1133	0.0768	0.1873	0.0131	0.0389	0.0003
STD		0.0088	0.0077	0.0032	0.0018	0.0114	0.0020	0.0046	0.0004	0.0041	0.0037	0.0111	0.0017	0.0046	0.0001

Tabelle 446: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.62). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.69	0.0616	0.0384	0.0072	0.0050	0.0504	0.0241	0.0037	0.0011	0.2059	0.2032	0.0525	0.0263	0.0042	0.0011
STD		0.0138	0.0146	0.0060	0.0069	0.0083	0.0073	0.0010	0.0006	0.0029	0.0035	0.0076	0.0042	0.0012	0.0003

Tabelle 447: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.75). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.74	0.0791	0.0288	0.0090	0.0014	0.0814	0.0250	0.0084	0.0010	0.1867	0.1789	0.0863	0.0248	0.0095	0.0010
STD		0.0161	0.0080	0.0048	0.0006	0.0098	0.0074	0.0020	0.0005	0.0031	0.0040	0.0106	0.0050	0.0022	0.0004

Tabelle 448: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.92). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.1037	0.0215	0.0137	0.0008	0.1202	0.0198	0.0172	0.0006	0.1605	0.1437	0.1254	0.0194	0.0186	0.0006
STD		0.0089	0.0055	0.0022	0.0004	0.0100	0.0055	0.0028	0.0003	0.0031	0.0039	0.0102	0.0039	0.0032	0.0002

Tabelle 449: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=1.08). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.1208	0.0293	0.0190	0.0016	0.1479	0.0185	0.0254	0.0008	0.1370	0.1119	0.1551	0.0189	0.0273	0.0006
STD		0.0099	0.0022	0.0031	0.0005	0.0137	0.0028	0.0045	0.0002	0.0042	0.0038	0.0130	0.0036	0.0046	0.0002

Tabelle 450: Simulation SVR mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1374	0.0350	0.0244	0.0023	0.1699	0.0133	0.0327	0.0006	0.1073	0.0748	0.1773	0.0127	0.0348	0.0003
STD		0.0085	0.0065	0.0026	0.0011	0.0098	0.0033	0.0035	0.0003	0.0041	0.0036	0.0093	0.0017	0.0034	0.0001

Tabelle 451: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.71	0.0800	0.0721	0.0102	0.0097	0.0807	0.0643	0.0104	0.0071	0.1895	0.1900	0.0858	0.0730	0.0115	0.0080
STD_{Ridge}		0.0138	0.0129	0.0032	0.0050	0.0199	0.0125	0.0052	0.0028	0.0107	0.0099	0.0224	0.0147	0.0062	0.0033
$MW_{Elastic\ Net}$	0.70	0.0783	0.0742	0.0104	0.0104	0.0798	0.0653	0.0106	0.0070	0.1968	0.1978	0.0820	0.0710	0.0107	0.0077
$STD_{Elastic\ Net}$		0.0121	0.0155	0.0034	0.0083	0.0141	0.0080	0.0035	0.0019	0.0125	0.0111	0.0118	0.0117	0.0028	0.0024
MW_{Lasso}	0.70	0.0776	0.0773	0.0099	0.0090	0.0791	0.0729	0.0102	0.0075	0.1973	0.1982	0.0823	0.0715	0.0107	0.0078
STD_{Lasso}		0.0087	0.0163	0.0029	0.0045	0.0109	0.0108	0.0031	0.0019	0.0128	0.0113	0.0114	0.0105	0.0025	0.0024
MW_{Lasso+}	0.70	0.1019	0.0711	0.0166	0.0092	0.0872	0.0637	0.0114	0.0063	0.1996	0.1983	0.0942	0.0719	0.0132	0.0079
STD_{Lasso+}		0.0206	0.0259	0.0081	0.0088	0.0125	0.0154	0.0029	0.0028	0.0085	0.0111	0.0161	0.0113	0.0037	0.0025

Tabelle 452: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.76	0.0821	0.0669	0.0101	0.0078	0.0853	0.0618	0.0108	0.0066	0.1627	0.1608	0.0916	0.0655	0.0130	0.0069
STD_{Ridge}		0.0137	0.0171	0.0036	0.0040	0.0176	0.0135	0.0049	0.0028	0.0098	0.0096	0.0213	0.0121	0.0061	0.0024
$MW_{Elastic\ Net}$	0.75	0.0757	0.0647	0.0092	0.0071	0.0754	0.0603	0.0093	0.0063	0.1670	0.1684	0.0786	0.0649	0.0095	0.0067
$STD_{Elastic\ Net}$		0.0166	0.0142	0.0037	0.0027	0.0188	0.0108	0.0044	0.0021	0.0105	0.0097	0.0189	0.0086	0.0043	0.0017
MW_{Lasso}	0.75	0.0742	0.0659	0.0089	0.0070	0.0745	0.0621	0.0091	0.0063	0.1671	0.1687	0.0769	0.0647	0.0091	0.0065
STD_{Lasso}		0.0177	0.0150	0.0041	0.0027	0.0203	0.0114	0.0046	0.0021	0.0106	0.0097	0.0174	0.0081	0.0039	0.0016
MW_{Lasso+}	0.75	0.1181	0.0625	0.0204	0.0066	0.1126	0.0594	0.0189	0.0060	0.1794	0.1686	0.1218	0.0645	0.0210	0.0065
STD_{Lasso+}		0.0251	0.0106	0.0072	0.0029	0.0232	0.0102	0.0058	0.0021	0.0063	0.0095	0.0231	0.0101	0.0066	0.0018

Tabelle 453: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.82	0.0975	0.0653	0.0143	0.0077	0.0968	0.0519	0.0141	0.0052	0.1337	0.1271	0.1052	0.0507	0.0153	0.0043
STD_{Ridge}		0.0065	0.0177	0.0020	0.0036	0.0076	0.0146	0.0022	0.0023	0.0083	0.0094	0.0132	0.0097	0.0034	0.0017
$MW_{Elastic\ Net}$	0.81	0.0822	0.0739	0.0108	0.0093	0.0742	0.0561	0.0093	0.0060	0.1360	0.1349	0.0764	0.0548	0.0094	0.0049

STD$_{Elastic\ Net}$		0.0113	0.0126	0.0029	0.0030	0.0105	0.0096	0.0027	0.0018	0.0083	0.0087	0.0103	0.0070	0.0022	0.0010
MW$_{Lasso}$	0.81	0.0820	0.0704	0.0110	0.0083	0.0731	0.0535	0.0093	0.0055	0.1359	0.1351	0.0742	0.0542	0.0090	0.0048
STD$_{Lasso}$		0.0104	0.0122	0.0029	0.0025	0.0101	0.0100	0.0027	0.0017	0.0084	0.0088	0.0098	0.0080	0.0022	0.0010
MW$_{Lasso\ +}$	0.81	0.1471	0.0684	0.0291	0.0081	0.1539	0.0533	0.0309	0.0055	0.1599	0.1351	0.1622	0.0533	0.0333	0.0047
STD$_{Lasso\ +}$		0.0181	0.0107	0.0071	0.0022	0.0179	0.0086	0.0069	0.0015	0.0055	0.0087	0.0207	0.0080	0.0068	0.0010

Tabelle 454: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.39). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.86	0.1151	0.0840	0.0192	0.0135	0.1090	0.0469	0.0182	0.0057	0.1134	0.1018	0.1189	0.0449	0.0199	0.0036
STD$_{Ridge}$		0.0156	0.0194	0.0043	0.0051	0.0153	0.0095	0.0043	0.0021	0.0072	0.0094	0.0212	0.0084	0.0056	0.0014
MW$_{Elastic\ Net}$	0.85	0.0893	0.0822	0.0122	0.0129	0.0788	0.0476	0.0105	0.0060	0.1144	0.1097	0.0890	0.0464	0.0121	0.0037
STD$_{Elastic\ Net}$		0.0138	0.0137	0.0032	0.0028	0.0139	0.0125	0.0031	0.0019	0.0072	0.0085	0.0184	0.0103	0.0036	0.0017
MW$_{Lasso}$	0.85	0.0892	0.0823	0.0124	0.0132	0.0780	0.0476	0.0106	0.0062	0.1141	0.1098	0.0869	0.0465	0.0118	0.0037
STD$_{Lasso}$		0.0169	0.0208	0.0038	0.0059	0.0159	0.0143	0.0036	0.0029	0.0072	0.0085	0.0187	0.0113	0.0037	0.0017
MW$_{Lasso\ +}$	0.85	0.1596	0.0823	0.0338	0.0132	0.1818	0.0504	0.0405	0.0066	0.1463	0.1098	0.1948	0.0449	0.0452	0.0036
STD$_{Lasso\ +}$		0.0182	0.0173	0.0069	0.0043	0.0211	0.0121	0.0082	0.0024	0.0055	0.0086	0.0233	0.0108	0.0092	0.0017

Tabelle 455: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.48). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.91	0.1421	0.1039	0.0290	0.0220	0.1245	0.0366	0.0249	0.0054	0.0845	0.0645	0.1363	0.0324	0.0262	0.0023
STD$_{Ridge}$		0.0189	0.0329	0.0071	0.0162	0.0173	0.0087	0.0055	0.0029	0.0055	0.0088	0.0179	0.0065	0.0053	0.0009
MW$_{Elastic\ Net}$	0.90	0.1297	0.1112	0.0237	0.0228	0.1044	0.0418	0.0184	0.0061	0.0843	0.0716	0.1107	0.0318	0.0185	0.0022
STD$_{Elastic\ Net}$		0.0192	0.0259	0.0067	0.0096	0.0168	0.0088	0.0054	0.0022	0.0054	0.0082	0.0182	0.0091	0.0051	0.0012
MW$_{Lasso}$	0.90	0.1291	0.1052	0.0240	0.0225	0.1031	0.0394	0.0185	0.0055	0.0840	0.0717	0.1090	0.0318	0.0181	0.0022
STD$_{Lasso}$		0.0195	0.0282	0.0068	0.0100	0.0161	0.0110	0.0051	0.0025	0.0055	0.0081	0.0181	0.0090	0.0050	0.0012
MW$_{Lasso\ +}$	0.90	0.1775	0.1002	0.0423	0.0201	0.2141	0.0390	0.0549	0.0052	0.1263	0.0717	0.2344	0.0320	0.0627	0.0022
STD$_{Lasso\ +}$		0.0122	0.0225	0.0062	0.0093	0.0186	0.0097	0.0091	0.0020	0.0058	0.0082	0.0237	0.0099	0.0107	0.0014

Tabelle 456: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.19). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.71	0.0630	0.0385	0.0065	0.0027	0.0576	0.0340	0.0051	0.0020	0.1945	0.1922	0.0583	0.0417	0.0051	0.0028
STD_{Ridge}		0.0176	0.0112	0.0036	0.0014	0.0198	0.0103	0.0029	0.0011	0.0097	0.0109	0.0182	0.0115	0.0030	0.0016
$MW_{Elastic\ Net}$	0.70	0.0552	0.0337	0.0051	0.0023	0.0491	0.0284	0.0041	0.0015	0.1974	0.1957	0.0499	0.0410	0.0039	0.0027
$STD_{Elastic\ Net}$		0.0139	0.0088	0.0023	0.0019	0.0131	0.0055	0.0020	0.0006	0.0116	0.0109	0.0143	0.0082	0.0022	0.0011
MW_{Lasso}	0.70	0.0555	0.0354	0.0051	0.0026	0.0494	0.0296	0.0042	0.0017	0.1975	0.1958	0.0497	0.0411	0.0039	0.0027
STD_{Lasso}		0.0139	0.0098	0.0023	0.0020	0.0133	0.0082	0.0021	0.0008	0.0117	0.0110	0.0142	0.0080	0.0022	0.0010
MW_{Lasso+}	0.70	0.0820	0.0371	0.0102	0.0023	0.0731	0.0323	0.0080	0.0018	0.2023	0.1957	0.0803	0.0413	0.0097	0.0027
STD_{Lasso+}		0.0143	0.0101	0.0032	0.0011	0.0152	0.0080	0.0032	0.0008	0.0070	0.0110	0.0211	0.0084	0.0044	0.0011

Tabelle 457: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.75	0.0653	0.0397	0.0066	0.0027	0.0668	0.0380	0.0068	0.0025	0.1726	0.1686	0.0666	0.0381	0.0067	0.0023
STD_{Ridge}		0.0173	0.0085	0.0032	0.0010	0.0226	0.0073	0.0040	0.0009	0.0099	0.0120	0.0211	0.0081	0.0037	0.0009
$MW_{Elastic\ Net}$	0.75	0.0543	0.0378	0.0049	0.0022	0.0536	0.0372	0.0047	0.0021	0.1745	0.1725	0.0528	0.0383	0.0044	0.0024
$STD_{Elastic\ Net}$		0.0171	0.0084	0.0028	0.0007	0.0185	0.0078	0.0028	0.0007	0.0115	0.0119	0.0168	0.0098	0.0024	0.0011
MW_{Lasso}	0.75	0.0551	0.0375	0.0048	0.0023	0.0543	0.0364	0.0047	0.0022	0.1743	0.1723	0.0521	0.0383	0.0044	0.0024
STD_{Lasso}		0.0161	0.0097	0.0026	0.0011	0.0175	0.0092	0.0027	0.0009	0.0116	0.0119	0.0164	0.0097	0.0023	0.0011
MW_{Lasso+}	0.75	0.1073	0.0376	0.0161	0.0022	0.1061	0.0371	0.0161	0.0022	0.1874	0.1723	0.1144	0.0386	0.0182	0.0024
STD_{Lasso+}		0.0187	0.0104	0.0051	0.0010	0.0220	0.0103	0.0058	0.0010	0.0060	0.0119	0.0214	0.0097	0.0059	0.0011

Tabelle 458: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.81	0.0804	0.0446	0.0097	0.0035	0.0799	0.0348	0.0096	0.0024	0.1417	0.1336	0.0902	0.0340	0.0112	0.0019
STD_{Ridge}		0.0147	0.0156	0.0036	0.0022	0.0158	0.0107	0.0037	0.0013	0.0094	0.0124	0.0199	0.0054	0.0042	0.0007
$MW_{Elastic\ Net}$	0.80	0.0677	0.0466	0.0073	0.0041	0.0650	0.0377	0.0069	0.0029	0.1416	0.1367	0.0706	0.0342	0.0075	0.0020

STD$_{Elastic\ Net}$		0.0172	0.0081	0.0033	0.0013	0.0160	0.0067	0.0030	0.0009	0.0105	0.0123	0.0171	0.0061	0.0030	0.0008
MW$_{Lasso}$	0.80	0.0680	0.0459	0.0073	0.0038	0.0651	0.0374	0.0069	0.0027	0.1413	0.1366	0.0697	0.0341	0.0072	0.0020
STD$_{Lasso}$		0.0165	0.0090	0.0032	0.0013	0.0154	0.0082	0.0029	0.0010	0.0106	0.0123	0.0165	0.0065	0.0029	0.0008
MW$_{Lasso\ +}$	0.80	0.1418	0.0477	0.0259	0.0041	0.1582	0.0392	0.0307	0.0031	0.1679	0.1366	0.1690	0.0355	0.0346	0.0021
STD$_{Lasso\ +}$		0.0161	0.0100	0.0059	0.0016	0.0221	0.0073	0.0081	0.0011	0.0054	0.0122	0.0240	0.0067	0.0084	0.0009

Tabelle 459: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.34). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.85	0.1002	0.0526	0.0149	0.0052	0.0959	0.0336	0.0144	0.0028	0.1193	0.1069	0.1075	0.0325	0.0160	0.0019
STD$_{Ridge}$		0.0134	0.0141	0.0037	0.0030	0.0143	0.0120	0.0040	0.0017	0.0085	0.0120	0.0176	0.0080	0.0044	0.0009
MW$_{Elastic\ Net}$	0.85	0.0879	0.0516	0.0119	0.0047	0.0794	0.0335	0.0106	0.0026	0.1184	0.1097	0.0888	0.0328	0.0115	0.0018
STD$_{Elastic\ Net}$		0.0147	0.0134	0.0039	0.0021	0.0128	0.0097	0.0035	0.0012	0.0095	0.0118	0.0138	0.0060	0.0033	0.0006
MW$_{Lasso}$	0.85	0.0875	0.0470	0.0117	0.0042	0.0784	0.0309	0.0104	0.0023	0.1180	0.1096	0.0874	0.0328	0.0112	0.0018
STD$_{Lasso}$		0.0137	0.0116	0.0037	0.0020	0.0121	0.0090	0.0033	0.0011	0.0095	0.0118	0.0135	0.0060	0.0032	0.0006
MW$_{Lasso\ +}$	0.85	0.1645	0.0492	0.0344	0.0047	0.1923	0.0333	0.0436	0.0027	0.1543	0.1097	0.2047	0.0335	0.0485	0.0019
STD$_{Lasso\ +}$		0.0126	0.0112	0.0054	0.0021	0.0209	0.0100	0.0083	0.0013	0.0050	0.0118	0.0212	0.0062	0.0090	0.0006

Tabelle 460: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.41). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.90	0.1251	0.0661	0.0219	0.0095	0.1126	0.0277	0.0195	0.0028	0.0937	0.0754	0.1248	0.0243	0.0213	0.0011
STD$_{Ridge}$		0.0134	0.0205	0.0042	0.0053	0.0137	0.0084	0.0041	0.0013	0.0075	0.0109	0.0128	0.0079	0.0042	0.0007
MW$_{Elastic\ Net}$	0.89	0.1141	0.0639	0.0188	0.0090	0.0964	0.0276	0.0156	0.0028	0.0922	0.0782	0.1049	0.0258	0.0166	0.0012
STD$_{Elastic\ Net}$		0.0138	0.0281	0.0039	0.0092	0.0126	0.0081	0.0032	0.0019	0.0080	0.0106	0.0120	0.0066	0.0034	0.0006
MW$_{Lasso}$	0.89	0.1139	0.0605	0.0186	0.0084	0.0956	0.0265	0.0153	0.0026	0.0918	0.0781	0.1035	0.0256	0.0163	0.0012
STD$_{Lasso}$		0.0130	0.0270	0.0039	0.0092	0.0118	0.0079	0.0032	0.0019	0.0081	0.0106	0.0120	0.0067	0.0033	0.0006
MW$_{Lasso\ +}$	0.89	0.1841	0.0642	0.0439	0.0084	0.2247	0.0290	0.0593	0.0028	0.1381	0.0782	0.2394	0.0260	0.0644	0.0012
STD$_{Lasso\ +}$		0.0095	0.0267	0.0052	0.0085	0.0165	0.0078	0.0086	0.0019	0.0047	0.0107	0.0177	0.0067	0.0093	0.0006

Tabelle 461: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.41). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.70	0.0584	0.0388	0.0053	0.0026	0.0563	0.0369	0.0050	0.0024	0.1936	0.1927	0.0621	0.0432	0.0059	0.0032
STD_{Ridge}		0.0098	0.0121	0.0018	0.0013	0.0097	0.0131	0.0017	0.0014	0.0072	0.0084	0.0127	0.0096	0.0024	0.0013
$MW_{Elastic\ Net}$	0.69	0.0584	0.0461	0.0053	0.0039	0.0582	0.0408	0.0052	0.0027	0.1981	0.1983	0.0588	0.0490	0.0054	0.0038
$STD_{Elastic\ Net}$		0.0146	0.0118	0.0025	0.0024	0.0122	0.0129	0.0023	0.0013	0.0082	0.0083	0.0116	0.0126	0.0025	0.0017
MW_{Lasso}	0.70	0.0580	0.0470	0.0053	0.0037	0.0582	0.0412	0.0052	0.0026	0.1984	0.1986	0.0577	0.0492	0.0052	0.0039
STD_{Lasso}		0.0136	0.0110	0.0025	0.0018	0.0125	0.0121	0.0023	0.0012	0.0083	0.0083	0.0116	0.0119	0.0025	0.0016
MW_{Lasso+}	0.69	0.0766	0.0452	0.0088	0.0034	0.0721	0.0408	0.0079	0.0026	0.2030	0.1986	0.0810	0.0492	0.0099	0.0039
STD_{Lasso+}		0.0143	0.0110	0.0035	0.0017	0.0162	0.0133	0.0030	0.0013	0.0068	0.0083	0.0166	0.0118	0.0043	0.0016

Tabelle 462: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.51). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.76	0.0657	0.0399	0.0068	0.0029	0.0666	0.0379	0.0068	0.0026	0.1702	0.1669	0.0721	0.0402	0.0075	0.0026
STD_{Ridge}		0.0137	0.0127	0.0026	0.0018	0.0144	0.0119	0.0029	0.0016	0.0075	0.0093	0.0152	0.0082	0.0028	0.0011
$MW_{Elastic\ Net}$	0.74	0.0545	0.0395	0.0046	0.0026	0.0560	0.0377	0.0049	0.0023	0.1745	0.1734	0.0608	0.0453	0.0060	0.0034
$STD_{Elastic\ Net}$		0.0119	0.0098	0.0022	0.0016	0.0139	0.0081	0.0028	0.0009	0.0085	0.0095	0.0133	0.0112	0.0028	0.0013
MW_{Lasso}	0.75	0.0524	0.0369	0.0045	0.0024	0.0543	0.0354	0.0048	0.0021	0.1747	0.1737	0.0599	0.0450	0.0059	0.0033
STD_{Lasso}		0.0154	0.0129	0.0026	0.0018	0.0177	0.0111	0.0032	0.0012	0.0086	0.0095	0.0131	0.0112	0.0028	0.0014
MW_{Lasso+}	0.75	0.1042	0.0354	0.0150	0.0021	0.1036	0.0343	0.0150	0.0020	0.1867	0.1737	0.1116	0.0453	0.0175	0.0034
STD_{Lasso+}		0.0181	0.0115	0.0046	0.0012	0.0215	0.0102	0.0052	0.0010	0.0065	0.0095	0.0212	0.0115	0.0059	0.0014

Tabelle 463: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.61). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.80	0.0810	0.0402	0.0096	0.0033	0.0811	0.0331	0.0094	0.0025	0.1476	0.1412	0.0881	0.0362	0.0104	0.0023
STD_{Ridge}		0.0137	0.0137	0.0029	0.0018	0.0140	0.0108	0.0030	0.0013	0.0074	0.0096	0.0171	0.0092	0.0034	0.0011
$MW_{Elastic\ Net}$	0.79	0.0646	0.0425	0.0065	0.0032	0.0629	0.0371	0.0062	0.0025	0.1514	0.1481	0.0693	0.0385	0.0072	0.0025

STD$_{Elastic\ Net}$		0.0143	0.0100	0.0028	0.0014	0.0150	0.0070	0.0029	0.0009	0.0086	0.0101	0.0162	0.0070	0.0031	0.0008
MW$_{Lasso}$	0.79	0.0635	0.0402	0.0063	0.0028	0.0616	0.0353	0.0061	0.0022	0.1515	0.1484	0.0680	0.0381	0.0071	0.0024
STD$_{Lasso}$		0.0143	0.0078	0.0027	0.0011	0.0148	0.0058	0.0028	0.0007	0.0087	0.0102	0.0160	0.0069	0.0031	0.0008
MW$_{Lasso\ +}$	0.79	0.1312	0.0408	0.0224	0.0028	0.1370	0.0364	0.0242	0.0023	0.1721	0.1484	0.1467	0.0385	0.0276	0.0025
STD$_{Lasso\ +}$		0.0180	0.0106	0.0057	0.0013	0.0217	0.0081	0.0068	0.0009	0.0063	0.0101	0.0208	0.0073	0.0071	0.0009

Tabelle 464: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.75). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.85	0.1010	0.0464	0.0147	0.0042	0.0964	0.0306	0.0138	0.0023	0.1195	0.1074	0.1096	0.0287	0.0158	0.0015
STD$_{Ridge}$		0.0139	0.0083	0.0032	0.0020	0.0153	0.0056	0.0033	0.0009	0.0069	0.0095	0.0175	0.0058	0.0045	0.0007
MW$_{Elastic\ Net}$	0.84	0.0874	0.0461	0.0115	0.0041	0.0804	0.0317	0.0104	0.0023	0.1223	0.1145	0.0883	0.0303	0.0112	0.0016
STD$_{Elastic\ Net}$		0.0127	0.0150	0.0033	0.0027	0.0127	0.0075	0.0031	0.0012	0.0081	0.0101	0.0152	0.0046	0.0034	0.0005
MW$_{Lasso}$	0.84	0.0869	0.0450	0.0114	0.0039	0.0795	0.0314	0.0102	0.0022	0.1223	0.1148	0.0868	0.0304	0.0110	0.0016
STD$_{Lasso}$		0.0134	0.0156	0.0034	0.0027	0.0134	0.0079	0.0033	0.0012	0.0081	0.0102	0.0155	0.0046	0.0034	0.0005
MW$_{Lasso\ +}$	0.84	0.1617	0.0451	0.0331	0.0040	0.1821	0.0315	0.0397	0.0023	0.1543	0.1147	0.1906	0.0307	0.0436	0.0016
STD$_{Lasso\ +}$		0.0127	0.0165	0.0053	0.0027	0.0180	0.0087	0.0075	0.0012	0.0063	0.0101	0.0210	0.0049	0.0088	0.0005

Tabelle 465: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.92). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.90	0.1293	0.0702	0.0235	0.0101	0.1177	0.0270	0.0210	0.0027	0.0920	0.0726	0.1290	0.0230	0.0227	0.0011
STD$_{Ridge}$		0.0158	0.0214	0.0049	0.0063	0.0147	0.0040	0.0044	0.0012	0.0060	0.0087	0.0156	0.0060	0.0048	0.0006
MW$_{Elastic\ Net}$	0.89	0.1178	0.0618	0.0199	0.0086	0.1019	0.0278	0.0170	0.0028	0.0938	0.0793	0.1118	0.0262	0.0178	0.0012
STD$_{Elastic\ Net}$		0.0176	0.0234	0.0053	0.0058	0.0169	0.0067	0.0048	0.0015	0.0068	0.0093	0.0168	0.0067	0.0047	0.0006
MW$_{Lasso}$	0.89	0.1168	0.0629	0.0194	0.0091	0.1006	0.0281	0.0165	0.0029	0.0937	0.0796	0.1102	0.0265	0.0174	0.0013
STD$_{Lasso}$		0.0168	0.0217	0.0047	0.0061	0.0167	0.0057	0.0043	0.0014	0.0070	0.0093	0.0169	0.0066	0.0046	0.0006
MW$_{Lasso\ +}$	0.89	0.1828	0.0625	0.0425	0.0089	0.2206	0.0284	0.0558	0.0028	0.1362	0.0796	0.2342	0.0263	0.0617	0.0012
STD$_{Lasso\ +}$		0.0137	0.0221	0.0059	0.0062	0.0211	0.0065	0.0091	0.0014	0.0063	0.0093	0.0201	0.0066	0.0095	0.0006

Tabelle 466: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.72	0.0577	0.0391	0.0053	0.0025	0.0561	0.0387	0.0050	0.0024	0.1865	0.1852	0.0596	0.0422	0.0055	0.0029
STD_{Ridge}		0.0110	0.0090	0.0021	0.0011	0.0112	0.0094	0.0019	0.0010	0.0073	0.0086	0.0139	0.0090	0.0024	0.0011
$MW_{Elastic\ Net}$	0.71	0.0513	0.0443	0.0042	0.0032	0.0514	0.0436	0.0043	0.0031	0.1899	0.1900	0.0559	0.0465	0.0051	0.0035
$STD_{Elastic\ Net}$		0.0141	0.0135	0.0020	0.0020	0.0179	0.0130	0.0025	0.0020	0.0080	0.0088	0.0167	0.0128	0.0030	0.0018
MW_{Lasso}	0.70	0.0580	0.0470	0.0053	0.0037	0.0582	0.0412	0.0052	0.0026	0.1984	0.1986	0.0577	0.0492	0.0052	0.0039
STD_{Lasso}		0.0136	0.0110	0.0025	0.0018	0.0125	0.0121	0.0023	0.0012	0.0083	0.0083	0.0116	0.0119	0.0025	0.0016
MW_{Lasso+}	0.71	0.0860	0.0449	0.0106	0.0033	0.0842	0.0451	0.0102	0.0032	0.1970	0.1903	0.0886	0.0470	0.0116	0.0036
STD_{Lasso+}		0.0208	0.0139	0.0041	0.0021	0.0222	0.0141	0.0039	0.0021	0.0054	0.0087	0.0220	0.0135	0.0046	0.0019

Tabelle 467: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.75	0.0584	0.0373	0.0054	0.0023	0.0587	0.0361	0.0053	0.0022	0.1742	0.1717	0.0658	0.0380	0.0065	0.0024
STD_{Ridge}		0.0121	0.0092	0.0022	0.0011	0.0130	0.0081	0.0024	0.0009	0.0074	0.0091	0.0154	0.0087	0.0028	0.0014
$MW_{Elastic\ Net}$	0.74	0.0540	0.0408	0.0045	0.0031	0.0555	0.0397	0.0048	0.0029	0.1774	0.1768	0.0589	0.0421	0.0056	0.0031
$STD_{Elastic\ Net}$		0.0119	0.0159	0.0019	0.0023	0.0149	0.0153	0.0026	0.0021	0.0082	0.0093	0.0167	0.0126	0.0033	0.0016
MW_{Lasso}	0.74	0.0533	0.0397	0.0044	0.0029	0.0546	0.0385	0.0047	0.0027	0.1776	0.1771	0.0578	0.0417	0.0054	0.0030
STD_{Lasso}		0.0103	0.0150	0.0017	0.0021	0.0139	0.0143	0.0025	0.0019	0.0083	0.0093	0.0158	0.0127	0.0031	0.0016
MW_{Lasso+}	0.74	0.0997	0.0399	0.0142	0.0029	0.0989	0.0396	0.0140	0.0028	0.1884	0.1770	0.1049	0.0418	0.0159	0.0030
STD_{Lasso+}		0.0247	0.0157	0.0052	0.0023	0.0262	0.0150	0.0051	0.0022	0.0052	0.0093	0.0241	0.0129	0.0058	0.0017

Tabelle 468: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.85). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.81	0.0824	0.0473	0.0102	0.0045	0.0803	0.0357	0.0098	0.0029	0.1406	0.1335	0.0910	0.0347	0.0111	0.0021
STD_{Ridge}		0.0134	0.0122	0.0030	0.0025	0.0136	0.0069	0.0031	0.0015	0.0073	0.0096	0.0176	0.0069	0.0036	0.0011
$MW_{Elastic\ Net}$	0.80	0.0718	0.0412	0.0078	0.0031	0.0702	0.0339	0.0076	0.0022	0.1433	0.1390	0.0743	0.0343	0.0079	0.0019

STD$_{Elastic\ Net}$		0.0137	0.0119	0.0027	0.0016	0.0131	0.0089	0.0028	0.0010	0.0080	0.0098	0.0175	0.0073	0.0035	0.0008
MW$_{Lasso}$	0.80	0.0710	0.0399	0.0076	0.0030	0.0691	0.0325	0.0074	0.0021	0.1434	0.1394	0.0728	0.0347	0.0077	0.0020
STD$_{Lasso}$		0.0143	0.0117	0.0027	0.0017	0.0135	0.0078	0.0027	0.0010	0.0081	0.0098	0.0168	0.0073	0.0033	0.0008
MW$_{Lasso\ +}$	0.80	0.1419	0.0399	0.0259	0.0029	0.1526	0.0329	0.0290	0.0021	0.1668	0.1393	0.1569	0.0350	0.0310	0.0020
STD$_{Lasso\ +}$		0.0179	0.0116	0.0063	0.0017	0.0228	0.0077	0.0078	0.0011	0.0051	0.0098	0.0246	0.0071	0.0082	0.0008

Tabelle 469: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW$_{Ridge}$	0.85	0.1024	0.0522	0.0148	0.0053	0.0970	0.0316	0.0139	0.0026	0.1185	0.1068	0.1078	0.0291	0.0153	0.0015
STD$_{Ridge}$		0.0135	0.0126	0.0034	0.0023	0.0144	0.0084	0.0034	0.0011	0.0068	0.0095	0.0171	0.0060	0.0043	0.0007
MW$_{Elastic\ Net}$	0.84	0.0886	0.0439	0.0117	0.0043	0.0809	0.0296	0.0105	0.0023	0.1206	0.1124	0.0911	0.0295	0.0116	0.0015
STD$_{Elastic\ Net}$		0.0147	0.0156	0.0034	0.0036	0.0158	0.0095	0.0034	0.0016	0.0075	0.0095	0.0181	0.0066	0.0041	0.0007
MW$_{Lasso}$	0.84	0.0877	0.0427	0.0115	0.0040	0.0801	0.0290	0.0103	0.0022	0.1206	0.1127	0.0898	0.0298	0.0114	0.0016
STD$_{Lasso}$		0.0135	0.0134	0.0031	0.0029	0.0144	0.0096	0.0031	0.0014	0.0076	0.0095	0.0173	0.0065	0.0039	0.0007
MW$_{Lasso\ +}$	0.84	0.1621	0.0433	0.0332	0.0040	0.1824	0.0297	0.0400	0.0022	0.1531	0.1127	0.1941	0.0304	0.0444	0.0016
STD$_{Lasso\ +}$		0.0142	0.0132	0.0057	0.0026	0.0202	0.0103	0.0081	0.0013	0.0052	0.0095	0.0220	0.0071	0.0090	0.0008

Tabelle 470: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=1.3). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW$_{Ridge}$	0.91	0.1354	0.0784	0.0258	0.0160	0.1192	0.0259	0.0220	0.0034	0.0839	0.0629	0.1314	0.0195	0.0241	0.0008
STD$_{Ridge}$		0.0159	0.0241	0.0050	0.0142	0.0151	0.0070	0.0045	0.0016	0.0055	0.0084	0.0152	0.0060	0.0047	0.0006
MW$_{Elastic\ Net}$	0.90	0.1285	0.0707	0.0229	0.0090	0.1088	0.0276	0.0190	0.0027	0.0850	0.0679	0.1184	0.0215	0.0202	0.0009
STD$_{Elastic\ Net}$		0.0172	0.0198	0.0054	0.0052	0.0162	0.0062	0.0049	0.0013	0.0061	0.0084	0.0158	0.0064	0.0047	0.0006
MW$_{Lasso}$	0.90	0.1268	0.0707	0.0224	0.0092	0.1064	0.0278	0.0183	0.0028	0.0848	0.0681	0.1164	0.0215	0.0197	0.0009
STD$_{Lasso}$		0.0157	0.0189	0.0048	0.0049	0.0147	0.0072	0.0042	0.0014	0.0062	0.0084	0.0151	0.0063	0.0045	0.0005
MW$_{Lasso\ +}$	0.90	0.1906	0.0670	0.0460	0.0086	0.2324	0.0269	0.0614	0.0027	0.1303	0.0681	0.2451	0.0214	0.0674	0.0009
STD$_{Lasso\ +}$		0.0114	0.0214	0.0051	0.0053	0.0192	0.0065	0.0089	0.0015	0.0059	0.0084	0.0186	0.0063	0.0095	0.0005

Tabelle 471: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.70	0.0319	0.0203	0.0017	0.0006	0.0316	0.0204	0.0017	0.0007	0.1915	0.1908	0.0349	0.0241	0.0019	0.0009
STD_{Ridge}		0.0082	0.0050	0.0008	0.0003	0.0078	0.0053	0.0008	0.0003	0.0037	0.0042	0.0073	0.0051	0.0008	0.0003
$MW_{Elastic\ Net}$	0.70	0.0433	0.0319	0.0028	0.0017	0.0416	0.0290	0.0026	0.0014	0.1943	0.1936	0.0396	0.0313	0.0024	0.0015
$STD_{Elastic\ Net}$		0.0093	0.0121	0.0012	0.0011	0.0086	0.0114	0.0009	0.0010	0.0051	0.0054	0.0094	0.0063	0.0010	0.0006
MW_{Lasso}	0.70	0.0430	0.0314	0.0028	0.0016	0.0415	0.0288	0.0026	0.0013	0.1942	0.1935	0.0394	0.0315	0.0024	0.0016
STD_{Lasso}		0.0110	0.0110	0.0013	0.0009	0.0100	0.0109	0.0011	0.0009	0.0052	0.0054	0.0095	0.0064	0.0010	0.0006
MW_{Lasso+}	0.70	0.0911	0.0317	0.0127	0.0017	0.0807	0.0292	0.0089	0.0014	0.2014	0.1935	0.0827	0.0315	0.0093	0.0016
STD_{Lasso+}		0.0082	0.0106	0.0049	0.0011	0.0112	0.0093	0.0020	0.0008	0.0049	0.0055	0.0129	0.0065	0.0024	0.0006

Tabelle 472: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.75	0.0421	0.0205	0.0028	0.0007	0.0433	0.0204	0.0029	0.0007	0.1700	0.1681	0.0460	0.0234	0.0031	0.0009
STD_{Ridge}		0.0080	0.0061	0.0010	0.0003	0.0073	0.0065	0.0010	0.0004	0.0038	0.0045	0.0062	0.0049	0.0009	0.0003
$MW_{Elastic\ Net}$	0.75	0.0430	0.0304	0.0028	0.0015	0.0432	0.0301	0.0028	0.0015	0.1716	0.1703	0.0470	0.0318	0.0033	0.0016
$STD_{Elastic\ Net}$		0.0124	0.0062	0.0013	0.0006	0.0119	0.0062	0.0012	0.0006	0.0048	0.0055	0.0123	0.0040	0.0014	0.0004
MW_{Lasso}	0.75	0.0428	0.0300	0.0028	0.0015	0.0432	0.0298	0.0028	0.0015	0.1715	0.1703	0.0465	0.0319	0.0032	0.0016
STD_{Lasso}		0.0123	0.0060	0.0013	0.0006	0.0119	0.0059	0.0012	0.0005	0.0048	0.0055	0.0124	0.0040	0.0014	0.0004
MW_{Lasso+}	0.75	0.1183	0.0283	0.0182	0.0014	0.1188	0.0280	0.0175	0.0013	0.1871	0.1703	0.1200	0.0319	0.0182	0.0016
STD_{Lasso+}		0.0085	0.0074	0.0032	0.0006	0.0113	0.0071	0.0031	0.0006	0.0050	0.0055	0.0132	0.0044	0.0034	0.0005

Tabelle 473: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.81	0.0637	0.0233	0.0059	0.0010	0.0605	0.0184	0.0055	0.0007	0.1396	0.1344	0.0685	0.0188	0.0063	0.0006
STD_{Ridge}		0.0062	0.0078	0.0012	0.0007	0.0050	0.0061	0.0011	0.0005	0.0036	0.0047	0.0069	0.0039	0.0013	0.0003
$MW_{Elastic\ Net}$	0.80	0.0586	0.0299	0.0054	0.0018	0.0560	0.0250	0.0051	0.0013	0.1399	0.1365	0.0596	0.0270	0.0053	0.0012

STD$_{Elastic\ Net}$		0.0105	0.0072	0.0016	0.0010	0.0098	0.0059	0.0015	0.0006	0.0042	0.0054	0.0105	0.0047	0.0016	0.0004
MW$_{Lasso}$	0.80	0.0582	0.0302	0.0053	0.0018	0.0555	0.0253	0.0050	0.0013	0.1398	0.1365	0.0589	0.0271	0.0052	0.0012
STD$_{Lasso}$		0.0111	0.0080	0.0017	0.0011	0.0102	0.0065	0.0016	0.0007	0.0042	0.0053	0.0106	0.0047	0.0016	0.0004
MW$_{Lasso\ +}$	0.80	0.1493	0.0319	0.0280	0.0021	0.1647	0.0261	0.0321	0.0014	0.1685	0.1365	0.1675	0.0264	0.0338	0.0012
STD$_{Lasso\ +}$		0.0112	0.0084	0.0040	0.0011	0.0158	0.0066	0.0046	0.0007	0.0052	0.0054	0.0136	0.0050	0.0044	0.0004

Tabelle 474: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.33). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.85	0.0878	0.0366	0.0107	0.0025	0.0826	0.0243	0.0100	0.0013	0.1188	0.1100	0.0914	0.0237	0.0109	0.0010
STD$_{Ridge}$		0.0085	0.0097	0.0019	0.0013	0.0088	0.0054	0.0019	0.0006	0.0036	0.0050	0.0089	0.0044	0.0020	0.0004
MW$_{Elastic\ Net}$	0.85	0.0782	0.0393	0.0089	0.0030	0.0700	0.0252	0.0078	0.0016	0.1172	0.1108	0.0774	0.0244	0.0083	0.0011
STD$_{Elastic\ Net}$		0.0095	0.0104	0.0022	0.0015	0.0086	0.0057	0.0019	0.0006	0.0038	0.0050	0.0095	0.0044	0.0019	0.0004
MW$_{Lasso}$	0.85	0.0777	0.0398	0.0088	0.0031	0.0695	0.0255	0.0078	0.0016	0.1171	0.1108	0.0769	0.0244	0.0083	0.0011
STD$_{Lasso}$		0.0098	0.0112	0.0023	0.0016	0.0087	0.0062	0.0020	0.0007	0.0038	0.0050	0.0096	0.0044	0.0020	0.0004
MW$_{Lasso\ +}$	0.85	0.1663	0.0401	0.0349	0.0031	0.1927	0.0258	0.0435	0.0016	0.1552	0.1107	0.2009	0.0237	0.0471	0.0010
STD$_{Lasso\ +}$		0.0095	0.0105	0.0036	0.0015	0.0164	0.0059	0.0051	0.0007	0.0054	0.0050	0.0133	0.0046	0.0050	0.0005

Tabelle 475: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.89	0.1133	0.0446	0.0173	0.0036	0.0984	0.0205	0.0148	0.0014	0.0930	0.0790	0.1088	0.0179	0.0161	0.0006
STD$_{Ridge}$		0.0069	0.0080	0.0017	0.0015	0.0068	0.0034	0.0016	0.0005	0.0032	0.0043	0.0065	0.0035	0.0019	0.0003
MW$_{Elastic\ Net}$	0.89	0.1047	0.0460	0.0151	0.0036	0.0872	0.0210	0.0123	0.0013	0.0912	0.0798	0.0970	0.0183	0.0133	0.0007
STD$_{Elastic\ Net}$		0.0096	0.0103	0.0025	0.0016	0.0083	0.0050	0.0019	0.0005	0.0033	0.0044	0.0068	0.0036	0.0019	0.0003
MW$_{Lasso}$	0.89	0.1046	0.0462	0.0151	0.0036	0.0871	0.0212	0.0122	0.0014	0.0912	0.0798	0.0968	0.0184	0.0133	0.0007
STD$_{Lasso}$		0.0097	0.0100	0.0025	0.0015	0.0082	0.0049	0.0019	0.0005	0.0033	0.0044	0.0068	0.0036	0.0019	0.0003
MW$_{Lasso\ +}$	0.89	0.1846	0.0434	0.0430	0.0034	0.2259	0.0200	0.0578	0.0013	0.1391	0.0798	0.2359	0.0179	0.0630	0.0006
STD$_{Lasso\ +}$		0.0063	0.0124	0.0029	0.0018	0.0130	0.0054	0.0048	0.0006	0.0055	0.0044	0.0106	0.0036	0.0053	0.0003

Tabelle 476: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.71	0.0542	0.0304	0.0043	0.0015	0.0512	0.0289	0.0040	0.0013	0.1898	0.1873	0.0536	0.0303	0.0041	0.0014
STD_{Ridge}		0.0089	0.0080	0.0012	0.0007	0.0100	0.0079	0.0013	0.0006	0.0043	0.0049	0.0117	0.0057	0.0016	0.0005
$MW_{Elastic\ Net}$	0.71	0.0433	0.0313	0.0031	0.0016	0.0395	0.0288	0.0026	0.0014	0.1925	0.1915	0.0394	0.0304	0.0025	0.0014
$STD_{Elastic\ Net}$		0.0102	0.0069	0.0013	0.0007	0.0109	0.0091	0.0012	0.0008	0.0046	0.0048	0.0116	0.0058	0.0013	0.0006
MW_{Lasso}	0.70	0.0427	0.0306	0.0030	0.0015	0.0391	0.0282	0.0025	0.0013	0.1926	0.1917	0.0390	0.0306	0.0024	0.0014
STD_{Lasso}		0.0112	0.0072	0.0013	0.0007	0.0115	0.0088	0.0011	0.0008	0.0046	0.0048	0.0112	0.0056	0.0013	0.0006
MW_{Lasso+}	0.70	0.0878	0.0307	0.0118	0.0015	0.0797	0.0282	0.0092	0.0013	0.1999	0.1917	0.0822	0.0306	0.0097	0.0015
STD_{Lasso+}		0.0186	0.0068	0.0083	0.0007	0.0117	0.0078	0.0025	0.0007	0.0053	0.0048	0.0153	0.0057	0.0032	0.0006

Tabelle 477: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.76	0.0606	0.0269	0.0054	0.0012	0.0609	0.0266	0.0055	0.0012	0.1659	0.1616	0.0671	0.0290	0.0062	0.0013
STD_{Ridge}		0.0084	0.0061	0.0012	0.0006	0.0098	0.0059	0.0013	0.0006	0.0046	0.0055	0.0112	0.0048	0.0018	0.0004
$MW_{Elastic\ Net}$	0.75	0.0498	0.0307	0.0038	0.0015	0.0494	0.0300	0.0038	0.0015	0.1680	0.1658	0.0504	0.0294	0.0039	0.0014
$STD_{Elastic\ Net}$		0.0112	0.0095	0.0016	0.0009	0.0121	0.0092	0.0017	0.0009	0.0048	0.0053	0.0126	0.0074	0.0019	0.0008
MW_{Lasso}	0.75	0.0488	0.0299	0.0036	0.0015	0.0486	0.0292	0.0036	0.0014	0.1681	0.1659	0.0495	0.0293	0.0038	0.0014
STD_{Lasso}		0.0110	0.0095	0.0016	0.0009	0.0120	0.0093	0.0018	0.0009	0.0048	0.0054	0.0126	0.0075	0.0019	0.0008
MW_{Lasso+}	0.75	0.1179	0.0314	0.0194	0.0015	0.1181	0.0310	0.0184	0.0015	0.1836	0.1659	0.1202	0.0295	0.0192	0.0015
STD_{Lasso+}		0.0166	0.0074	0.0081	0.0008	0.0159	0.0072	0.0041	0.0007	0.0050	0.0054	0.0163	0.0073	0.0043	0.0007

Tabelle 478: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.80	0.0730	0.0313	0.0075	0.0017	0.0726	0.0261	0.0075	0.0013	0.1431	0.1363	0.0809	0.0253	0.0087	0.0011
STD_{Ridge}		0.0086	0.0072	0.0017	0.0009	0.0097	0.0058	0.0019	0.0007	0.0045	0.0058	0.0108	0.0038	0.0020	0.0004
$MW_{Elastic\ Net}$	0.80	0.0617	0.0317	0.0057	0.0018	0.0594	0.0274	0.0054	0.0014	0.1446	0.1404	0.0642	0.0260	0.0059	0.0011

STD$_{Elastic\ Net}$		0.0116	0.0120	0.0019	0.0015	0.0124	0.0092	0.0020	0.0011	0.0049	0.0057	0.0111	0.0058	0.0020	0.0006
MW$_{Lasso}$	0.80	0.0606	0.0321	0.0055	0.0018	0.0582	0.0278	0.0052	0.0014	0.1446	0.1405	0.0633	0.0260	0.0058	0.0011
STD$_{Lasso}$		0.0114	0.0120	0.0018	0.0015	0.0122	0.0091	0.0019	0.0010	0.0049	0.0057	0.0110	0.0059	0.0019	0.0006
MW$_{Lasso\ +}$	0.80	0.1398	0.0334	0.0250	0.0019	0.1507	0.0292	0.0283	0.0015	0.1690	0.1404	0.1568	0.0264	0.0302	0.0011
STD$_{Lasso\ +}$		0.0097	0.0113	0.0033	0.0015	0.0181	0.0082	0.0050	0.0010	0.0051	0.0057	0.0163	0.0054	0.0051	0.0006

Tabelle 479: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.72). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.85	0.0953	0.0289	0.0124	0.0015	0.0895	0.0189	0.0115	0.0009	0.1189	0.1081	0.0978	0.0205	0.0125	0.0007
STD$_{Ridge}$		0.0088	0.0080	0.0020	0.0010	0.0085	0.0048	0.0019	0.0005	0.0043	0.0057	0.0090	0.0029	0.0020	0.0003
MW$_{Elastic\ Net}$	0.84	0.0834	0.0324	0.0094	0.0020	0.0755	0.0221	0.0084	0.0012	0.1196	0.1119	0.0822	0.0218	0.0093	0.0008
STD$_{Elastic\ Net}$		0.0076	0.0065	0.0017	0.0009	0.0077	0.0046	0.0017	0.0006	0.0048	0.0057	0.0082	0.0030	0.0018	0.0003
MW$_{Lasso}$	0.84	0.0827	0.0321	0.0093	0.0020	0.0747	0.0218	0.0083	0.0012	0.1196	0.1120	0.0816	0.0217	0.0092	0.0008
STD$_{Lasso}$		0.0080	0.0065	0.0018	0.0008	0.0079	0.0047	0.0017	0.0005	0.0048	0.0057	0.0082	0.0029	0.0018	0.0003
MW$_{Lasso\ +}$	0.84	0.1632	0.0336	0.0330	0.0021	0.1876	0.0233	0.0408	0.0013	0.1539	0.1119	0.1950	0.0220	0.0442	0.0008
STD$_{Lasso\ +}$		0.0076	0.0066	0.0030	0.0010	0.0156	0.0044	0.0053	0.0006	0.0057	0.0057	0.0148	0.0028	0.0057	0.0002

Tabelle 480: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.86). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.89	0.1180	0.0396	0.0186	0.0029	0.1041	0.0185	0.0161	0.0011	0.0953	0.0796	0.1138	0.0167	0.0175	0.0005
STD$_{Ridge}$		0.0064	0.0138	0.0017	0.0018	0.0060	0.0049	0.0016	0.0006	0.0040	0.0053	0.0064	0.0037	0.0017	0.0002
MW$_{Elastic\ Net}$	0.89	0.1064	0.0406	0.0156	0.0031	0.0897	0.0200	0.0128	0.0012	0.0954	0.0829	0.0997	0.0192	0.0141	0.0007
STD$_{Elastic\ Net}$		0.0079	0.0112	0.0019	0.0017	0.0072	0.0047	0.0017	0.0006	0.0044	0.0053	0.0062	0.0040	0.0017	0.0003
MW$_{Lasso}$	0.89	0.1064	0.0406	0.0155	0.0030	0.0893	0.0201	0.0128	0.0012	0.0954	0.0830	0.0991	0.0192	0.0140	0.0007
STD$_{Lasso}$		0.0075	0.0095	0.0019	0.0014	0.0068	0.0041	0.0016	0.0005	0.0045	0.0054	0.0061	0.0040	0.0017	0.0003
MW$_{Lasso\ +}$	0.89	0.1827	0.0400	0.0411	0.0029	0.2218	0.0205	0.0546	0.0012	0.1389	0.0830	0.2304	0.0193	0.0591	0.0007
STD$_{Lasso\ +}$		0.0063	0.0123	0.0030	0.0016	0.0145	0.0046	0.0059	0.0006	0.0063	0.0054	0.0130	0.0040	0.0060	0.0003

Tabelle 481: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.55). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.72	0.0556	0.0270	0.0042	0.0012	0.0538	0.0261	0.0040	0.0011	0.1860	0.1834	0.0546	0.0288	0.0041	0.0013
STD_{Ridge}		0.0111	0.0052	0.0015	0.0004	0.0112	0.0042	0.0014	0.0003	0.0045	0.0050	0.0107	0.0041	0.0015	0.0003
$MW_{Elastic\ Net}$	0.72	0.0451	0.0317	0.0031	0.0015	0.0438	0.0307	0.0029	0.0014	0.1885	0.1872	0.0457	0.0315	0.0031	0.0016
$STD_{Elastic\ Net}$		0.0108	0.0077	0.0014	0.0008	0.0092	0.0083	0.0011	0.0008	0.0046	0.0048	0.0115	0.0075	0.0014	0.0007
MW_{Lasso}	0.72	0.0445	0.0304	0.0029	0.0017	0.0436	0.0286	0.0028	0.0013	0.1887	0.1874	0.0443	0.0313	0.0030	0.0016
STD_{Lasso}		0.0101	0.0096	0.0013	0.0012	0.0089	0.0100	0.0011	0.0009	0.0050	0.0051	0.0122	0.0077	0.0014	0.0007
MW_{Lasso+}	0.72	0.0973	0.0306	0.0209	0.0015	0.0862	0.0296	0.0108	0.0014	0.1976	0.1874	0.0912	0.0316	0.0117	0.0016
STD_{Lasso+}		0.0352	0.0081	0.0334	0.0008	0.0124	0.0085	0.0033	0.0008	0.0058	0.0051	0.0127	0.0075	0.0032	0.0007

Tabelle 482: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.76	0.0586	0.0260	0.0049	0.0010	0.0592	0.0259	0.0050	0.0010	0.1681	0.1643	0.0637	0.0283	0.0056	0.0012
STD_{Ridge}		0.0081	0.0066	0.0011	0.0005	0.0086	0.0067	0.0012	0.0005	0.0046	0.0054	0.0101	0.0041	0.0015	0.0004
$MW_{Elastic\ Net}$	0.75	0.0512	0.0294	0.0038	0.0013	0.0521	0.0290	0.0039	0.0013	0.1704	0.1681	0.0536	0.0302	0.0041	0.0015
$STD_{Elastic\ Net}$		0.0090	0.0042	0.0011	0.0003	0.0103	0.0041	0.0012	0.0003	0.0047	0.0052	0.0099	0.0045	0.0014	0.0004
MW_{Lasso}	0.75	0.0499	0.0288	0.0036	0.0013	0.0506	0.0285	0.0037	0.0013	0.1705	0.1682	0.0525	0.0301	0.0039	0.0014
STD_{Lasso}		0.0087	0.0043	0.0012	0.0003	0.0097	0.0043	0.0013	0.0003	0.0048	0.0052	0.0100	0.0043	0.0014	0.0004
MW_{Lasso+}	0.75	0.1144	0.0278	0.0184	0.0012	0.1136	0.0275	0.0172	0.0012	0.1853	0.1682	0.1181	0.0304	0.0185	0.0015
STD_{Lasso+}		0.0184	0.0045	0.0083	0.0004	0.0158	0.0045	0.0040	0.0004	0.0052	0.0052	0.0139	0.0039	0.0037	0.0004

Tabelle 483: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.80	0.0726	0.0296	0.0075	0.0015	0.0720	0.0247	0.0074	0.0012	0.1426	0.1360	0.0798	0.0249	0.0084	0.0010
STD_{Ridge}		0.0075	0.0070	0.0014	0.0008	0.0086	0.0048	0.0015	0.0005	0.0046	0.0058	0.0100	0.0034	0.0018	0.0004
$MW_{Elastic\ Net}$	0.80	0.0640	0.0325	0.0058	0.0019	0.0624	0.0277	0.0057	0.0014	0.1443	0.1397	0.0674	0.0262	0.0063	0.0011

STD$_{Elastic\ Net}$		0.0078	0.0104	0.0013	0.0014	0.0089	0.0069	0.0015	0.0009	0.0047	0.0055	0.0098	0.0042	0.0018	0.0004
MW$_{Lasso}$	0.80	0.0632	0.0331	0.0057	0.0020	0.0616	0.0280	0.0055	0.0015	0.1444	0.1399	0.0666	0.0262	0.0062	0.0011
STD$_{Lasso}$		0.0082	0.0114	0.0014	0.0014	0.0091	0.0078	0.0016	0.0009	0.0047	0.0055	0.0099	0.0043	0.0018	0.0005
MW$_{Lasso\ +}$	0.80	0.1428	0.0319	0.0259	0.0019	0.1532	0.0276	0.0288	0.0015	0.1689	0.1399	0.1583	0.0266	0.0307	0.0012
STD$_{Lasso\ +}$		0.0113	0.0089	0.0039	0.0012	0.0167	0.0058	0.0047	0.0008	0.0052	0.0055	0.0154	0.0042	0.0048	0.0004

Tabelle 484: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.95). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.85	0.0915	0.0280	0.0115	0.0014	0.0855	0.0188	0.0106	0.0008	0.1198	0.1097	0.0946	0.0200	0.0118	0.0007
STD$_{Ridge}$		0.0076	0.0080	0.0016	0.0006	0.0072	0.0055	0.0015	0.0004	0.0044	0.0057	0.0085	0.0026	0.0019	0.0003
MW$_{Elastic\ Net}$	0.84	0.0841	0.0367	0.0098	0.0024	0.0763	0.0245	0.0088	0.0014	0.1211	0.1132	0.0837	0.0225	0.0095	0.0008
STD$_{Elastic\ Net}$		0.0090	0.0071	0.0019	0.0010	0.0092	0.0030	0.0019	0.0005	0.0045	0.0055	0.0086	0.0031	0.0019	0.0003
MW$_{Lasso}$	0.84	0.0839	0.0382	0.0098	0.0026	0.0758	0.0256	0.0087	0.0015	0.1211	0.1134	0.0830	0.0225	0.0094	0.0008
STD$_{Lasso}$		0.0085	0.0090	0.0018	0.0012	0.0087	0.0042	0.0018	0.0006	0.0046	0.0056	0.0083	0.0032	0.0018	0.0003
MW$_{Lasso\ +}$	0.84	0.1627	0.0369	0.0328	0.0023	0.1881	0.0257	0.0410	0.0014	0.1547	0.1133	0.1941	0.0227	0.0435	0.0009
STD$_{Lasso\ +}$		0.0088	0.0084	0.0033	0.0011	0.0165	0.0037	0.0056	0.0006	0.0057	0.0056	0.0144	0.0032	0.0058	0.0003

Tabelle 485: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=1.2). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.90	0.1229	0.0460	0.0202	0.0040	0.1054	0.0188	0.0169	0.0013	0.0892	0.0726	0.1150	0.0157	0.0183	0.0005
STD$_{Ridge}$		0.0054	0.0128	0.0017	0.0022	0.0050	0.0042	0.0014	0.0006	0.0039	0.0051	0.0064	0.0034	0.0017	0.0002
MW$_{Elastic\ Net}$	0.90	0.1135	0.0410	0.0176	0.0033	0.0952	0.0192	0.0145	0.0012	0.0898	0.0756	0.1053	0.0182	0.0159	0.0006
STD$_{Elastic\ Net}$		0.0068	0.0088	0.0016	0.0015	0.0064	0.0042	0.0015	0.0004	0.0040	0.0050	0.0061	0.0031	0.0017	0.0002
MW$_{Lasso}$	0.90	0.1132	0.0415	0.0175	0.0033	0.0947	0.0196	0.0143	0.0012	0.0897	0.0757	0.1046	0.0183	0.0158	0.0006
STD$_{Lasso}$		0.0067	0.0078	0.0016	0.0013	0.0062	0.0040	0.0015	0.0004	0.0040	0.0050	0.0061	0.0032	0.0017	0.0002
MW$_{Lasso\ +}$	0.90	0.1859	0.0406	0.0425	0.0033	0.2280	0.0195	0.0574	0.0012	0.1351	0.0757	0.2385	0.0183	0.0629	0.0006
STD$_{Lasso\ +}$		0.0094	0.0089	0.0036	0.0011	0.0174	0.0048	0.0068	0.0004	0.0064	0.0050	0.0124	0.0033	0.0062	0.0002

Tabelle 486: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.70	0.0319	0.0203	0.0017	0.0006	0.0316	0.0204	0.0017	0.0007	0.1915	0.1908	0.0349	0.0241	0.0019	0.0009
STD_{Ridge}		0.0082	0.0050	0.0008	0.0003	0.0078	0.0053	0.0008	0.0003	0.0037	0.0042	0.0073	0.0051	0.0008	0.0003
$MW_{Elastic\ Net}$	0.71	0.0314	0.0212	0.0014	0.0007	0.0303	0.0214	0.0014	0.0007	0.1918	0.1915	0.0295	0.0240	0.0014	0.0010
$STD_{Elastic\ Net}$		0.0061	0.0059	0.0004	0.0004	0.0067	0.0061	0.0005	0.0004	0.0036	0.0038	0.0063	0.0049	0.0005	0.0003
MW_{Lasso}	0.70	0.0321	0.0209	0.0015	0.0007	0.0307	0.0208	0.0014	0.0007	0.1913	0.1910	0.0294	0.0244	0.0014	0.0010
STD_{Lasso}		0.0075	0.0054	0.0006	0.0004	0.0074	0.0052	0.0006	0.0003	0.0040	0.0043	0.0062	0.0050	0.0005	0.0003
MW_{Lasso+}	0.70	0.0972	0.0199	0.0121	0.0006	0.0871	0.0202	0.0104	0.0007	0.2011	0.1910	0.0877	0.0243	0.0104	0.0010
STD_{Lasso+}		0.0118	0.0045	0.0027	0.0003	0.0122	0.0045	0.0020	0.0003	0.0034	0.0042	0.0099	0.0050	0.0022	0.0003

Tabelle 487: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.75	0.0421	0.0205	0.0028	0.0007	0.0433	0.0204	0.0029	0.0007	0.1700	0.1681	0.0460	0.0234	0.0031	0.0009
STD_{Ridge}		0.0080	0.0061	0.0010	0.0003	0.0073	0.0065	0.0010	0.0004	0.0038	0.0045	0.0062	0.0049	0.0009	0.0003
$MW_{Elastic\ Net}$	0.75	0.0363	0.0199	0.0020	0.0007	0.0368	0.0198	0.0021	0.0007	0.1700	0.1688	0.0388	0.0224	0.0023	0.0009
$STD_{Elastic\ Net}$		0.0064	0.0054	0.0006	0.0003	0.0061	0.0057	0.0007	0.0003	0.0036	0.0041	0.0058	0.0053	0.0007	0.0004
MW_{Lasso}	0.75	0.0374	0.0213	0.0022	0.0007	0.0377	0.0211	0.0022	0.0007	0.1696	0.1684	0.0387	0.0236	0.0023	0.0009
STD_{Lasso}		0.0084	0.0068	0.0008	0.0004	0.0078	0.0070	0.0008	0.0004	0.0040	0.0046	0.0057	0.0052	0.0007	0.0004
MW_{Lasso+}	0.75	0.1244	0.0204	0.0196	0.0007	0.1234	0.0205	0.0196	0.0007	0.1880	0.1683	0.1251	0.0233	0.0199	0.0009
STD_{Lasso+}		0.0071	0.0066	0.0019	0.0004	0.0093	0.0072	0.0023	0.0004	0.0038	0.0045	0.0076	0.0054	0.0025	0.0004

Tabelle 488: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.81	0.0637	0.0233	0.0059	0.0010	0.0605	0.0184	0.0055	0.0007	0.1396	0.1344	0.0685	0.0188	0.0063	0.0006
STD_{Ridge}		0.0062	0.0078	0.0012	0.0007	0.0050	0.0061	0.0011	0.0005	0.0036	0.0047	0.0069	0.0039	0.0013	0.0003
$MW_{Elastic\ Net}$	0.81	0.0578	0.0240	0.0049	0.0010	0.0538	0.0191	0.0045	0.0008	0.1388	0.1347	0.0613	0.0190	0.0053	0.0006

STD$_{Elastic\ Net}$		0.0074	0.0069	0.0012	0.0006	0.0063	0.0057	0.0011	0.0004	0.0037	0.0047	0.0070	0.0043	0.0012	0.0003
MW$_{Lasso}$	0.81	0.0578	0.0241	0.0049	0.0011	0.0538	0.0192	0.0045	0.0008	0.1388	0.1347	0.0612	0.0190	0.0052	0.0006
STD$_{Lasso}$		0.0074	0.0070	0.0012	0.0006	0.0063	0.0058	0.0011	0.0004	0.0037	0.0047	0.0069	0.0043	0.0012	0.0003
MW$_{Lasso\ +}$	0.81	0.1559	0.0237	0.0301	0.0009	0.1712	0.0191	0.0349	0.0007	0.1708	0.1346	0.1759	0.0187	0.0368	0.0006
STD$_{Lasso\ +}$		0.0051	0.0058	0.0022	0.0005	0.0071	0.0047	0.0030	0.0003	0.0041	0.0047	0.0067	0.0040	0.0029	0.0003

Tabelle 489: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.32). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.84	0.0821	0.0227	0.0095	0.0009	0.0750	0.0154	0.0086	0.0005	0.1215	0.1133	0.0836	0.0149	0.0093	0.0004
STD$_{Ridge}$		0.0096	0.0031	0.0020	0.0002	0.0082	0.0030	0.0018	0.0002	0.0033	0.0046	0.0084	0.0026	0.0018	0.0001
MW$_{Elastic\ Net}$	0.84	0.0772	0.0217	0.0084	0.0009	0.0690	0.0149	0.0074	0.0005	0.1207	0.1136	0.0773	0.0152	0.0081	0.0004
STD$_{Elastic\ Net}$		0.0087	0.0052	0.0018	0.0004	0.0073	0.0038	0.0015	0.0003	0.0034	0.0047	0.0085	0.0027	0.0017	0.0001
MW$_{Lasso}$	0.84	0.0770	0.0217	0.0084	0.0009	0.0689	0.0149	0.0074	0.0005	0.1207	0.1136	0.0772	0.0152	0.0081	0.0004
STD$_{Lasso}$		0.0086	0.0055	0.0017	0.0004	0.0072	0.0039	0.0015	0.0003	0.0034	0.0047	0.0084	0.0027	0.0017	0.0001
MW$_{Lasso\ +}$	0.84	0.1706	0.0204	0.0360	0.0007	0.1971	0.0143	0.0449	0.0004	0.1607	0.1136	0.2046	0.0149	0.0483	0.0004
STD$_{Lasso\ +}$		0.0038	0.0043	0.0018	0.0003	0.0060	0.0031	0.0027	0.0002	0.0042	0.0046	0.0076	0.0025	0.0032	0.0001

Tabelle 490: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.90	0.1162	0.0369	0.0179	0.0025	0.0994	0.0158	0.0151	0.0009	0.0918	0.0770	0.1083	0.0133	0.0163	0.0004
STD$_{Ridge}$		0.0114	0.0103	0.0025	0.0012	0.0088	0.0049	0.0020	0.0004	0.0028	0.0042	0.0084	0.0033	0.0020	0.0002
MW$_{Elastic\ Net}$	0.89	0.1123	0.0362	0.0169	0.0024	0.0940	0.0157	0.0139	0.0009	0.0908	0.0773	0.1028	0.0134	0.0150	0.0004
STD$_{Elastic\ Net}$		0.0112	0.0087	0.0025	0.0010	0.0084	0.0040	0.0020	0.0003	0.0028	0.0042	0.0082	0.0033	0.0019	0.0002
MW$_{Lasso}$	0.89	0.1123	0.0364	0.0169	0.0025	0.0940	0.0158	0.0139	0.0009	0.0909	0.0773	0.1028	0.0134	0.0150	0.0004
STD$_{Lasso}$		0.0112	0.0090	0.0025	0.0011	0.0084	0.0041	0.0020	0.0004	0.0028	0.0043	0.0082	0.0033	0.0019	0.0002
MW$_{Lasso\ +}$	0.89	0.1921	0.0357	0.0456	0.0024	0.2379	0.0158	0.0627	0.0009	0.1429	0.0773	0.2484	0.0134	0.0687	0.0004
STD$_{Lasso\ +}$		0.0040	0.0089	0.0015	0.0010	0.0064	0.0044	0.0026	0.0004	0.0043	0.0042	0.0079	0.0033	0.0032	0.0002

Tabelle 491: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.71	0.0415	0.0204	0.0024	0.0007	0.0405	0.0204	0.0024	0.0006	0.1934	0.1918	0.0434	0.0207	0.0027	0.0007
STD_{Ridge}		0.0117	0.0069	0.0012	0.0004	0.0124	0.0081	0.0012	0.0004	0.0035	0.0046	0.0102	0.0047	0.0011	0.0003
$MW_{Elastic\ Net}$	0.70	0.0374	0.0205	0.0021	0.0007	0.0343	0.0203	0.0018	0.0007	0.1943	0.1937	0.0344	0.0227	0.0018	0.0009
$STD_{Elastic\ Net}$		0.0075	0.0046	0.0006	0.0003	0.0089	0.0057	0.0007	0.0003	0.0036	0.0043	0.0065	0.0047	0.0006	0.0004
MW_{Lasso}	0.70	0.0373	0.0203	0.0020	0.0007	0.0340	0.0202	0.0017	0.0007	0.1944	0.1937	0.0341	0.0227	0.0017	0.0009
STD_{Lasso}		0.0073	0.0044	0.0006	0.0003	0.0090	0.0054	0.0006	0.0003	0.0036	0.0043	0.0064	0.0046	0.0006	0.0004
MW_{Lasso+}	0.70	0.0874	0.0215	0.0114	0.0007	0.0804	0.0210	0.0092	0.0007	0.2025	0.1937	0.0832	0.0228	0.0097	0.0009
STD_{Lasso+}		0.0192	0.0047	0.0077	0.0003	0.0127	0.0057	0.0024	0.0003	0.0032	0.0042	0.0116	0.0046	0.0021	0.0004

Tabelle 492: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.76	0.0556	0.0180	0.0044	0.0006	0.0568	0.0174	0.0046	0.0005	0.1649	0.1611	0.0605	0.0188	0.0049	0.0006
STD_{Ridge}		0.0099	0.0051	0.0014	0.0005	0.0103	0.0050	0.0016	0.0004	0.0037	0.0053	0.0109	0.0029	0.0016	0.0002
$MW_{Elastic\ Net}$	0.76	0.0459	0.0191	0.0030	0.0006	0.0461	0.0187	0.0031	0.0006	0.1651	0.1628	0.0487	0.0187	0.0033	0.0006
$STD_{Elastic\ Net}$		0.0101	0.0061	0.0012	0.0003	0.0109	0.0060	0.0014	0.0003	0.0039	0.0052	0.0105	0.0033	0.0014	0.0002
MW_{Lasso}	0.76	0.0453	0.0190	0.0030	0.0006	0.0455	0.0185	0.0030	0.0006	0.1651	0.1628	0.0481	0.0186	0.0032	0.0006
STD_{Lasso}		0.0098	0.0066	0.0012	0.0003	0.0106	0.0065	0.0013	0.0003	0.0039	0.0052	0.0106	0.0032	0.0014	0.0002
MW_{Lasso+}	0.76	0.1287	0.0190	0.0206	0.0006	0.1311	0.0186	0.0214	0.0006	0.1847	0.1628	0.1325	0.0191	0.0223	0.0006
STD_{Lasso+}		0.0119	0.0054	0.0033	0.0003	0.0116	0.0053	0.0032	0.0003	0.0034	0.0052	0.0122	0.0035	0.0035	0.0002

Tabelle 493: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.80	0.0707	0.0211	0.0073	0.0008	0.0699	0.0176	0.0072	0.0006	0.1424	0.1357	0.0786	0.0185	0.0081	0.0006
STD_{Ridge}		0.0091	0.0054	0.0016	0.0006	0.0097	0.0037	0.0017	0.0004	0.0036	0.0054	0.0090	0.0047	0.0018	0.0003
$MW_{Elastic\ Net}$	0.80	0.0624	0.0219	0.0056	0.0009	0.0601	0.0184	0.0054	0.0006	0.1420	0.1373	0.0658	0.0184	0.0060	0.0006

STD$_{Elastic\ Net}$		0.0084	0.0059	0.0014	0.0005	0.0089	0.0043	0.0015	0.0003	0.0039	0.0055	0.0094	0.0041	0.0016	0.0003
MW$_{Lasso}$	0.80	0.0619	0.0222	0.0055	0.0009	0.0593	0.0187	0.0053	0.0007	0.1419	0.1373	0.0647	0.0183	0.0059	0.0006
STD$_{Lasso}$		0.0089	0.0059	0.0015	0.0004	0.0094	0.0048	0.0016	0.0003	0.0038	0.0055	0.0099	0.0040	0.0017	0.0003
MW$_{Lasso\ +}$	0.80	0.1522	0.0225	0.0284	0.0008	0.1681	0.0196	0.0332	0.0007	0.1715	0.1373	0.1706	0.0190	0.0351	0.0006
STD$_{Lasso\ +}$		0.0055	0.0060	0.0022	0.0004	0.0086	0.0048	0.0034	0.0003	0.0034	0.0055	0.0094	0.0047	0.0039	0.0003

Tabelle 494: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.72). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.85	0.0935	0.0239	0.0120	0.0012	0.0875	0.0162	0.0112	0.0007	0.1183	0.1074	0.0989	0.0173	0.0126	0.0005
STD$_{Ridge}$		0.0079	0.0086	0.0018	0.0008	0.0081	0.0065	0.0019	0.0005	0.0033	0.0052	0.0078	0.0056	0.0019	0.0003
MW$_{Elastic\ Net}$	0.85	0.0853	0.0228	0.0102	0.0010	0.0773	0.0158	0.0092	0.0006	0.1176	0.1087	0.0879	0.0173	0.0103	0.0005
STD$_{Elastic\ Net}$		0.0071	0.0060	0.0015	0.0006	0.0073	0.0044	0.0015	0.0003	0.0035	0.0053	0.0080	0.0061	0.0018	0.0003
MW$_{Lasso}$	0.85	0.0845	0.0230	0.0101	0.0011	0.0761	0.0160	0.0090	0.0006	0.1174	0.1088	0.0865	0.0172	0.0101	0.0005
STD$_{Lasso}$		0.0074	0.0057	0.0016	0.0007	0.0074	0.0042	0.0015	0.0003	0.0035	0.0053	0.0081	0.0060	0.0018	0.0003
MW$_{Lasso\ +}$	0.85	0.1739	0.0233	0.0372	0.0011	0.2040	0.0164	0.0475	0.0006	0.1574	0.1088	0.2105	0.0177	0.0506	0.0006
STD$_{Lasso\ +}$		0.0037	0.0067	0.0020	0.0006	0.0072	0.0052	0.0036	0.0003	0.0032	0.0053	0.0071	0.0063	0.0039	0.0004

Tabelle 495: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.86). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW$_{Ridge}$	0.89	0.1185	0.0276	0.0185	0.0015	0.1052	0.0134	0.0162	0.0006	0.0950	0.0788	0.1163	0.0137	0.0179	0.0004
STD$_{Ridge}$		0.0064	0.0048	0.0020	0.0006	0.0063	0.0037	0.0019	0.0002	0.0030	0.0047	0.0066	0.0038	0.0020	0.0002
MW$_{Elastic\ Net}$	0.89	0.1118	0.0302	0.0168	0.0017	0.0958	0.0148	0.0141	0.0007	0.0939	0.0800	0.1065	0.0141	0.0155	0.0004
STD$_{Elastic\ Net}$		0.0074	0.0068	0.0020	0.0007	0.0067	0.0045	0.0018	0.0003	0.0031	0.0048	0.0067	0.0042	0.0019	0.0002
MW$_{Lasso}$	0.89	0.1117	0.0301	0.0167	0.0017	0.0956	0.0147	0.0141	0.0007	0.0939	0.0800	0.1061	0.0141	0.0154	0.0004
STD$_{Lasso}$		0.0073	0.0061	0.0020	0.0007	0.0067	0.0041	0.0018	0.0003	0.0031	0.0048	0.0067	0.0042	0.0019	0.0002
MW$_{Lasso\ +}$	0.89	0.1892	0.0307	0.0443	0.0018	0.2329	0.0153	0.0606	0.0007	0.1432	0.0800	0.2453	0.0143	0.0662	0.0004
STD$_{Lasso\ +}$		0.0044	0.0055	0.0022	0.0006	0.0085	0.0042	0.0041	0.0002	0.0031	0.0049	0.0070	0.0042	0.0039	0.0002

Tabelle 496: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.70	0.0397	0.0204	0.0023	0.0007	0.0395	0.0214	0.0023	0.0007	0.1940	0.1925	0.0427	0.0215	0.0026	0.0008
STD_{Ridge}		0.0110	0.0045	0.0011	0.0003	0.0112	0.0060	0.0011	0.0004	0.0030	0.0038	0.0089	0.0039	0.0010	0.0003
$MW_{Elastic\ Net}$	0.70	0.0357	0.0207	0.0019	0.0007	0.0323	0.0201	0.0016	0.0006	0.1956	0.1949	0.0334	0.0211	0.0017	0.0007
$STD_{Elastic\ Net}$		0.0075	0.0049	0.0005	0.0003	0.0089	0.0050	0.0007	0.0003	0.0034	0.0041	0.0074	0.0057	0.0007	0.0004
MW_{Lasso}	0.70	0.0355	0.0212	0.0019	0.0007	0.0317	0.0208	0.0016	0.0007	0.1957	0.1950	0.0325	0.0213	0.0016	0.0008
STD_{Lasso}		0.0066	0.0044	0.0005	0.0003	0.0082	0.0046	0.0006	0.0003	0.0034	0.0041	0.0073	0.0054	0.0007	0.0004
MW_{Lasso+}	0.70	0.0903	0.0209	0.0191	0.0007	0.0788	0.0204	0.0091	0.0007	0.2036	0.1950	0.0812	0.0216	0.0094	0.0008
STD_{Lasso+}		0.0386	0.0049	0.0328	0.0003	0.0121	0.0051	0.0026	0.0003	0.0031	0.0041	0.0125	0.0052	0.0023	0.0004

Tabelle 497: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.76	0.0524	0.0173	0.0039	0.0005	0.0537	0.0170	0.0041	0.0005	0.1678	0.1645	0.0572	0.0180	0.0044	0.0005
STD_{Ridge}		0.0097	0.0059	0.0014	0.0003	0.0101	0.0058	0.0015	0.0003	0.0037	0.0052	0.0108	0.0028	0.0015	0.0002
$MW_{Elastic\ Net}$	0.75	0.0446	0.0180	0.0029	0.0006	0.0450	0.0177	0.0029	0.0005	0.1685	0.1663	0.0477	0.0196	0.0032	0.0006
$STD_{Elastic\ Net}$		0.0099	0.0043	0.0011	0.0003	0.0100	0.0040	0.0012	0.0002	0.0037	0.0049	0.0104	0.0025	0.0013	0.0002
MW_{Lasso}	0.75	0.0445	0.0184	0.0028	0.0006	0.0447	0.0181	0.0028	0.0006	0.1684	0.1664	0.0470	0.0195	0.0031	0.0006
STD_{Lasso}		0.0099	0.0046	0.0011	0.0003	0.0100	0.0043	0.0012	0.0003	0.0037	0.0049	0.0103	0.0025	0.0013	0.0002
MW_{Lasso+}	0.75	0.1255	0.0183	0.0197	0.0006	0.1252	0.0181	0.0198	0.0006	0.1867	0.1663	0.1270	0.0200	0.0208	0.0006
STD_{Lasso+}		0.0147	0.0049	0.0044	0.0003	0.0135	0.0047	0.0035	0.0002	0.0033	0.0049	0.0121	0.0028	0.0035	0.0002

Tabelle 498: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW_{Ridge}	0.80	0.0707	0.0221	0.0073	0.0008	0.0696	0.0178	0.0073	0.0006	0.1421	0.1353	0.0780	0.0178	0.0080	0.0005
STD_{Ridge}		0.0080	0.0035	0.0015	0.0003	0.0089	0.0023	0.0016	0.0002	0.0031	0.0044	0.0086	0.0030	0.0017	0.0002
$MW_{Elastic\ Net}$	0.80	0.0620	0.0212	0.0056	0.0009	0.0599	0.0182	0.0055	0.0007	0.1425	0.1377	0.0661	0.0186	0.0061	0.0006

$STD_{Elastic\ Net}$		0.0088	0.0084	0.0015	0.0007	0.0092	0.0070	0.0016	0.0005	0.0036	0.0053	0.0100	0.0044	0.0017	0.0003
MW_{Lasso}	0.80	0.0616	0.0211	0.0055	0.0008	0.0595	0.0181	0.0054	0.0007	0.1425	0.1378	0.0653	0.0186	0.0060	0.0006
STD_{Lasso}		0.0084	0.0075	0.0015	0.0006	0.0089	0.0065	0.0016	0.0004	0.0037	0.0053	0.0100	0.0044	0.0017	0.0003
MW_{Lasso+}	0.80	0.1524	0.0226	0.0285	0.0009	0.1674	0.0198	0.0330	0.0007	0.1719	0.1378	0.1699	0.0193	0.0350	0.0006
STD_{Lasso+}		0.0060	0.0073	0.0025	0.0006	0.0097	0.0058	0.0037	0.0005	0.0033	0.0053	0.0095	0.0050	0.0040	0.0003

Tabelle 499: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.95). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW_{Ridge}	0.85	0.0915	0.0226	0.0115	0.0010	0.0859	0.0157	0.0107	0.0006	0.1199	0.1095	0.0968	0.0175	0.0120	0.0005
STD_{Ridge}		0.0084	0.0070	0.0019	0.0006	0.0085	0.0059	0.0019	0.0004	0.0034	0.0052	0.0081	0.0051	0.0019	0.0002
$MW_{Elastic\ Net}$	0.84	0.0836	0.0221	0.0099	0.0010	0.0760	0.0159	0.0090	0.0006	0.1195	0.1111	0.0861	0.0176	0.0099	0.0005
$STD_{Elastic\ Net}$		0.0068	0.0094	0.0015	0.0007	0.0069	0.0069	0.0015	0.0004	0.0034	0.0053	0.0082	0.0061	0.0018	0.0003
MW_{Lasso}	0.84	0.0832	0.0226	0.0098	0.0010	0.0754	0.0162	0.0089	0.0006	0.1194	0.1111	0.0852	0.0176	0.0098	0.0005
STD_{Lasso}		0.0068	0.0091	0.0014	0.0007	0.0069	0.0067	0.0014	0.0004	0.0035	0.0053	0.0081	0.0061	0.0018	0.0003
MW_{Lasso+}	0.84	0.1727	0.0221	0.0366	0.0010	0.2017	0.0161	0.0465	0.0006	0.1587	0.1111	0.2075	0.0181	0.0494	0.0006
STD_{Lasso+}		0.0031	0.0093	0.0018	0.0006	0.0063	0.0071	0.0034	0.0004	0.0032	0.0053	0.0072	0.0064	0.0039	0.0004

Tabelle 500: Simulation Ridge, Elastic Net, Lasso (abgeschnitten) und Lasso (umskaliert(+)) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=1.2). *

	Acc	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**	**MSE_w**	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
MW_{Ridge}	0.90	0.1256	0.0289	0.0205	0.0016	0.1093	0.0129	0.0176	0.0006	0.0895	0.0722	0.1190	0.0119	0.0190	0.0003
STD_{Ridge}		0.0066	0.0075	0.0021	0.0008	0.0061	0.0041	0.0019	0.0002	0.0028	0.0046	0.0064	0.0036	0.0020	0.0002
$MW_{Elastic\ Net}$	0.90	0.1179	0.0267	0.0186	0.0013	0.0998	0.0124	0.0155	0.0005	0.0888	0.0735	0.1104	0.0124	0.0168	0.0003
$STD_{Elastic\ Net}$		0.0072	0.0053	0.0020	0.0004	0.0062	0.0034	0.0018	0.0002	0.0029	0.0046	0.0064	0.0035	0.0019	0.0002
MW_{Lasso}	0.90	0.1170	0.0280	0.0184	0.0014	0.0985	0.0128	0.0152	0.0005	0.0886	0.0735	0.1092	0.0125	0.0165	0.0003
STD_{Lasso}		0.0080	0.0060	0.0021	0.0005	0.0069	0.0041	0.0019	0.0002	0.0029	0.0047	0.0068	0.0038	0.0019	0.0002
MW_{Lasso+}	0.90	0.1927	0.0284	0.0460	0.0015	0.2392	0.0134	0.0636	0.0006	0.1399	0.0736	0.2522	0.0128	0.0695	0.0003
STD_{Lasso+}		0.0048	0.0060	0.0020	0.0004	0.0086	0.0042	0.0039	0.0002	0.0031	0.0047	0.0069	0.0037	0.0038	0.0002

Tabelle 501: Simulation SPLS (abgeschnitten) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0829	0.0589	0.0110	0.0056	0.0817	0.0553	0.0108	0.0050	0.1873	0.1882	0.0789	0.0664	0.0095	0.0072
STD		0.0200	0.0166	0.0050	0.0031	0.0196	0.0137	0.0053	0.0026	0.0137	0.0115	0.0097	0.0095	0.0026	0.0021

Tabelle 502: Simulation SPLS (abgeschnitten) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.27). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0738	0.0690	0.0083	0.0071	0.0740	0.0625	0.0083	0.0060	0.1635	0.1639	0.0761	0.0605	0.0088	0.0057
STD		0.0123	0.0105	0.0028	0.0026	0.0137	0.0102	0.0030	0.0018	0.0130	0.0116	0.0129	0.0089	0.0031	0.0016

Tabelle 503: Simulation SPLS (abgeschnitten) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.82	0.0971	0.0747	0.0138	0.0101	0.0933	0.0557	0.0132	0.0063	0.1314	0.1269	0.0957	0.0512	0.0131	0.0044
STD		0.0118	0.0202	0.0029	0.0063	0.0127	0.0145	0.0031	0.0032	0.0107	0.0107	0.0066	0.0100	0.0020	0.0018

Tabelle 504: Simulation SPLS (abgeschnitten) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.38). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.1083	0.0750	0.0180	0.0105	0.0980	0.0459	0.0160	0.0049	0.1099	0.1009	0.1059	0.0444	0.0170	0.0035
STD		0.0107	0.0232	0.0039	0.0066	0.0090	0.0164	0.0033	0.0035	0.0091	0.0108	0.0103	0.0092	0.0031	0.0014

Tabelle 505: Simulation SPLS (abgeschnitten) mit 500 Objekten (n1=250, n2=250), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.46). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1387	0.1161	0.0268	0.0271	0.1146	0.0419	0.0215	0.0069	0.0824	0.0667	0.1225	0.0338	0.0221	0.0024
STD		0.0191	0.0222	0.0060	0.0094	0.0146	0.0055	0.0044	0.0018	0.0066	0.0100	0.0154	0.0069	0.0042	0.0009

Tabelle 506: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.19). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0538	0.0425	0.0046	0.0028	0.0514	0.0422	0.0041	0.0028	0.1932	0.1921	0.0525	0.0461	0.0043	0.0032
STD		0.0124	0.0116	0.0024	0.0015	0.0104	0.0112	0.0019	0.0014	0.0120	0.0104	0.0098	0.0072	0.0014	0.0010

Tabelle 507: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.23). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0576	0.0408	0.0053	0.0027	0.0582	0.0403	0.0053	0.0027	0.1686	0.1665	0.0596	0.0436	0.0054	0.0031
STD		0.0103	0.0110	0.0017	0.0014	0.0108	0.0113	0.0018	0.0015	0.0118	0.0121	0.0117	0.0061	0.0020	0.0010

Tabelle 508: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.82	0.0769	0.0475	0.0090	0.0044	0.0727	0.0363	0.0085	0.0030	0.1359	0.1299	0.0801	0.0335	0.0095	0.0020
STD		0.0157	0.0141	0.0034	0.0029	0.0152	0.0090	0.0033	0.0017	0.0103	0.0125	0.0167	0.0058	0.0032	0.0007

Tabelle 509 Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.34). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.0961	0.0583	0.0143	0.0058	0.0859	0.0369	0.0127	0.0032	0.1132	0.1028	0.0962	0.0294	0.0136	0.0016
STD		0.0158	0.0168	0.0038	0.0035	0.0137	0.0140	0.0034	0.0022	0.0092	0.0124	0.0177	0.0110	0.0043	0.0011

Tabelle 510: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1282	0.0733	0.0228	0.0097	0.1069	0.0296	0.0184	0.0031	0.0844	0.0675	0.1166	0.0244	0.0197	0.0012
STD		0.0193	0.0108	0.0064	0.0029	0.0157	0.0112	0.0051	0.0017	0.0074	0.0116	0.0162	0.0103	0.0056	0.0010

Tabelle 511: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0439	0.0381	0.0030	0.0026	0.0441	0.0363	0.0031	0.0024	0.1843	0.1846	0.0527	0.0428	0.0043	0.0030
STD		0.0126	0.0113	0.0015	0.0013	0.0139	0.0111	0.0017	0.0013	0.0083	0.0088	0.0108	0.0106	0.0015	0.0016

Tabelle 512: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.53). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.77	0.0555	0.0437	0.0051	0.0033	0.0551	0.0410	0.0049	0.0029	0.1602	0.1588	0.0595	0.0386	0.0056	0.0025
STD		0.0131	0.0112	0.0019	0.0019	0.0144	0.0106	0.0020	0.0018	0.0082	0.0097	0.0150	0.0095	0.0024	0.0014

Tabelle 513: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0769	0.0496	0.0091	0.0052	0.0724	0.0362	0.0085	0.0033	0.1335	0.1284	0.0805	0.0356	0.0093	0.0024
STD		0.0146	0.0185	0.0029	0.0040	0.0148	0.0118	0.0029	0.0023	0.0076	0.0101	0.0152	0.0092	0.0030	0.0015

Tabelle 514 Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.78). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.86	0.1001	0.0550	0.0146	0.0061	0.0886	0.0304	0.0126	0.0028	0.1085	0.0982	0.0980	0.0284	0.0136	0.0016
STD		0.0113	0.0164	0.0036	0.0033	0.0117	0.0067	0.0033	0.0015	0.0068	0.0098	0.0150	0.0062	0.0038	0.0009

Tabelle 515: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.1 und Acc=0.97 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1296	0.0725	0.0237	0.0107	0.1082	0.0264	0.0190	0.0031	0.0827	0.0654	0.1177	0.0214	0.0202	0.0010
STD		0.0155	0.0153	0.0046	0.0045	0.0143	0.0056	0.0039	0.0012	0.0056	0.0085	0.0153	0.0047	0.0043	0.0005

Tabelle 516: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0462	0.0372	0.0035	0.0026	0.0468	0.0362	0.0036	0.0024	0.1833	0.1834	0.0534	0.0429	0.0044	0.0029
STD		0.0108	0.0123	0.0014	0.0016	0.0132	0.0132	0.0017	0.0017	0.0082	0.0088	0.0121	0.0111	0.0016	0.0017

Tabelle 517: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.68). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0499	0.0431	0.0043	0.0033	0.0506	0.0409	0.0044	0.0029	0.1653	0.1642	0.0572	0.0392	0.0052	0.0025
STD		0.0141	0.0134	0.0020	0.0019	0.0154	0.0119	0.0021	0.0015	0.0081	0.0095	0.0143	0.0103	0.0023	0.0014

Tabelle 518: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.87). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0753	0.0478	0.0090	0.0049	0.0711	0.0357	0.0084	0.0031	0.1334	0.1283	0.0801	0.0346	0.0092	0.0023
STD		0.0153	0.0175	0.0031	0.0041	0.0157	0.0111	0.0031	0.0022	0.0076	0.0100	0.0163	0.0089	0.0032	0.0014

Tabelle 519 Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.85 (mu1=0, mu2=1.0). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0960	0.0465	0.0133	0.0045	0.0859	0.0288	0.0117	0.0024	0.1143	0.1054	0.0944	0.0294	0.0125	0.0018
STD		0.0128	0.0195	0.0037	0.0033	0.0116	0.0111	0.0032	0.0018	0.0070	0.0098	0.0150	0.0083	0.0037	0.0013

Tabelle 520: Simulation SPLS (abgeschnitten) mit 1000 Objekten (n1=500, n2=500), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.35). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.92	0.1360	0.0882	0.0264	0.0172	0.1116	0.0247	0.0207	0.0032	0.0754	0.0561	0.1215	0.0187	0.0220	0.0008
STD		0.0192	0.0264	0.0057	0.0126	0.0154	0.0048	0.0043	0.0013	0.0052	0.0079	0.0144	0.0046	0.0043	0.0005

Tabelle 521: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.18). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.71	0.0415	0.0329	0.0025	0.0018	0.0396	0.0307	0.0022	0.0015	0.1926	0.1918	0.0413	0.0328	0.0026	0.0017
STD		0.0094	0.0085	0.0011	0.0008	0.0080	0.0087	0.0009	0.0006	0.0050	0.0055	0.0076	0.0079	0.0009	0.0007

Tabelle 522: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0444	0.0304	0.0030	0.0016	0.0461	0.0299	0.0032	0.0015	0.1706	0.1694	0.0468	0.0316	0.0034	0.0016
STD		0.0125	0.0062	0.0017	0.0007	0.0126	0.0057	0.0018	0.0006	0.0047	0.0058	0.0127	0.0061	0.0017	0.0006

Tabelle 523: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.28). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0631	0.0308	0.0059	0.0020	0.0606	0.0243	0.0056	0.0014	0.1382	0.1340	0.0662	0.0277	0.0063	0.0013
STD		0.0115	0.0036	0.0019	0.0007	0.0111	0.0032	0.0019	0.0004	0.0040	0.0056	0.0128	0.0028	0.0020	0.0003

Tabelle 524: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.33). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0805	0.0383	0.0094	0.0027	0.0731	0.0246	0.0084	0.0014	0.1156	0.1082	0.0828	0.0252	0.0094	0.0011
STD		0.0091	0.0085	0.0017	0.0015	0.0096	0.0043	0.0018	0.0006	0.0037	0.0052	0.0096	0.0031	0.0020	0.0003

Tabelle 525: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1092	0.0446	0.0161	0.0038	0.0918	0.0187	0.0131	0.0013	0.0898	0.0772	0.1020	0.0173	0.0146	0.0006
STD		0.0067	0.0077	0.0018	0.0012	0.0067	0.0040	0.0015	0.0005	0.0033	0.0042	0.0066	0.0024	0.0018	0.0002

Tabelle 526: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0425	0.0277	0.0027	0.0012	0.0424	0.0270	0.0027	0.0012	0.1857	0.1844	0.0444	0.0294	0.0029	0.0015
STD		0.0113	0.0080	0.0012	0.0007	0.0104	0.0090	0.0011	0.0008	0.0049	0.0051	0.0081	0.0040	0.0008	0.0004

Tabelle 527: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.77	0.0511	0.0291	0.0037	0.0015	0.0509	0.0272	0.0037	0.0013	0.1615	0.1588	0.0539	0.0263	0.0042	0.0011
STD		0.0073	0.0045	0.0011	0.0006	0.0072	0.0038	0.0010	0.0005	0.0049	0.0055	0.0088	0.0020	0.0012	0.0002

Tabelle 528: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.6). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0643	0.0303	0.0061	0.0017	0.0615	0.0250	0.0058	0.0013	0.1387	0.1338	0.0683	0.0223	0.0064	0.0009
STD		0.0081	0.0131	0.0012	0.0016	0.0078	0.0100	0.0012	0.0012	0.0048	0.0057	0.0087	0.0043	0.0013	0.0003

Tabelle 529: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.72). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0845	0.0299	0.0101	0.0018	0.0751	0.0187	0.0089	0.0009	0.1145	0.1060	0.0854	0.0184	0.0100	0.0006
STD		0.0071	0.0113	0.0015	0.0016	0.0061	0.0063	0.0014	0.0007	0.0045	0.0055	0.0063	0.0030	0.0015	0.0002

Tabelle 530: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.9). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.89	0.1105	0.0451	0.0167	0.0040	0.0923	0.0193	0.0136	0.0014	0.0913	0.0779	0.1016	0.0151	0.0148	0.0004
STD		0.0039	0.0166	0.0014	0.0030	0.0040	0.0065	0.0012	0.0010	0.0040	0.0050	0.0055	0.0033	0.0014	0.0002

Tabelle 531: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.52). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.72	0.0413	0.0256	0.0026	0.0010	0.0404	0.0248	0.0025	0.0010	0.1885	0.1872	0.0437	0.0287	0.0028	0.0014
STD		0.0082	0.0067	0.0009	0.0005	0.0086	0.0069	0.0008	0.0005	0.0049	0.0051	0.0071	0.0043	0.0007	0.0004

Tabelle 532: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.75 (mu1=0, mu2=0.66). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0507	0.0282	0.0036	0.0014	0.0503	0.0268	0.0036	0.0013	0.1632	0.1606	0.0536	0.0260	0.0041	0.0011
STD		0.0069	0.0048	0.0009	0.0005	0.0073	0.0041	0.0009	0.0004	0.0050	0.0055	0.0087	0.0024	0.0011	0.0002

Tabelle 533: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.81). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.81	0.0653	0.0310	0.0063	0.0017	0.0622	0.0251	0.0060	0.0012	0.1377	0.1326	0.0693	0.0214	0.0065	0.0008
STD		0.0088	0.0088	0.0012	0.0008	0.0086	0.0061	0.0012	0.0006	0.0048	0.0057	0.0083	0.0046	0.0013	0.0003

Tabelle 534: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.96). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0832	0.0294	0.0101	0.0017	0.0742	0.0184	0.0089	0.0009	0.1152	0.1067	0.0851	0.0185	0.0099	0.0006
STD		0.0077	0.0076	0.0016	0.0008	0.0065	0.0057	0.0015	0.0004	0.0045	0.0056	0.0064	0.0027	0.0015	0.0002

Tabelle 535: Simulation SPLS (abgeschnitten) mit 2000 Objekten (n1=1000, n2=1000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.22). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1198	0.0464	0.0197	0.0046	0.0975	0.0178	0.0155	0.0014	0.0841	0.0691	0.1065	0.0142	0.0165	0.0004
STD		0.0065	0.0173	0.0017	0.0035	0.0062	0.0057	0.0014	0.0009	0.0039	0.0048	0.0051	0.0033	0.0013	0.0002

Tabelle 536: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.16). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.68	0.0307	0.0239	0.0015	0.0009	0.0281	0.0234	0.0012	0.0008	0.2011	0.2013	0.0284	0.0239	0.0012	0.0009
STD		0.0063	0.0028	0.0006	0.0003	0.0064	0.0047	0.0005	0.0004	0.0037	0.0042	0.0055	0.0037	0.0004	0.0003

Tabelle 537: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.75 (mu1=0, mu2=0.2). *.

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.75	0.0407	0.0224	0.0025	0.0008	0.0408	0.0220	0.0025	0.0008	0.1684	0.1671	0.0414	0.0225	0.0026	0.0008
STD		0.0065	0.0060	0.0008	0.0005	0.0063	0.0059	0.0008	0.0004	0.0037	0.0044	0.0062	0.0045	0.0008	0.0003

Tabelle 538: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.26). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.79	0.0539	0.0225	0.0043	0.0008	0.0518	0.0193	0.0041	0.0006	0.1475	0.1442	0.0566	0.0195	0.0045	0.0006
STD		0.0055	0.0048	0.0010	0.0003	0.0051	0.0041	0.0009	0.0002	0.0034	0.0043	0.0057	0.0034	0.0009	0.0002

Tabelle 539: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.85 (mu1=0, mu2=0.32). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.84	0.0789	0.0227	0.0090	0.0011	0.0708	0.0146	0.0080	0.0006	0.1195	0.1117	0.0805	0.0149	0.0089	0.0004
STD		0.0093	0.0084	0.0020	0.0008	0.0073	0.0045	0.0016	0.0004	0.0030	0.0040	0.0074	0.0036	0.0016	0.0002

Tabelle 540: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.4). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1148	0.0382	0.0176	0.0031	0.0968	0.0153	0.0146	0.0010	0.0897	0.0753	0.1062	0.0133	0.0159	0.0004
STD		0.0110	0.0077	0.0026	0.0013	0.0083	0.0031	0.0020	0.0004	0.0025	0.0037	0.0084	0.0033	0.0021	0.0002

Tabelle 541: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.36). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0317	0.0187	0.0016	0.0007	0.0286	0.0180	0.0013	0.0006	0.1964	0.1960	0.0298	0.0216	0.0014	0.0008
STD		0.0055	0.0051	0.0005	0.0005	0.0066	0.0050	0.0005	0.0002	0.0038	0.0044	0.0060	0.0052	0.0006	0.0004

Tabelle 542: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.75 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0464	0.0161	0.0030	0.0004	0.0459	0.0154	0.0030	0.0004	0.1626	0.1604	0.0483	0.0172	0.0033	0.0005
STD		0.0083	0.0049	0.0011	0.0003	0.0086	0.0049	0.0012	0.0003	0.0039	0.0050	0.0097	0.0033	0.0013	0.0002

Tabelle 543: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.58). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0608	0.0219	0.0053	0.0008	0.0586	0.0183	0.0052	0.0006	0.1442	0.1400	0.0628	0.0174	0.0055	0.0005
STD		0.0079	0.0060	0.0013	0.0005	0.0083	0.0051	0.0013	0.0003	0.0038	0.0051	0.0086	0.0029	0.0015	0.0002

Tabelle 544: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.85 (mu1=0, mu2=0.71). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0853	0.0235	0.0101	0.0010	0.0771	0.0155	0.0091	0.0006	0.1175	0.1088	0.0869	0.0160	0.0100	0.0004
STD		0.0087	0.0075	0.0019	0.0005	0.0087	0.0052	0.0018	0.0003	0.0035	0.0048	0.0081	0.0033	0.0017	0.0002

Tabelle 545: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.94). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.91	0.1265	0.0269	0.0212	0.0015	0.1041	0.0101	0.0170	0.0004	0.0813	0.0643	0.1136	0.0094	0.0183	0.0002
STD		0.0073	0.0061	0.0025	0.0009	0.0064	0.0017	0.0020	0.0002	0.0027	0.0039	0.0058	0.0016	0.0018	0.0000

Tabelle 546: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.5). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.70	0.0332	0.0198	0.0016	0.0006	0.0312	0.0194	0.0014	0.0006	0.1932	0.1927	0.0308	0.0215	0.0015	0.0008
STD		0.0059	0.0074	0.0005	0.0003	0.0066	0.0076	0.0006	0.0003	0.0039	0.0045	0.0066	0.0048	0.0007	0.0004

Tabelle 547: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.65 (mu1=0, mu2=0.65). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.76	0.0432	0.0142	0.0026	0.0004	0.0431	0.0138	0.0027	0.0004	0.1661	0.1642	0.0458	0.0173	0.0030	0.0005
STD		0.0089	0.0042	0.0010	0.0002	0.0093	0.0041	0.0011	0.0002	0.0040	0.0050	0.0099	0.0035	0.0012	0.0002

Tabelle 548: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.80	0.0621	0.0214	0.0057	0.0008	0.0594	0.0177	0.0054	0.0006	0.1404	0.1357	0.0660	0.0173	0.0060	0.0005
STD		0.0081	0.0064	0.0014	0.0005	0.0083	0.0050	0.0013	0.0003	0.0038	0.0051	0.0086	0.0030	0.0015	0.0002

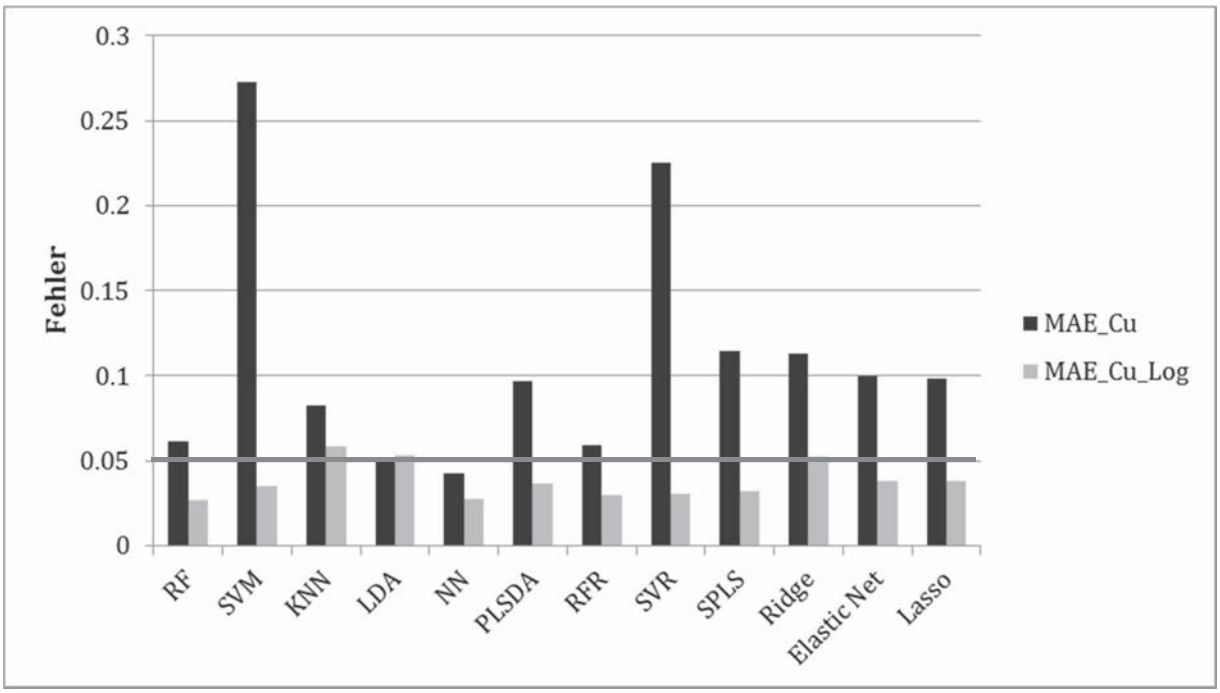
Tabelle 549: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.95 (mu1=0, mu2=0.95). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.85	0.0848	0.0246	0.0100	0.0012	0.0766	0.0160	0.0090	0.0006	0.1177	0.1091	0.0867	0.0160	0.0100	0.0004
STD		0.0090	0.0055	0.0019	0.0005	0.0090	0.0043	0.0018	0.0003	0.0035	0.0048	0.0081	0.0031	0.0017	0.0002

Tabelle 550: Simulation SPLS (abgeschnitten) mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2). *

	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log	MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
MW	0.90	0.1205	0.0275	0.0190	0.0016	0.1007	0.0115	0.0155	0.0005	0.0873	0.0717	0.1100	0.0101	0.0169	0.0002
STD		0.0049	0.0052	0.0017	0.0008	0.0049	0.0030	0.0015	0.0002	0.0029	0.0041	0.0063	0.0020	0.0018	0.0001

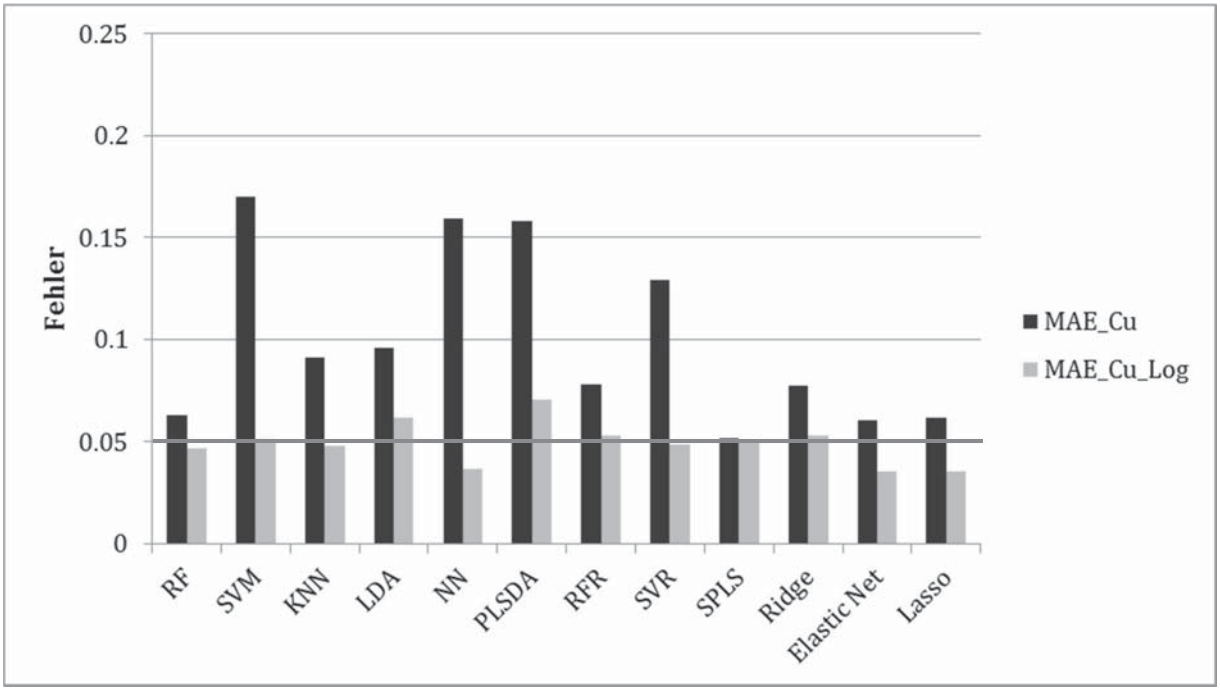
9.1.4 Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels realer Datensätze



Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.86	0.0642	0.0430	0.0066	0.0034	0.0472	0.0204
SVM	0.87	0.1946	0.0382	0.0548	0.0023	0.2557	0.0295
KNN	0.86	0.0944	0.0846	0.0142	0.0112	0.0853	0.0508
LDA	0.84	0.0734	0.0671	0.0068	0.0095	0.0511	0.0407
NN	0.90	0.1026	0.0754	0.0210	0.0117	0.0512	0.0329
PLSDA	0.84	0.0925	0.0527	0.0120	0.0054	0.0986	0.0339
RFR	0.87	0.0660	0.0636	0.0095	0.0067	0.0505	0.0303
SVR	0.86	0.1537	0.0412	0.0356	0.0038	0.2190	0.0280
SPLS	0.80	0.1203	0.0534	0.0205	0.0046	0.1140	0.0337
Ridge	0.85	0.1160	0.0659	0.0169	0.0096	0.1055	0.0448
Elastic Net	0.85	0.1009	0.0495	0.0130	0.0039	0.0965	0.0336
Lasso	0.85	0.1023	0.0478	0.0132	0.0038	0.0952	0.0293
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0046	0.0013	0.0988	0.0965	0.0618	0.0268	0.0054	0.0012
0.0853	0.0016	0.1885	0.1046	0.2726	0.0348	0.0950	0.0020
0.0125	0.0051	0.1192	0.1150	0.0824	0.0584	0.0111	0.0053
0.0042	0.0047	0.1120	0.1144	0.0509	0.0529	0.0039	0.0045
0.0083	0.0038	0.0914	0.0910	0.0426	0.0277	0.0032	0.0013
0.0130	0.0030	0.1234	0.1110	0.0970	0.0363	0.0127	0.0020
0.0062	0.0025	0.0989	0.0959	0.0590	0.0295	0.0053	0.0013
0.0599	0.0018	0.1628	0.1052	0.2250	0.0305	0.0644	0.0016

0.0179	0.0027	0.1446	0.1323	0.1142	0.0323	0.0179	0.0026
0.0158	0.0049	0.1191	0.1090	0.1129	0.0521	0.0175	0.0041
0.0126	0.0021	0.1154	0.1066	0.0997	0.0379	0.0137	0.0021
0.0124	0.0018	0.1152	0.1068	0.0982	0.0379	0.0137	0.0021

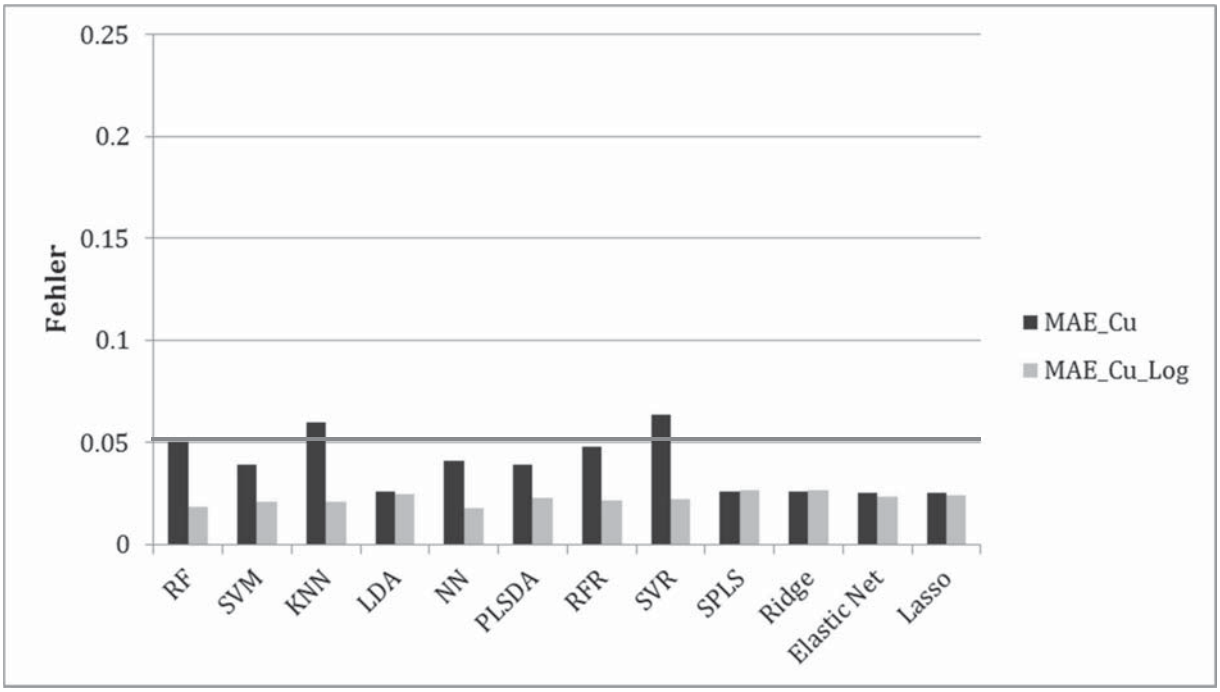
Abbildung 82: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem QSAR Datensatz (Dragon) mit 1055 Molekülen und einer mittleren Acc über alle Techniken von 0.85. Nur die Techniken LDA und NN bringen bereits vor der Kalibrierung gute Wahrscheinlichkeitsschätzer hervor (unter der roten Linie). Alle Techniken mit Ausnahme der LDA profitieren von der Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Die übrigen Fehlermaße sind in der Tabelle darunter aufgeführt.



Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.66	0.1546	/	0.0593	/	0.0727	/
SVM	0.61	0.2662	/	0.0957	/	0.1759	/
KNN	0.67	/	/	/	/	/	/
LDA	0.66	/	/	/	/	/	/
NN	0.63	0.2512	/	0.0855	/	0.1785	/
PLSDA	0.62	0.2925	/	0.1358	/	0.1631	/
RFR	0.64	/	/	/	/	/	/
SVR	0.61	0.2240	/	0.0763	/	0.1350	/
SPLS	0.70	/	/	/	/	/	/
Ridge	0.69	/	/	/	/	/	/
Elastic Net	0.70	/	/	/	/	/	/
Lasso	0.69	/	/	/	/	/	/
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**

0.0136	/	0.2204	0.2100	0.0635	0.0465	0.0084	0.0033
0.0470	/	0.2526	0.2141	0.1703	0.0512	0.0423	0.0037
/	/	0.2221	0.2128	0.0914	0.0479	0.0107	0.0031
/	/	0.2201	0.2132	0.0956	0.0618	0.0140	0.0051
0.0423	/	0.2518	0.2127	0.1592	0.0366	0.0335	0.0021
0.0460	/	0.2787	0.2133	0.1583	0.0704	0.0394	0.0067
/	/	0.2162	0.2094	0.0780	0.0532	0.0087	0.0042
0.0346	/	0.2411	0.2136	0.1291	0.0487	0.0294	0.0035
/	/	0.2132	0.2111	0.0522	0.0505	0.0042	0.0040
/	/	0.2096	0.2102	0.0774	0.0530	0.0087	0.0045
/	/	0.2089	0.2095	0.0606	0.0354	0.0059	0.0018
/	/	0.2092	0.2096	0.0620	0.0353	0.0059	0.0020

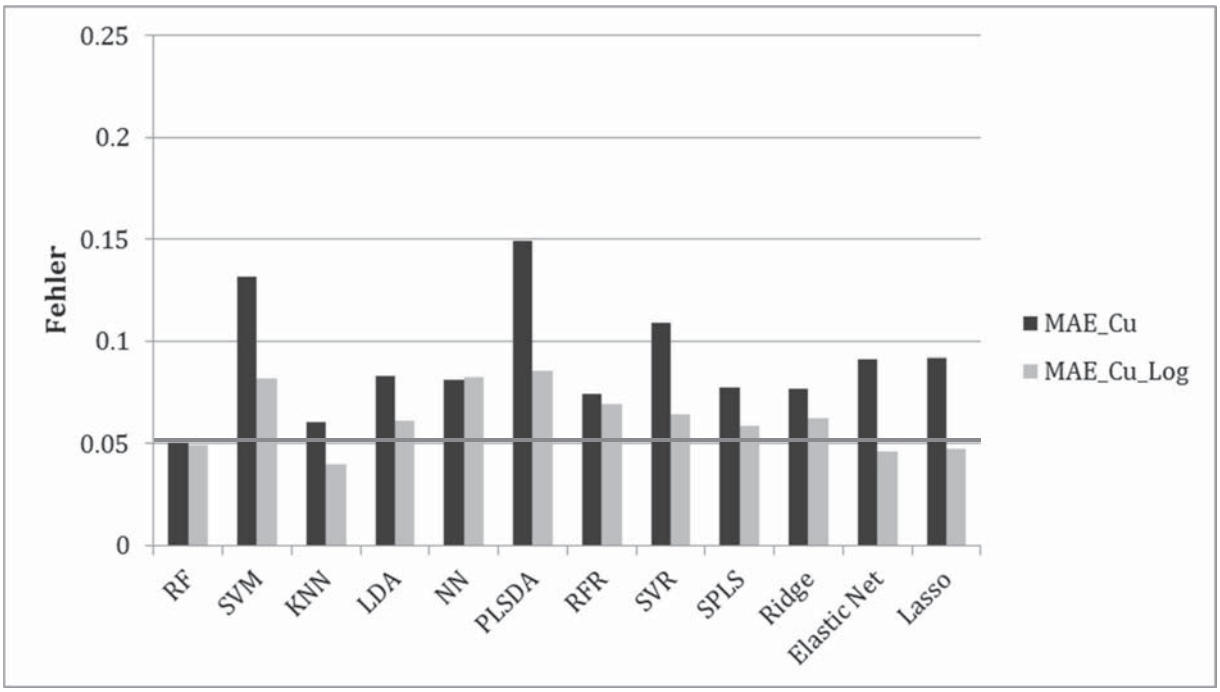
Abbildung 83: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem Liver Datensatz (MACCS) mit 951 Molekülen und einer mittleren Acc über alle Techniken von 0.66. Fast alle Techniken bringen vor der Kalibrierung schlechte Wahrscheinlichkeitsschätzer hervor (unter der roten Linie). Alle Techniken mit Ausnahme der SPLS profitieren von der Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Die übrigen Fehlermaße sind in der Tabelle darunter aufgeführt. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.



Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.58	0.0635	/	0.0063	/	0.0543	/
SVM	0.57	0.1429	/	0.0381	/	0.0492	/
KNN	0.56	0.1282	/	0.0238	/	0.07293227	/
LDA	0.55	/	/	/	/	/	/
NN	0.57	0.0899	/	0.0119	/	0.0483	/

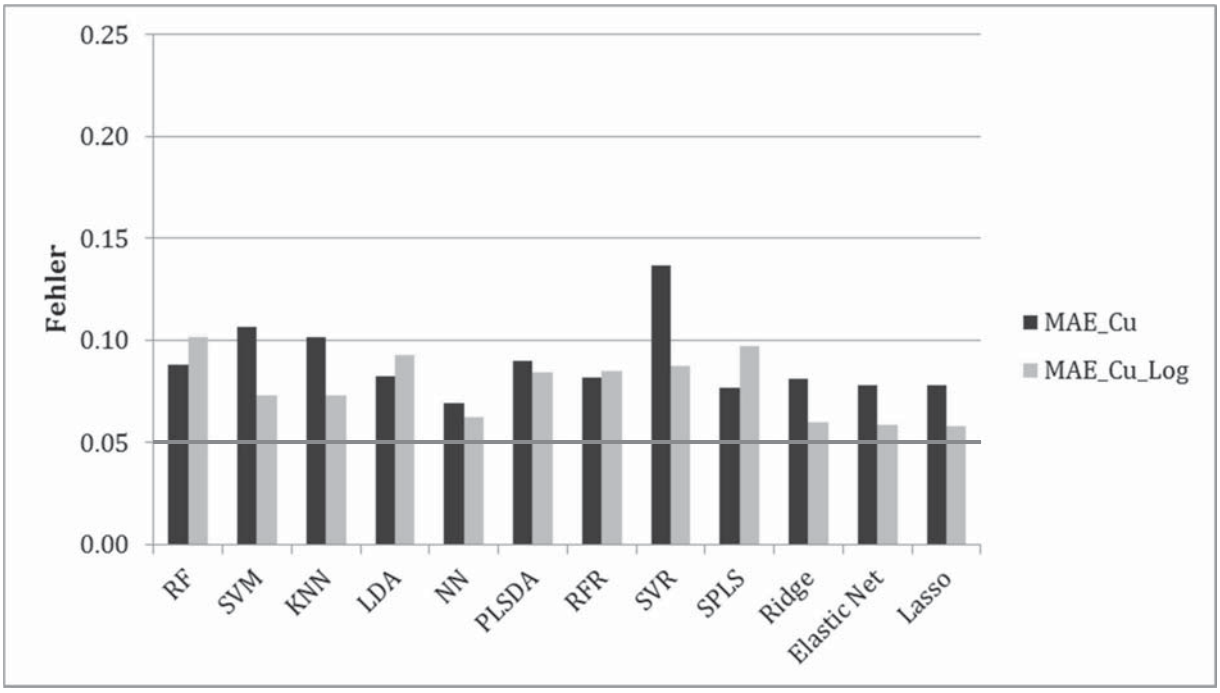
PLSDA	0.56	0.1668	/	0.0567	/	0.0458	/
RFR	0.58	0.1044	/	0.0171	/	0.0503	/
SVR	0.57	0.1936	/	0.0743	/	0.0734	/
SPLS	0.54	/	/	/	/	/	/
Ridge	0.56	/	/	/	/	/	/
Elastic Net	0.57	/	/	/	/	/	/
Lasso	0.57	/	/	/	/	/	/
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0043	/	0.2443	0.2393	0.0505	0.0183	0.0037	0.0005
0.0050	/	0.2487	0.2442	0.0395	0.0211	0.0025	0.0006
0.0075	/	0.2494	0.2427	0.0599	0.0210	0.0044	0.0006
/	/	0.2462	0.2452	0.0263	0.0247	0.0012	0.0008
0.0043	/	0.2473	0.2431	0.0409	0.0180	0.0028	0.0005
0.0053	/	0.2494	0.2442	0.0394	0.0231	0.0025	0.0009
0.0046	/	0.2427	0.2385	0.0482	0.0214	0.0032	0.0006
0.0097	/	0.2526	0.2438	0.0636	0.0225	0.0066	0.0008
/	/	0.2468	0.2465	0.0264	0.0269	0.0010	0.0012
/	/	0.2438	0.2436	0.0259	0.0267	0.0009	0.0010
/	/	0.2446	0.2441	0.0253	0.0238	0.0010	0.0010
/	/	0.2446	0.2440	0.0257	0.0240	0.0010	0.0011

Abbildung 84: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem Cancer Datensatz (MACCS) mit 7747 Objekten und einer mittleren Acc über alle Techniken von 0.57. Auffällig bei diesem Datensatz ist, dass fast alle Techniken bereits vor der Kalibrierung gute Wahrscheinlichkeitsschätzer hervorbringen. Dennoch profitieren alle von der Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Die übrigen Fehlermaße sind in der Tabelle darunter aufgeführt. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.



Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.76	0.0418	0.0493	0.0032	0.0039	0.0382	0.0399
SVM	0.75	0.1358	0.1178	0.0241	0.0216	0.1336	0.1032
KNN	0.74	0.1051	/	0.0179	/	0.0671	/
LDA	0.79	0.0972	/	0.0202	/	0.0714	/
NN	0.78	0.1133	/	0.0192	/	0.0950	/
PLSDA	0.73	0.1338	0.0682	0.0369	0.0068	0.1237	0.0714
RFR	0.74	0.0907	0.1025	0.0114	0.0176	0.0887	0.0846
SVR	0.75	0.1106	0.0842	0.0165	0.0140	0.1172	0.0605
SPLS	0.75	0.0799	0.0671	0.0104	0.0070	0.0779	0.0521
Ridge	0.74	0.0747	0.0620	0.0075	0.0066	0.0708	0.0534
Elastic Net	0.78	0.0721	0.0423	0.0082	0.0029	0.0790	0.0425
Lasso	0.78	0.0730	0.0523	0.0085	0.0040	0.0800	0.0495
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0030	0.0029	0.1605	0.1650	0.0506	0.0491	0.0037	0.0039
0.0241	0.0163	0.1898	0.1781	0.1318	0.0819	0.0232	0.0094
0.0090	/	0.1766	0.1783	0.0603	0.0401	0.0054	0.0023
0.0147	/	0.1496	0.1512	0.0836	0.0610	0.0125	0.0055
0.0153	/	0.1637	0.1655	0.0812	0.0824	0.0104	0.0094
0.0322	0.0081	0.2435	0.1979	0.1491	0.0857	0.0355	0.0116
0.0105	0.0144	0.1643	0.1675	0.0743	0.0697	0.0084	0.0084
0.0169	0.0083	0.1806	0.1732	0.1092	0.0646	0.0158	0.0074
0.0091	0.0043	0.1809	0.1839	0.0776	0.0585	0.0084	0.0056
0.0068	0.0055	0.1718	0.1755	0.0768	0.0623	0.0093	0.0066
0.0091	0.0027	0.1574	0.1559	0.0912	0.0463	0.0118	0.0028
0.0094	0.0035	0.1560	0.1543	0.0918	0.0473	0.0122	0.0031

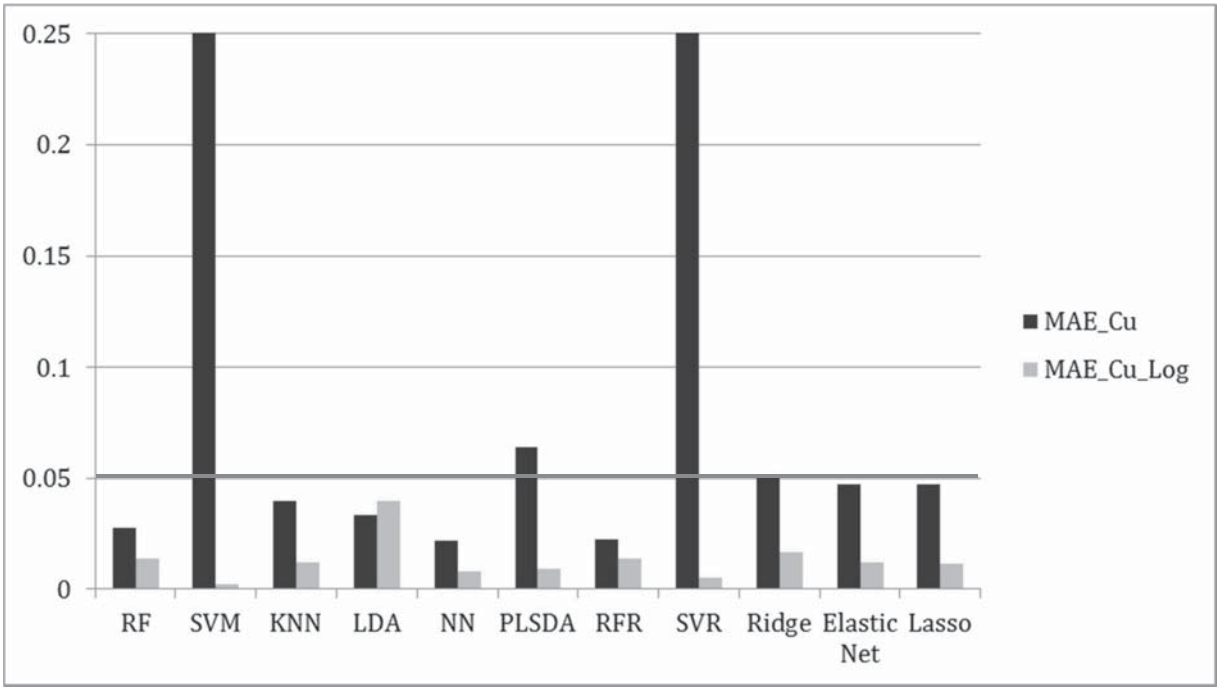
Abbildung 85: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem hERG Datensatz (Maccs) mit 561 Molekülen und einer mittleren Acc über alle Techniken von 0.76. Alle Techniken mit Ausnahme der NN profitieren von der Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Im Gegensatz zu den Ergebnissen aus den Simulationsstudien und den übrigen Datensätzen, sind die kalibrierten und unkalibrierten Fehler der LDA und der NN erhöht. Die übrigen Fehlermaße sind in der Tabelle darunter aufgeführt. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.



Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.80	0.0272	0.0187	0.0013	0.0006	0.0243	0.0182
SVM	0.77	0.1199	0.0319	0.0207	0.0018	0.1434	0.0304
KNN	0.78	0.0341	0.0269	0.0016	0.0012	0.0310	0.0191
LDA	0.72	0.0259	0.0193	0.0008	0.0007	0.0254	0.0207
NN	0.77	0.0365	0.0441	0.0020	0.0028	0.0333	0.0340
PLSDA	0.72	0.0354	0.0324	0.0020	0.0012	0.0337	0.0323
RFR	0.76	0.0293	0.0292	0.0013	0.0013	0.0288	0.0247
SVR	0.77	0.0920	0.0284	0.0112	0.0012	0.1146	0.0252
SPLS	0.72	0.0510	0.0274	0.0035	0.0014	0.0467	0.0318
Ridge	0.73	0.0527	0.0336	0.0031	0.0015	0.0513	0.0336
Elastic Net	0.73	0.0568	0.0290	0.0036	0.0012	0.0562	0.0298
Lasso	0.73	0.0554	0.0281	0.0035	0.0011	0.0550	0.0288
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0011	0.0005	0.1412	0.1405	0.0302	0.0166	0.0012	0.0005
0.0264	0.0016	0.1958	0.1686	0.1519	0.0320	0.0289	0.0016
0.0012	0.0007	0.1798	0.1797	0.0267	0.0193	0.0010	0.0005
0.0008	0.0008	0.1908	0.1920	0.0300	0.0281	0.0013	0.0012
0.0017	0.0020	0.1562	0.1575	0.0368	0.0343	0.0019	0.0018
0.0017	0.0012	0.1838	0.1776	0.0377	0.0323	0.0022	0.0013

0.0013	0.0010	0.1407	0.1404	0.0287	0.0231	0.0013	0.0008
0.0155	0.0009	0.1704	0.1552	0.1204	0.0237	0.0168	0.0010
0.0027	0.0015	0.2120	0.2112	0.0462	0.0328	0.0032	0.0019
0.0030	0.0015	0.1775	0.1759	0.0486	0.0289	0.0031	0.0012
0.0036	0.0013	0.1778	0.1762	0.0481	0.0300	0.0031	0.0013
0.0036	0.0011	0.1778	0.1762	0.0477	0.0296	0.0030	0.0012

Abbildung 86: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem BBB Datensatz (Chemical/Physical Properties) mit 325 Molekülen und einer mittleren Acc über alle Techniken von 0.75. Alle Techniken mit Ausnahme von RF, LDA und SPLS profitieren von der Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Im Gegensatz zu den Ergebnissen aus den Simulationsstudien und den übrigen Datensätzen, sind die kalibrierten und unkalibrierten Fehler aller Techniken erhöht. In den Simulationsstudien stieg der Fehler mit abnehmender Datensatzgröße an. Dies kann auch hier der Grund sein, denn der Datensatz besteht lediglich aus knapp über 300 Objekten. Die übrigen Fehlermaße sind in der Tabelle darunter aufgeführt.



Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.97	0.1047	0.0743	0.0169	0.0126	0.0210	0.0072
SVM	0.99	0.2548	0.1139	0.0874	0.0245	0.4203	0.0033
KNN	0.95	0.0727	NA	0.0073	NA	0.0377	NA
LDA	0.91	0.0712	0.0552	0.0086	0.0059	0.0312	0.0102
NN	0.98	0.1040	0.1050	0.0160	0.0203	0.0252	0.0048
PLSDA	0.94	0.1175	0.0425	0.0226	0.0026	0.0720	0.0102
RFR	0.97	0.0871	0.0750	0.0123	0.0111	0.0258	0.0069
SVR	0.99	0.2048	0.1422	0.0550	0.0359	0.3290	0.0043
Ridge	0.93	0.1101	0.0440	0.0190	0.0036	0.0545	0.0108
Elastic Net	0.94	0.1185	0.0628	0.0247	0.0063	0.0493	0.0121
Lasso	0.94	0.1209	0.0665	0.0251	0.0077	0.0501	0.0120

MSE_w	MSE_w_Log	Brier	Brier_Log	MAE_Cu	MAE_Cu_Log	MSE_Cu	MSE_Cu_Log
0.0028	0.0006	0.0286	0.0261	0.0275	0.0139	0.0018	0.0004
0.1936	0.0005	0.2066	0.0100	0.4329	0.0025	0.2036	0.0000
0.0029	NA	0.0411	0.0385	0.0397	0.0119	0.0030	0.0003
0.0017	0.0008	0.0689	0.0693	0.0331	0.0396	0.0022	0.0020
0.0018	0.0004	0.0148	0.0135	0.0219	0.0078	0.0011	0.0002
0.0082	0.0005	0.0573	0.0473	0.0642	0.0094	0.0055	0.0002
0.0021	0.0007	0.0277	0.0262	0.0227	0.0136	0.0013	0.0004
0.1127	0.0006	0.1228	0.0094	0.3275	0.0050	0.1139	0.0001
0.0059	0.0007	0.0557	0.0501	0.0510	0.0165	0.0046	0.0005
0.0056	0.0008	0.0517	0.0465	0.0473	0.0118	0.0040	0.0002
0.0058	0.0009	0.0516	0.0463	0.0474	0.0118	0.0040	0.0002

Abbildung 87: MAE_Cu und MAE_Cu_Log der auf der x-Achse aufgeführten Techniken basierend auf dem Musk2 Datensatz (Shape/Conformation) mit 6598 Molekülen und einer mittleren Acc über alle Techniken von 0.96. Alle Techniken mit Ausnahme der LDA profitieren von der Rekalibrierung der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer. Die Techniken RF, KNN, LDA, NN, RFR, sowie Ridge, Elastic Net und Lasso, bringen bereits unkalibriert gute Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer hervor. Im Gegensatz zu den Ergebnissen aus den Simulationsstudien und den übrigen Datensätzen, sind die unkalibrierten Fehler des RF und des RFR etwas niedriger. Die übrigen Fehlermaße sind in der Tabelle darunter aufgeführt.

Tabelle 551: Auflistung aller berechneten Fehlermaße auf Grundlage des Ames Datensatzes (MACCS) mit 6512 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.80	0.0272	0.0187	0.0013	0.0006	0.0243	0.0182
SVM	0.76	0.1199	0.0319	0.0207	0.0018	0.1434	0.0304
KNN	0.74	0.0341	0.0269	0.0016	0.0012	0.0310	0.0191
LDA	0.71	0.0259	0.0193	0.0008	0.0007	0.0254	0.0207
NN	0.78	0.0365	0.0441	0.0020	0.0028	0.0333	0.0340
PLSDA	0.74	0.0354	0.0324	0.0020	0.0012	0.0337	0.0323
RFR	0.80	0.0293	0.0292	0.0013	0.0013	0.0288	0.0247
SVR	0.78	0.0920	0.0284	0.0112	0.0012	0.1146	0.0252
SPLS	0.68	0.0510	0.0274	0.0035	0.0014	0.0467	0.0318
Ridge	0.74	0.0527	0.0336	0.0031	0.0015	0.0513	0.0336
Elastic Net	0.74	0.0568	0.0290	0.0036	0.0012	0.0562	0.0298
Lasso	0.74	0.0554	0.0281	0.0035	0.0011	0.0550	0.0288
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0011	0.0005	0.1412	0.1405	0.0302	0.0166	0.0012	0.0005
0.0264	0.0016	0.1958	0.1686	0.1519	0.0320	0.0289	0.0016
0.0012	0.0007	0.1798	0.1797	0.0267	0.0193	0.0010	0.0005
0.0008	0.0008	0.1908	0.1920	0.0300	0.0281	0.0013	0.0012
0.0017	0.0020	0.1562	0.1575	0.0368	0.0343	0.0019	0.0018
0.0017	0.0012	0.1838	0.1776	0.0377	0.0323	0.0022	0.0013

0.0013	0.0010	0.1407	0.1404	0.0287	0.0231	0.0013	0.0008
0.0155	0.0009	0.1704	0.1552	0.1204	0.0237	0.0168	0.0010
0.0027	0.0015	0.2120	0.2112	0.0462	0.0328	0.0032	0.0019
0.0030	0.0015	0.1775	0.1759	0.0486	0.0289	0.0031	0.0012
0.0036	0.0013	0.1778	0.1762	0.0481	0.0300	0.0031	0.0013
0.0036	0.0011	0.1778	0.1762	0.0477	0.0296	0.0030	0.0012

Tabelle 552: Auflistung aller berechneten Fehlermaße auf Grundlage des CYP1A2 Datensatzes (MACCS) mit 7485 Molekülen.

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.81	0.0260	0.0280	0.0013	0.0010	0.0217	0.0249
SVM	0.79	0.1410	0.0269	0.0232	0.0010	0.1561	0.0259
KNN	0.78	0.0458	0.0365	0.0028	0.0020	0.0399	0.0301
LDA	0.78	0.0171	0.0203	0.0004	0.0006	0.0163	0.0181
NN	0.81	0.0492	0.0442	0.0031	0.0031	0.0438	0.0283
PLSDA	0.79	0.0449	0.0270	0.0026	0.0014	0.0436	0.0257
RFR	0.81	0.0313	0.0349	0.0015	0.0016	0.0281	0.0297
SVR	0.80	0.0948	0.0442	0.0114	0.0025	0.1208	0.0374
SPLS	0.73	0.0561	0.0385	0.0036	0.0021	0.0556	0.0395
Ridge	0.79	0.0467	0.0342	0.0032	0.0020	0.0447	0.0305
Elastic Net	0.79	0.0455	0.0351	0.0028	0.0022	0.0435	0.0314
Lasso	0.79	0.0457	0.0331	0.0029	0.0019	0.0436	0.0300
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0011	0.0008	0.1333	0.1334	0.0266	0.0228	0.0012	0.0008
0.0271	0.0010	0.1756	0.1487	0.1562	0.0283	0.0290	0.0011
0.0021	0.0014	0.1511	0.1510	0.0392	0.0323	0.0020	0.0015
0.0004	0.0005	0.1572	0.1578	0.0130	0.0188	0.0003	0.0005
0.0026	0.0018	0.1390	0.1404	0.0425	0.0383	0.0025	0.0025
0.0026	0.0012	0.1602	0.1525	0.0411	0.0298	0.0025	0.0011
0.0012	0.0012	0.1321	0.1323	0.0289	0.0293	0.0013	0.0012
0.0169	0.0018	0.1587	0.1438	0.1214	0.0372	0.0174	0.0018
0.0036	0.0023	0.1794	0.1777	0.0495	0.0358	0.0038	0.0020
0.0030	0.0016	0.1514	0.1502	0.0437	0.0269	0.0026	0.0011
0.0026	0.0017	0.1519	0.1507	0.0443	0.0288	0.0025	0.0012
0.0027	0.0016	0.1519	0.1508	0.0444	0.0288	0.0025	0.0012

Tabelle 553: Auflistung aller berechneten Fehlermaße auf Grundlage des CYP1A2 Datensatzes (MOE) mit 7485 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.81	0.0304	0.0264	0.0014	0.0013	0.0263	0.0205
SVM	0.79	0.1465	0.0357	0.0239	0.0016	0.1602	0.0338
KNN	0.78	0.0421	0.0368	0.0028	0.0025	0.0389	0.0298

LDA	0.75	0.0201	0.0169	0.0007	0.0006	0.0191	0.0169
NN	0.81	0.0370	0.0430	0.0026	0.0041	0.0318	0.0327
PLSDA	0.81	0.0590	0.0338	0.0045	0.0021	0.0648	0.0320
RFR	0.81	0.0253	0.0275	0.0011	0.0014	0.0219	0.0216
SVR	0.81	0.2467	0.0302	0.1294	0.0020	0.1637	0.0226
SPLS	0.71	0.0671	0.0167	0.0061	0.0004	0.0513	0.0139
Ridge	0.80	0.0664	0.0311	0.0057	0.0013	0.0679	0.0268
Elastic Net	0.81	0.0695	0.0316	0.0060	0.0016	0.0705	0.0251
Lasso	0.81	0.0693	0.0276	0.0054	0.0015	0.0688	0.0228
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0012	0.0009	0.1317	0.1317	0.0321	0.0219	0.0014	0.0007
0.0280	0.0015	0.1761	0.1486	0.1629	0.0335	0.0300	0.0014
0.0024	0.0017	0.1465	0.1471	0.0352	0.0362	0.0017	0.0018
0.0006	0.0006	0.1662	0.1667	0.0172	0.0214	0.0004	0.0007
0.0019	0.0026	0.1343	0.1365	0.0331	0.0355	0.0020	0.0024
0.0052	0.0020	0.1677	0.1439	0.0643	0.0326	0.0054	0.0017
0.0009	0.0009	0.1321	0.1327	0.0245	0.0223	0.0009	0.0009
0.0306	0.0011	0.1697	0.1397	0.1628	0.0310	0.0314	0.0014
0.0041	0.0003	0.1879	0.1847	0.0499	0.0179	0.0040	0.0005
0.0060	0.0011	0.1512	0.1464	0.0715	0.0272	0.0064	0.0011
0.0061	0.0013	0.1482	0.1437	0.0698	0.0257	0.0061	0.0010
0.0054	0.0012	0.1472	0.1431	0.0683	0.0251	0.0058	0.0010

Tabelle 554: Auflistung aller berechneten Fehlermaße auf Grundlage des CYP1A2 Datensatzes (E-State) mit 7485 Molekülen.

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.81	0.0279	0.0328	0.0013	0.0016	0.0219	0.0263
SVM	0.79	0.1238	0.0287	0.0196	0.0011	0.1564	0.0266
KNN	0.77	0.0331	0.0337	0.0017	0.0015	0.0289	0.0304
LDA	0.76	0.0227	0.0372	0.0010	0.0021	0.0207	0.0328
NN	0.81	0.0635	0.0462	0.0057	0.0044	0.0546	0.0316
PLSDA	0.79	0.0340	0.0313	0.0016	0.0014	0.0337	0.0315
RFR	0.80	0.0267	0.0316	0.0010	0.0014	0.0250	0.0272
SVR	0.80	0.1209	0.0317	0.0204	0.0013	0.1401	0.0256
SPLS	0.75	0.0705	0.0355	0.0062	0.0014	0.0695	0.0363
Ridge	0.79	0.0461	0.0222	0.0026	0.0007	0.0459	0.0205
Elastic Net	0.79	0.0459	0.0225	0.0026	0.0007	0.0444	0.0224
Lasso	0.79	0.0464	0.0249	0.0026	0.0008	0.0452	0.0237
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0011	0.0011	0.1321	0.1327	0.0274	0.0232	0.0012	0.0009
0.0276	0.0010	0.1788	0.1510	0.1571	0.0257	0.0297	0.0009
0.0013	0.0012	0.1565	0.1567	0.0268	0.0286	0.0010	0.0011
0.0009	0.0015	0.1615	0.1621	0.0207	0.0331	0.0007	0.0014

0.0042	0.0024	0.1403	0.1409	0.0502	0.0440	0.0039	0.0030
0.0016	0.0013	0.1601	0.1509	0.0366	0.0263	0.0019	0.0009
0.0009	0.0010	0.1342	0.1352	0.0218	0.0251	0.0007	0.0010
0.0266	0.0009	0.1690	0.1436	0.1434	0.0290	0.0274	0.0012
0.0062	0.0015	0.1705	0.1662	0.0680	0.0332	0.0063	0.0015
0.0025	0.0006	0.1509	0.1491	0.0430	0.0168	0.0023	0.0004
0.0025	0.0007	0.1505	0.1490	0.0397	0.0159	0.0020	0.0004
0.0025	0.0007	0.1504	0.1489	0.0390	0.0161	0.0019	0.0004

Tabelle 555: Auflistung aller berechneten Fehlermaße auf Grundlage des Factor Xa Datensatzes (Maccs) mit 435 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.95	0.1503	/	0.0386	/	0.0603	/
SVM	0.92	/	0.1738	/	0.0464	/	0.0553
KNN	0.90	0.1238	0.1749	0.0261	0.0675	0.0578	0.0359
LDA	0.95	/	/	/	/	/	/
NN	0.92	0.1636	0.1943	0.0377	0.0657	0.0485	0.0307
PLSDA	0.89	0.0947	0.0999	0.0152	0.0326	0.0761	0.0286
RFR	0.94	0.1256	/	0.0241	/	0.0624	/
SVR	0.94	0.1369	/	0.0218	/	0.1261	/
SPLS	0.82	0.1201	0.1160	0.0236	0.0316	0.0938	0.0492
Ridge	0.94	0.1119	/	0.0202	/	0.0771	/
Elastic Net	0.94	0.1361	0.2094	0.0277	0.1043	0.0866	0.0371
Lasso	0.94	0.1495	0.2503	0.0353	0.1160	0.0852	0.0413
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0117	/	0.0507	0.0429	0.0661	0.0216	0.0084	0.0008
/	0.0103	0.1193	0.0679	0.2018	0.0536	0.0561	0.0045
0.0091	0.0087	0.0687	0.0684	0.0431	0.0322	0.0034	0.0018
NA	/	0.0466	0.0461	0.0364	0.0532	0.0020	0.0035
0.0079	0.0067	0.0587	0.0623	0.0463	0.0402	0.0032	0.0033
0.0090	0.0027	0.1041	0.0716	0.0854	0.0444	0.0135	0.0026
0.0100	/	0.0533	0.0450	0.0682	0.0244	0.0091	0.0010
0.0184	/	0.0685	0.0523	0.1299	0.0361	0.0202	0.0017
0.0157	0.0097	0.1220	0.1228	0.0846	0.0565	0.0118	0.0056
0.0128	/	0.0555	0.0448	0.0893	0.0255	0.0120	0.0011
0.0163	0.0083	0.0575	0.0458	0.0916	0.0368	0.0148	0.0024
0.0174	0.0111	0.0572	0.0462	0.0876	0.0381	0.0139	0.0025

Tabelle 556: Auflistung aller berechneten Fehlermaße auf Grundlage des Factor Xa Datensatzes (MOE) mit 435 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.90	0.1541	0.1091	0.0336	0.0221	0.0742	0.0475
SVM	0.92	0.1490	0.1194	0.0294	0.0336	0.1566	0.0343

KNN	0.88	0.1492	0.2026	0.0587	0.0679	0.0470	0.0639
LDA	0.87	0.1820	0.1678	0.0497	0.0467	0.0780	0.0596
NN	0.92	0.1471	0.1506	0.0368	0.0350	0.0458	0.0295
PLSDA	0.92	0.0928	0.0382	0.0128	0.0021	0.0862	0.0162
RFR	0.89	0.1424	0.1140	0.0282	0.0505	0.0650	0.0408
SVR	0.93	0.1636	/	0.0320	/	0.1597	/
SPLS	0.81	0.1346	0.0633	0.0266	0.0069	0.1119	0.0696
Ridge	0.93	0.1269	0.1278	0.0236	0.0309	0.0762	0.0433
Elastic Net	0.92	0.0942	0.1834	0.0141	0.0850	0.0639	0.0473
Lasso	0.92	0.1030	0.2059	0.0149	0.1055	0.0642	0.0362
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0154	0.0065	0.0763	0.0690	0.0728	0.0440	0.0114	0.0028
0.0303	0.0060	0.0847	0.0583	0.1676	0.0374	0.0330	0.0026
0.0116	0.0185	0.0881	0.0939	0.0432	0.0523	0.0032	0.0048
0.0101	0.0076	0.1030	0.1033	0.0774	0.0687	0.0092	0.0065
0.0067	0.0060	0.0665	0.0670	0.0391	0.0276	0.0028	0.0012
0.0107	0.0007	0.0804	0.0645	0.0980	0.0311	0.0132	0.0015
0.0099	0.0050	0.0770	0.0745	0.0586	0.0490	0.0073	0.0038
0.0303	NA	0.0807	0.0613	0.1586	0.0590	0.0287	0.0070
0.0183	0.0075	0.1504	0.1401	0.1113	0.0683	0.0189	0.0071
0.0117	0.0066	0.0727	0.0673	0.0821	0.0297	0.0100	0.0016
0.0087	0.0111	0.0671	0.0639	0.0711	0.0371	0.0091	0.0033
0.0087	0.0089	0.0678	0.0644	0.0695	0.0367	0.0088	0.0026

9.1.5 Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels Simulationsstudien

Tabelle 557: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.7 (mu1=0, mu2=0.195).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.72	/	0.0204	/	0.0006	/	0.0199
KNN	0.67	0.0464	0.0244	0.0034	0.0010	0.0310	0.0185
SVM	0.67	0.1663	0.0357	0.1062	0.0041	0.0592	0.0148
PLSDA	0.73	0.0455	0.0204	0.0026	0.0007	0.0414	0.0200
RFR	0.71	/	0.0206	/	0.0005	/	0.0202
SVR	0.67	0.1597	0.0446	0.1056	0.0049	0.0539	0.0236
SPLS	0.74	0.0374	0.0176	0.0021	0.0006	0.0369	0.0171
Ridge	0.73	0.0559	0.0230	0.0040	0.0007	0.0524	0.0228
Elastic Net	0.73	0.0468	0.0238	0.0028	0.0007	0.0437	0.0235
Lasso	0.73	0.0463	0.0236	0.0028	0.0007	0.0431	0.0234
LDA PC:30	0.73	0.0299	0.0382	0.0014	0.0019	0.0303	0.0361

NN	0.72	0.0485	/	0.0032	/	0.0478	/
Ensemble A	0.73	0.0733	0.0284	0.0068	0.0013	0.0717	0.0297
Ensemble B	0.73	/	0.0323	/	0.0020	/	0.0339
Ensemble C	0.73	/	0.0295	/	0.0018	/	0.0316
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
NA	0.0006	0.2075	0.1884	0.1223	0.0245	0.0192	0.0009
0.0018	0.0006	0.2117	0.2117	0.0271	0.0241	0.0013	0.0009
0.0052	0.0007	0.2134	0.2108	0.0560	0.0247	0.0044	0.0010
0.0024	0.0006	0.1794	0.1768	0.0390	0.0292	0.0022	0.0011
NA	0.0005	0.2194	0.1904	0.1427	0.0159	0.0284	0.0004
0.0049	0.0010	0.2142	0.2120	0.0560	0.0279	0.0043	0.0011
0.0020	0.0005	0.1814	0.1816	0.0301	0.0180	0.0015	0.0005
0.0037	0.0006	0.1783	0.1765	0.0475	0.0281	0.0034	0.0011
0.0026	0.0007	0.1777	0.1767	0.0392	0.0283	0.0023	0.0011
0.0025	0.0007	0.1776	0.1767	0.0387	0.0283	0.0022	0.0010
0.0015	0.0017	0.1751	0.1767	0.0288	0.0339	0.0012	0.0019
0.0031	NA	0.1838	0.1837	0.0407	0.0313	0.0024	0.0016
0.0067	0.0014	0.1825	0.1776	0.0752	0.0260	0.0075	0.0011
/	0.0023	/	0.1798	/	0.0400	/	0.0026
/	0.0020	/	0.1794	/	0.0395	/	0.0026

Tabelle 558 Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.8 (mu1=0, mu2=0.29).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE-w_Log**
RF	0.80	0.2434	0.0306	0.1147	0.0014	0.1754	0.0243
KNN	0.77	0.0751	/	0.0075	/	0.0807	/
SVM	0.77	0.1306	0.0198	0.0221	0.0006	0.1403	0.0198
PLSDA	0.82	0.0695	0.0240	0.0068	0.0008	0.0706	0.0201
RFR	0.80	/	0.0210	/	0.0007	/	0.0181
SVR	0.77	0.1165	0.0135	0.0182	0.0005	0.1239	0.0127
SPLS	0.82	0.0653	0.0246	0.0063	0.0014	0.0581	0.0185
Ridge	0.82	0.0754	0.0225	0.0080	0.0007	0.0690	0.0177
Elastic Net	0.82	0.0713	0.0163	0.0070	0.0004	0.0638	0.0133
Lasso	0.82	0.0713	0.0168	0.0070	0.0004	0.0638	0.0138
LDA PC:30	0.82	0.0125	0.0409	0.0003	0.0020	0.0099	0.0301
NN	0.82	0.0347	0.0432	0.0018	0.0031	0.0354	0.0339
Ensemble A	0.82	0.0958	0.0295	0.0121	0.0013	0.1020	0.0210
Ensemble B	0.82	/	0.0412	/	0.0021	/	0.0379
Ensemble C	0.82	/	0.0334	/	0.0016	/	0.0301
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0354	0.0011	0.1755	0.1378	0.1763	0.0192	0.0387	0.0007
0.0084	NA	0.1656	0.1580	0.0871	0.0193	0.0089	0.0006
0.0249	0.0006	0.1779	0.1538	0.1390	0.0214	0.0259	0.0007

0.0069	0.0006	0.1354	0.1239	0.0745	0.0190	0.0071	0.0005
/	0.0006	0.1799	0.1396	0.1828	0.0206	0.0411	0.0006
0.0199	0.0004	0.1771	0.1580	0.1265	0.0172	0.0210	0.0004
0.0053	0.0008	0.1303	0.1256	0.0630	0.0185	0.0057	0.0005
0.0072	0.0005	0.1306	0.1237	0.0763	0.0172	0.0079	0.0004
0.0060	0.0003	0.1298	0.1239	0.0701	0.0168	0.0067	0.0004
0.0060	0.0003	0.1298	0.1239	0.0702	0.0168	0.0067	0.0004
0.0002	0.0012	0.1231	0.1244	0.0144	0.0277	0.0003	0.0011
0.0017	0.0020	0.1294	0.1302	0.0329	0.0372	0.0016	0.0019
0.0132	0.0009	0.1362	0.1237	0.1095	0.0188	0.0141	0.0006
/	0.0019	/	0.1260	/	0.0339	/	0.0017
/	0.0014	/	0.1259	/	0.0332	/	0.0017

Tabelle 559 Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0 und Acc=0.9 (mu1=0, mu2=0.42).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE-w_Log**
RF	0.89	0.1786	0.0198	0.0392	0.0007	0.2215	0.0093
KNN	0.88	0.1097	/	0.0157	/	0.1092	/
SVM	0.89	0.1869	0.0214	0.0424	0.0008	0.2319	0.0117
PLSDA	0.91	0.1274	0.0260	0.0208	0.0011	0.1269	0.0109
RFR	0.89	/	0.0235	/	0.0008	/	0.0120
SVR	0.88	0.1765	0.0301	0.0376	0.0016	0.2142	0.0172
SPLS	0.91	0.1177	0.0397	0.0188	0.0025	0.0895	0.0180
Ridge	0.91	0.1239	0.0262	0.0220	0.0009	0.0986	0.0121
Elastic Net	0.91	0.1220	0.0238	0.0207	0.0014	0.0946	0.0105
Lasso	0.91	0.1220	0.0239	0.0207	0.0014	0.0947	0.0105
LDA	0.91	0.0349	0.0552	0.0025	0.0052	0.0175	0.0194
NN	0.91	0.0306	0.0494	0.0017	0.0036	0.0199	0.0177
Ensemble A	0.91	0.1376	0.0439	0.0226	0.0047	0.1360	0.0143
Ensemble B	0.91	/	0.0440	/	0.0029	/	0.0229
Ensemble C	0.91	/	0.0479	/	0.0032	/	0.0244
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0531	0.0003	0.1333	0.0786	0.2279	0.0121	0.0577	0.0002
0.0160	/	0.1030	0.0872	0.1197	0.0112	0.0174	0.0002
0.0584	0.0003	0.1396	0.0797	0.2385	0.0171	0.0631	0.0005
0.0206	0.0004	0.0915	0.0664	0.1175	0.0159	0.0180	0.0004
NA	0.0004	0.1359	0.0799	0.2284	0.0132	0.0587	0.0003
0.0507	0.0008	0.1379	0.0860	0.2202	0.0209	0.0556	0.0007
0.0139	0.0009	0.0812	0.0669	0.0996	0.0145	0.0153	0.0004
0.0169	0.0004	0.0830	0.0662	0.1105	0.0143	0.0179	0.0004
0.0155	0.0004	0.0821	0.0663	0.1056	0.0146	0.0167	0.0004
0.0156	0.0004	0.0821	0.0663	0.1056	0.0146	0.0167	0.0004
0.0009	0.0016	0.0659	0.0670	0.0138	0.0230	0.0003	0.0010

0.0006	0.0011	0.0704	0.0718	0.0146	0.0245	0.0003	0.0010
0.0220	0.0010	0.0870	0.0655	0.1414	0.0112	0.0232	0.0003
/	0.0012	/	0.0669	/	0.0243	/	0.0011
/	0.0015	/	0.0670	/	0.0251	/	0.0013

Tabelle 560: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.7 (mu1=0, mu2=0.45).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.72	0.0520	0.0192	0.0035	0.0006	0.0438	0.0168
KNN	0.71	0.0325	0.0312	0.0027	0.0014	0.0255	0.0276
SVM	0.67	0.0722	0.0229	0.0105	0.0008	0.0379	0.0159
PLSDA	0.72	0.0364	0.0251	0.0019	0.0008	0.0299	0.0240
RFR	0.72	0.0464	0.0286	0.0027	0.0014	0.0420	0.0237
SVR	0.67	0.0639	0.0406	0.0066	0.0036	0.0391	0.0175
SPLS	0.73	0.0341	0.0230	0.0017	0.0008	0.0337	0.0215
Ridge	0.73	0.0306	0.0252	0.0013	0.0010	0.0348	0.0230
Elastic Net	0.72	0.0290	0.0235	0.0012	0.0009	0.0266	0.0207
Lasso	0.72	0.0281	0.0229	0.0012	0.0008	0.0258	0.0204
LDA PC:30	0.73	0.0260	0.0267	0.0009	0.0014	0.0257	0.0175
NN	0.73	0.0716	/	0.0061	/	0.0739	/
Ensemble A	0.72	0.0271	0.0219	0.0014	0.0012	0.0269	0.0158
Ensemble B	0.72	/	0.0300	/	0.0014	/	0.0229
Ensemble C	0.72	/	0.0255	/	0.0010	/	0.0189
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0027	0.0004	0.1876	0.1875	0.0458	0.0200	0.0028	0.0007
0.0019	0.0013	0.1919	0.1923	0.0179	0.0235	0.0007	0.0009
0.0020	0.0004	0.2120	0.2117	0.0390	0.0200	0.0021	0.0006
0.0013	0.0007	0.1864	0.1854	0.0295	0.0210	0.0013	0.0007
0.0023	0.0009	0.1865	0.1863	0.0429	0.0197	0.0026	0.0006
0.0021	0.0006	0.2126	0.2122	0.0407	0.0215	0.0025	0.0007
0.0016	0.0007	0.1807	0.1813	0.0384	0.0286	0.0018	0.0011
0.0015	0.0008	0.1829	0.1832	0.0392	0.0211	0.0021	0.0007
0.0011	0.0007	0.1845	0.1853	0.0295	0.0194	0.0013	0.0006
0.0010	0.0006	0.1846	0.1853	0.0293	0.0191	0.0013	0.0006
0.0009	0.0005	0.1834	0.1851	0.0264	0.0205	0.0010	0.0006
0.0064	/	0.1952	0.1918	0.0636	0.0256	0.0053	0.0009
0.0014	0.0006	0.1839	0.1851	0.0301	0.0168	0.0013	0.0005
/	0.0007	/	0.1853	/	0.0227	/	0.0008
/	0.0006	/	0.1847	/	0.0193	/	0.0006

Tabelle 561: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.8 (mu1=0, mu2=0.65).

Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.81	0.0741	0.0173	0.0069	0.0004	0.0773	0.0169
KNN	0.81	0.0446	0.0286	0.0028	0.0014	0.0368	0.0242
SVM	0.77	0.1066	0.0237	0.0129	0.0008	0.1183	0.0246
PLSDA	0.81	0.0588	0.0348	0.0050	0.0024	0.0647	0.0294
RFR	0.81	0.0706	0.0182	0.0057	0.0005	0.0720	0.0174
SVR	0.77	0.0938	0.0278	0.0109	0.0011	0.1035	0.0266
SPLS	0.82	0.0602	0.0270	0.0055	0.0010	0.0578	0.0235
Ridge	0.82	0.0685	0.0357	0.0068	0.0020	0.0692	0.0298
Elastic Net	0.82	0.0611	0.0344	0.0056	0.0025	0.0592	0.0281
Lasso	0.81	0.0567	0.0368	0.0051	0.0023	0.0543	0.0307
LDA PC:30	0.82	0.0238	0.0253	0.0008	0.0011	0.0233	0.0237
NN	0.82	0.0453	0.0371	0.0024	0.0023	0.0474	0.0320
Ensemble A	0.82	0.0644	0.0275	0.0046	0.0013	0.0648	0.0212
Ensemble B	0.82	/	0.0268	/	0.0011	/	0.0233
Ensemble C	0.82	/	0.0225	/	0.0008	/	0.0193
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0074	0.0004	0.1406	0.1357	0.0814	0.0205	0.0077	0.0006
0.0019	0.0010	0.1378	0.1386	0.0296	0.0189	0.0012	0.0005
0.0158	0.0008	0.1757	0.1602	0.1224	0.0227	0.0177	0.0008
0.0056	0.0016	0.1422	0.1356	0.0671	0.0320	0.0062	0.0015
0.0058	0.0005	0.1396	0.1361	0.0737	0.0186	0.0063	0.0005
0.0129	0.0011	0.1727	0.1610	0.1133	0.0288	0.0147	0.0012
0.0053	0.0008	0.1343	0.1315	0.0636	0.0228	0.0056	0.0007
0.0068	0.0015	0.1377	0.1334	0.0779	0.0298	0.0077	0.0012
0.0055	0.0017	0.1383	0.1353	0.0698	0.0306	0.0066	0.0014
0.0049	0.0016	0.1379	0.1355	0.0646	0.0304	0.0060	0.0014
0.0007	0.0010	0.1325	0.1352	0.0232	0.0290	0.0009	0.0013
0.0025	0.0016	0.1393	0.1406	0.0447	0.0291	0.0024	0.0013
0.0047	0.0008	0.1364	0.1342	0.0685	0.0220	0.0054	0.0007
/	0.0009	/	0.1346	/	0.0237	/	0.0009
/	0.0006	/	0.1341	/	0.0222	/	0.0008

Tabelle 562: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.1 und Acc=0.9 (mu1=0, mu2=0.9).

Technik	Acc	MAE	MAE_Log	MSE	MSE_Log	MAE_w	MAE_w_Log
RF	0.89	0.1046	0.0209	0.0122	0.0009	0.0997	0.0142
KNN	0.89	0.0437	0.0349	0.0029	0.0022	0.0295	0.0166
SVM	0.87	0.1583	0.0267	0.0292	0.0019	0.1932	0.0160
PLSDA	0.89	0.1044	0.0351	0.0141	0.0018	0.1114	0.0225
RFR	0.89	0.0950	0.0267	0.0110	0.0016	0.0927	0.0147

SVR	0.87	0.1346	0.0275	0.0223	0.0018	0.1652	0.0163
SPLS	0.90	0.1087	0.0236	0.0153	0.0008	0.0928	0.0172
Ridge	0.89	0.1121	0.0397	0.0168	0.0031	0.0991	0.0230
Elastic Net	0.89	0.1049	0.0359	0.0146	0.0026	0.0917	0.0209
Lasso	0.89	0.1044	0.0368	0.0144	0.0027	0.0908	0.0211
LDA	0.89	0.0343	0.0444	0.0015	0.0033	0.0216	0.0218
NN	0.89	0.0365	0.0431	0.0019	0.0034	0.0281	0.0233
Ensemble A	0.89	0.0940	0.0403	0.0097	0.0028	0.0896	0.0199
Ensemble B	0.89	/	0.0365	/	0.0024	/	0.0194
Ensemble C	0.89	/	0.0363	/	0.0017	/	0.0198
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0113	0.0005	0.0918	0.0828	0.1048	0.0196	0.0122	0.0005
0.0016	0.0006	0.0835	0.0850	0.0281	0.0255	0.0012	0.0010
0.0399	0.0007	0.1350	0.0968	0.1923	0.0183	0.0413	0.0005
0.0155	0.0009	0.1000	0.0826	0.1079	0.0204	0.0145	0.0007
0.0104	0.0005	0.0920	0.0839	0.1000	0.0189	0.0112	0.0005
0.0306	0.0007	0.1256	0.0973	0.1689	0.0196	0.0319	0.0006
0.0130	0.0005	0.0907	0.0798	0.1005	0.0174	0.0140	0.0005
0.0147	0.0013	0.0938	0.0811	0.1108	0.0203	0.0162	0.0006
0.0126	0.0010	0.0940	0.0824	0.1054	0.0204	0.0149	0.0007
0.0124	0.0010	0.0938	0.0824	0.1044	0.0204	0.0147	0.0007
0.0007	0.0011	0.0797	0.0834	0.0155	0.0369	0.0005	0.0020
0.0010	0.0013	0.0860	0.0885	0.0287	0.0417	0.0011	0.0022
0.0092	0.0008	0.0886	0.0819	0.0933	0.0213	0.0099	0.0006
/	0.0008	/	0.0818	/	0.0174	/	0.0004
/	0.0007	/	0.0813	/	0.0177	/	0.0004

Tabelle 563: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.7 (mu1=0, mu2=0.55).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.71	0.0169	0.0234	0.0004	0.0010	0.0188	0.0173
KNN	0.70	0.0416	0.0336	0.0025	0.0033	0.0364	0.0148
SVM	0.65	0.0740	0.0891	0.0138	0.0269	0.0290	0.0185
PLSDA	0.71	0.0333	0.0226	0.0020	0.0008	0.0242	0.0189
RFR	0.72	0.0263	0.0279	0.0015	0.0012	0.0231	0.0281
SVR	0.65	0.0939	0.1008	0.0250	0.0284	0.0339	0.0221
SPLS	0.72	0.0412	0.0290	0.0022	0.0019	0.0381	0.0209
Ridge	0.71	0.0275	0.0266	0.0010	0.0014	0.0282	0.0201
Elastic Net	0.71	0.0309	0.0299	0.0016	0.0020	0.0270	0.0210
Lasso	0.71	0.0304	0.0289	0.0016	0.0020	0.0265	0.0198
LDA PC:30	0.71	0.0275	0.0266	0.0010	0.0014	0.0282	0.0201
NN	0.70	0.0617	/	0.0049	/	0.0646	/
Ensemble A	0.71	0.0237	0.0250	0.0010	0.0011	0.0226	0.0170

Ensemble B	0.71	/	0.0619	/	0.0128	/	0.0249
Ensemble C	0.71	/	0.0308	/	0.0019	/	0.0182
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0005	0.0005	0.1927	0.1940	0.0215	0.0178	0.0007	0.0005
0.0020	0.0005	0.2005	0.1998	0.0326	0.0161	0.0013	0.0004
0.0017	0.0009	0.2204	0.2207	0.0247	0.0235	0.0009	0.0008
0.0011	0.0005	0.1957	0.1951	0.0306	0.0225	0.0015	0.0008
0.0012	0.0012	0.1926	0.1941	0.0261	0.0285	0.0012	0.0011
0.0021	0.0013	0.2210	0.2212	0.0290	0.0241	0.0013	0.0009
0.0019	0.0011	0.1904	0.1910	0.0347	0.0264	0.0015	0.0010
0.0011	0.0008	0.1916	0.1922	0.0315	0.0185	0.0015	0.0006
0.0013	0.0010	0.1928	0.1935	0.0285	0.0191	0.0012	0.0006
0.0013	0.0009	0.1929	0.1936	0.0287	0.0192	0.0012	0.0006
0.0011	0.0008	0.1916	0.1922	0.0315	0.0185	0.0015	0.0006
0.0053	/	0.2039	0.2004	0.0640	0.0255	0.0053	0.0009
0.0009	0.0005	0.1924	0.1939	0.0231	0.0190	0.0008	0.0005
/	0.0011	/	0.1940	/	0.0241	/	0.0008
/	0.0006	/	0.1934	/	0.0195	/	0.0005

Tabelle 564: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.8 (mu1=0, mu2=0.8).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.80	0.0520	0.0231	0.0031	0.0012	0.0534	0.0172
KNN	0.80	0.0263	0.0256	0.0010	0.0010	0.0259	0.0195
SVM	0.76	0.0937	0.0259	0.0133	0.0009	0.1013	0.0264
PLSDA	0.79	0.0553	0.0155	0.0043	0.0005	0.0597	0.0129
RFR	0.80	0.0427	0.0208	0.0029	0.0006	0.0432	0.0194
SVR	0.76	0.0898	0.0283	0.0110	0.0010	0.0946	0.0290
SPLS	0.80	0.0621	0.0221	0.0058	0.0008	0.0632	0.0177
Ridge	0.80	0.0648	0.0199	0.0064	0.0006	0.0669	0.0162
Elastic Net	0.79	0.0597	0.0207	0.0048	0.0007	0.0579	0.0171
Lasso	0.79	0.0587	0.0199	0.0047	0.0007	0.0569	0.0165
LDA PC:30	0.80	0.0280	0.0400	0.0012	0.0024	0.0230	0.0323
NN	0.80	0.0729	0.0754	0.0061	0.0072	0.0642	0.0414
Ensemble A	0.80	0.0523	0.0370	0.0036	0.0022	0.0540	0.0296
Ensemble B	0.80	/	0.0272	/	0.0011	/	0.0251
Ensemble C	0.80	/	0.0230	/	0.0011	/	0.0212
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0032	0.0007	0.1425	0.1402	0.0549	0.0201	0.0036	0.0006
0.0009	0.0007	0.1440	0.1436	0.0196	0.0183	0.0006	0.0005
0.0149	0.0009	0.1801	0.1663	0.1031	0.0296	0.0155	0.0013
0.0048	0.0004	0.1496	0.1416	0.0612	0.0187	0.0050	0.0006
0.0029	0.0005	0.1427	0.1412	0.0474	0.0206	0.0028	0.0006
0.0125	0.0010	0.1779	0.1665	0.0975	0.0308	0.0136	0.0014

0.0060	0.0006	0.1427	0.1373	0.0712	0.0191	0.0069	0.0006
0.0068	0.0005	0.1443	0.1385	0.0734	0.0194	0.0071	0.0006
0.0047	0.0006	0.1456	0.1415	0.0615	0.0179	0.0052	0.0006
0.0046	0.0005	0.1457	0.1416	0.0612	0.0178	0.0052	0.0006
0.0009	0.0016	0.1399	0.1411	0.0222	0.0292	0.0008	0.0013
0.0046	0.0031	0.1492	0.1481	0.0653	0.0441	0.0048	0.0029
0.0037	0.0015	0.1423	0.1396	0.0576	0.0231	0.0042	0.0008
/	0.0009	/	0.1399	/	0.0239	/	0.0008
/	0.0009	/	0.1396	/	0.0218	/	0.0007

Tabelle 565: Simulation Hetero-Ensembles mit 4000 Objekten (n1=2000, n2=2000), 40 Variablen, r=0.2 und Acc=0.9 (mu1=0, mu2=1.2).

Technik	**Acc**	**MAE**	**MAE_Log**	**MSE**	**MSE_Log**	**MAE_w**	**MAE_w_Log**
RF	0.89	0.0597	0.0292	0.0047	0.0013	0.0524	0.0178
KNN	0.89	0.0297	0.0380	0.0016	0.0026	0.0243	0.0197
SVM	0.87	0.1401	0.0282	0.0252	0.0018	0.1747	0.0157
PLSDA	0.89	0.1033	0.0370	0.0139	0.0019	0.1102	0.0227
RFR	0.89	0.0607	0.0375	0.0054	0.0025	0.0515	0.0165
SVR	0.87	0.1275	0.0348	0.0201	0.0024	0.1616	0.0194
SPLS	0.89	0.1084	0.0202	0.0149	0.0006	0.0918	0.0148
Ridge	0.89	0.1131	0.0355	0.0167	0.0021	0.0993	0.0199
Elastic Net	0.89	0.1043	0.0253	0.0149	0.0011	0.0914	0.0172
Lasso	0.89	0.1011	0.0315	0.0143	0.0014	0.0870	0.0199
LDA PC:30	0.89	0.0391	0.0374	0.0018	0.0024	0.0232	0.0205
NN	0.89	0.0295	0.0501	0.0012	0.0047	0.0264	0.0185
Ensemble A	0.89	0.0757	0.0271	0.0074	0.0014	0.0754	0.0165
Ensemble B	0.89	/	0.0294	/	0.0012	/	0.0187
Ensemble C	0.89	/	0.0308	/	0.0019	/	0.0194
MSE_w	**MSE_w_Log**	**Brier**	**Brier_Log**	**MAE_Cu**	**MAE_Cu_Log**	**MSE_Cu**	**MSE_Cu_Log**
0.0036	0.0005	0.0843	0.0830	0.0567	0.0199	0.0040	0.0006
0.0010	0.0009	0.0833	0.0853	0.0194	0.0286	0.0005	0.0011
0.0353	0.0007	0.1312	0.0982	0.1789	0.0216	0.0362	0.0006
0.0152	0.0009	0.1007	0.0834	0.1074	0.0206	0.0144	0.0008
0.0040	0.0007	0.0844	0.0831	0.0544	0.0200	0.0039	0.0006
0.0293	0.0009	0.1253	0.0981	0.1664	0.0203	0.0306	0.0006
0.0125	0.0003	0.0912	0.0805	0.1004	0.0176	0.0139	0.0005
0.0145	0.0009	0.0939	0.0814	0.1099	0.0203	0.0159	0.0006
0.0129	0.0005	0.0939	0.0826	0.1037	0.0189	0.0145	0.0006
0.0122	0.0007	0.0938	0.0831	0.1004	0.0203	0.0138	0.0007
0.0008	0.0010	0.0805	0.0842	0.0158	0.0365	0.0005	0.0019
0.0008	0.0012	0.0845	0.0869	0.0250	0.0397	0.0010	0.0021
0.0071	0.0005	0.0868	0.0823	0.0809	0.0205	0.0077	0.0006
/	0.0005	/	0.0822	/	0.0169	/	0.0004
/	0.0008	/	0.0817	/	0.0160	/	0.0004

9.1.6 Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels realer Datensätze

Tabelle 566: Auflistung des MAE_Cu und des MAE_Cu_Log der Ensemble Typen A, B und C auf Grundlage des Ames Datensatzes (MACCS) mit 6512 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Ensemble Typ	MAE_Cu	MAE_Cu_Log
Ensemble A	0.0841	0.0160
Ensemble B	/	0.0650
Ensemble C	/	0.0641

Tabelle 567: Auflistung des MAE_Cu und des MAE_Cu_Log der Ensemble Typen A, B und C auf Grundlage des CYP1A2 Datensatzes (MACCS) mit 7485 Molekülen.

Ensemble Typ	MAE_Cu	MAE_Cu_Log
Ensemble A	0.0718	0.0186
Ensemble B	/	0.0523
Ensemble C	/	0.0507

Tabelle 568: Auflistung des MAE_Cu und des MAE_Cu_Log der Ensemble Typen A, B und C auf Grundlage des Factor Xa Datensatzes (MACCS) mit 435 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Ensemble Typ	MAE_Cu	MAE_Cu_Log
Ensemble A	0.1010	0.0224
Ensemble B	/	0.0477
Ensemble C	/	0.0467

Tabelle 569: Auflistung des MAE_Cu und des MAE_Cu_Log der Ensemble Typen A, B und C auf Grundlage des Liver Datensatzes (MACCS) mit 951 Molekülen. Bei der SPLS wurden die Kolonnen, dessen Varianz 0 war, entfernt.

Ensemble Typ	MAE_Cu	MAE_Cu_Log
Ensemble A	0.0695	0.0611
Ensemble B	/	0.0470
Ensemble C	/	0.0550

9.1.7 MOE Deskriptoren

Tabelle 570: Auflistung der verwendeten 181 translations- und rotationsinvarianten MOE-Deskriptoren.

Code	Class	Description
apol	2D	Sum of atomic polarizabilities
ASA	i3D	Water accessible surface area
ASA+	i3D	Positive accessible surface area
ASA-	i3D	Negative accessible surface area
ASA_H	i3D	Total hydrophobic surface area
ASA_P	i3D	Total polar surface area
a_acc	2D	Number of Hbond acceptor atoms
a_acid	2D	Number of acidic atoms
a_aro	2D	Number of aromatic atoms
a_base	2D	Number of basic atoms
a_count	2D	Number of atoms
a_don	2D	Number of Hbond donor atoms
a_heavy	2D	Number of heavy atoms
a_hyd	2D	Number of hydrophobic atoms
a_IC	2D	Atom information content (total)
a_ICM	2D	Atom information content (mean)
a_nB	2D	Number of boron atoms
a_nBr	2D	Number of bromine atoms
a_nC	2D	Number of carbon atoms
a_nCl	2D	Number of chlorine atoms
a_nF	2D	Number of fluorine atoms
a_nH	2D	Numberof hydrogen atoms
a_nI	2D	Number of iodine atoms
a_nN	2D	Number of nitrogen atoms
a_nO	2D	Number of oxygen atoms
a_nP	2D	Number of phosphorus atoms
a_nS	2D	Number of sulfur atoms
balabanJ	2D	Balaban averaged distance sum connectivity
bpol	2D	Difference of bonded atom polarizabilities
b_1rotN	2D	Number of rotatable single bonds
b_1rotR	2D	Fraction of rotatable single bonds
b_ar	2D	Number of aromatic bonds
b_count	2D	Number of bonds
b_double	2D	Number of double bonds
b_heavy	2D	Number of heavy-heavy bonds
b_rotN	2D	Number of rotatable bonds
b_rotR	2D	Fraction of rotatable bonds
b_single	2D	Number of single bonds
b_triple	2D	Number of triple bonds
CASA+	i3D	Charge weighted positive surface area
CASA	i3D	Charge weighted negative surface area
chi0	2D	Atomic connectivity index (order 0)

chi0v	2D	Atomic valence connectivity index (order 0)
chi0v_C	2D	Carbon valence connectivity index (order 0)
chi0_C	2D	Carbonconnectivity index (order 0)
chi1	2D	Atomic connectivity index (order 1)
chi1v	2D	Atomic valence connectivity index (order 1)
chi1v_C	2D	Carbon valence connectivity index (order 1)
chi1_C	2D	Carbon connectivity index (order 1)
DASA	i3D	Absolute difference in surface area
DCASA	i3D	Absolute difference in charge weighted areas
dens	i3D	Mass density (AMU/A^3)
density	2D	Mass density (AMU/A**3)
diameter	2D	Largest vertex eccentricity in graph
dipole	i3D	Dipole moment
E	i3D	Potential Energy
E_ang	i3D	Angle Bend Energy
E_ele	i3D	Electrostatic energy
E_nb	i3D	Non-bonded energy
E_oop	i3D	Out-of-plane Energy
E_sol	i3D	Solvation energy
E_stb	i3D	Stretch-bend energy
E_str	i3D	Bond stretch energy
E_tor	i3D	Torsion energy
E_vdw	i3D	Van der Waals energy
FASA+	i3D	Fractional positive accessible surface area
FASA-	i3D	Fractional negative accessible surface area
FASA_H	i3D	Fractional hydrophobic surface area
FASA_P	i3D	Fractional polar surface area
FCASA+	i3D	Fractional charge-weighted positive surface area
FCASA-	i3D	Fractional charge-weighted negative surface area
FCharge	2D	Sum of formal charges
glob	i3D	Molecular globularity
Kier1	2D	First kappa shape index
Kier2	2D	Second kappa shape index
Kier3	2D	Third kappa shape index
KierA1	2D	First alphamodified shape index
KierA2	2D	Second alpha modified shape index
KierA3	2D	Third alpha modified shape index
KierFlex	2D	Molecular flexibility
logP(o/w)	2D	Log octanol/water partition coefficient
mr	2D	Molar refractivity
PC+	2D	Total positive partial charge
PC-	2D	Total negative partial charge
PEOE_PC+	2D	Total positive partial charge
PEOE_PC-	2D	Total negative partial charge
PEOE_RPC+	2D	Relative positive partial charge
PEOE_RPC-	2D	Relative negative partial charge
PEOE_VSA+0	2D	Total positive 0 vdw surface area

PEOE_VSA+1	2D	Total positive 1 vdw surface area
PEOE_VSA+2	2D	Total positive 2 vdw surface area
PEOE_VSA+3	2D	Total positive 3 vdw surface area
PEOE_VSA+4	2D	Total positive 4 vdw surface area
PEOE_VSA+5	2D	Total positive 5 vdw surface area
PEOE_VSA+6	2D	Total positive 6 vdw surface area
PEOE_VSA-0	2D	Total negative 0 vdw surface area
PEOE_VSA-1	2D	Total negative 1 vdw surface area
PEOE_VSA-2	2D	Total negative 2 vdw surface area
PEOE_VSA-3	2D	Total negative 3 vdw surface area
PEOE_VSA-4	2D	Total negative 4 vdw surface area
PEOE_VSA-5	2D	Total negative 5 vdw surface area
PEOE_VSA-6	2D	Total negative 6 vdw surface area
PEOE_VSA_FHYD	2D	Fractional hydrophobic vdw surface area
PEOE_VSA_FNEG	2D	Fractional negative vdw surface area
PEOE_VSA_FPNEG	2D	Fractional polar negative vdw surface area
PEOE_VSA_FPOL	2D	Fractional polar vdw surface area
PEOE_VSA_FPOS	2D	Fractional positive vdw surface area
PEOE_VSA_FPPOS	2D	Fractional polar positive vdw surface area
PEOE_VSA_HYD	2D	Total hydrophobic vdw surface area
PEOE_VSA_NEG	2D	Total negative vdw surface area
PEOE_VSA_PNEG	2D	Total polar negative vdw surface area
PEOE_VSA_POL	2D	Total polar vdw surface area
PEOE_VSA_POS	2D	Total positive vdw surface area
PEOE_VSA_PPOS	2D	Total polar positive vdw surface area
petitjean	2D	(diameter radius) / diameter
petitjeanSC	2D	(diameter radius) / radius
pmi	i3D	Principal moment of inertia
Q_PC+	2D	Total positive partial charge
Q_PC-	2D	Total negative partial charge
Q_RPC+	2D	Relative positive partial charge
Q_RPC-	2D	Relative negative partial charge
Q_VSA_FHYD	2D	Fractional hydrophobic vdw surface area
Q_VSA_FNEG	2D	Fractional negative vdw surface area
Q_VSA_FPNEG	2D	Fractional polar negative vdw surface area
Q_VSA_FPOL	2D	Fractional polar vdw surface area
Q_VSA_FPOS	2D	Fractional positive vdw surface area
Q_VSA_FPPOS	2D	Fractional polar positive vdw surface area
Q_VSA_HYD	2D	Total hydrophobic vdw surface area
Q_VSA_NEG	2D	Total negative vdw surface area
Q_VSA_PNEG	2D	Total polar negative vdw surface area
Q_VSA_POL	2D	Total polar vdw surface area
Q_VSA_POS	2D	Total positive vdw surface area
Q_VSA_PPOS	2D	Total polar positive vdw surface area
radius	2D	Smallest vertex eccentricity in graph
reactive	2D	Molecule contains reactive groups
rgyr	i3D	Radius of gyration

RPC+	2D	Relative positive partial charge
RPC-	2D	Relative negative partial charge
SlogP	2D	Log Octanol/Water Partition Coefficient
SlogP_VSA0	2D	Bin 0 SlogP (-10,-0.40]
SlogP_VSA1	2D	Bin 1 SlogP (-0.40,-0.20]
SlogP_VSA2	2D	Bin 2 SlogP (-0.20,0.00]
SlogP_VSA3	2D	Bin 3 SlogP (0.00, 0.10]
SlogP_VSA4	2D	Bin 4 SlogP (0.10, 0.15]
SlogP_VSA5	2D	Bin 5 SlogP (0.15, 0.20]
SlogP_VSA6	2D	Bin 6 SlogP (0.20, 0.25]
SlogP_VSA7	2D	Bin 7 SlogP (0.25, 0.30]
SlogP_VSA8	2D	Bin 8 SlogP (0.30, 0.40]
SlogP_VSA9	2D	Bin 9 SlogP (0.40,10]
SMR	2D	Molar Refractivity
SMR_VSA0	2D	Bin 0 SMR (0.000,0.110]
SMR_VSA1	2D	Bin 1 SMR (0.110,0.260]
SMR_VSA2	2D	Bin 2 SMR(0.260,0.350]
SMR_VSA3	2D	Bin 3 SMR (0.350,0.390]
SMR_VSA4	2D	Bin 4 SMR (0.390,0.440]
SMR_VSA5	2D	Bin 5 SMR (0.440,0.485]
SMR_VSA6	2D	Bin 6 SMR (0.485,0.560]
SMR_VSA7	2D	Bin 7 SMR (0.560,10]
Standardabweichung_dim1	i3D	Standard dimension 1
Standardabweichung_dim2	i3D	Standard dimension 2
Standardabweichung_dim3	i3D	Standard dimension 3
TPSA	2D	Topological Polar Surface Area (A**2)
VAdjEq	2D	Vertex adjacency information (equal)
VAdjMa	2D	Vertex adjacency information (mag)
VDistEq	2D	Vertex distance equality index
VDistMa	2D	Vertex distance magnitude index
vdw_area	2D	Van der Waals surface area (A**2)
vdw_vol	2D	Van der Waals volume (A**3)
vol	i3D	Van der Waals volume
VSA	i3D	Van derWaals surface area
vsa_acc	2D	VDW acceptor surface area (A**2)
vsa_acid	2D	VDW acidic surface area (A**2)
vsa_base	2D	VDW basic surface area (A**2)
vsa_don	2D	VDW donor surface area (A**2)
vsa_hyd	2D	VDW hydrophobe surface area (A**2)
vsa_other	2D	VDW other surface area (A**2)
vsa_pol	2D	VDW polar surface area (A**2)
Weight	2D	Molecular weight (CRC)
weinerPath	2D	Weiner path number
weinerPol	2D	Weiner polarity number
zagreb	2D	Zagreb index

9.2 Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeits-schätzern versus CP

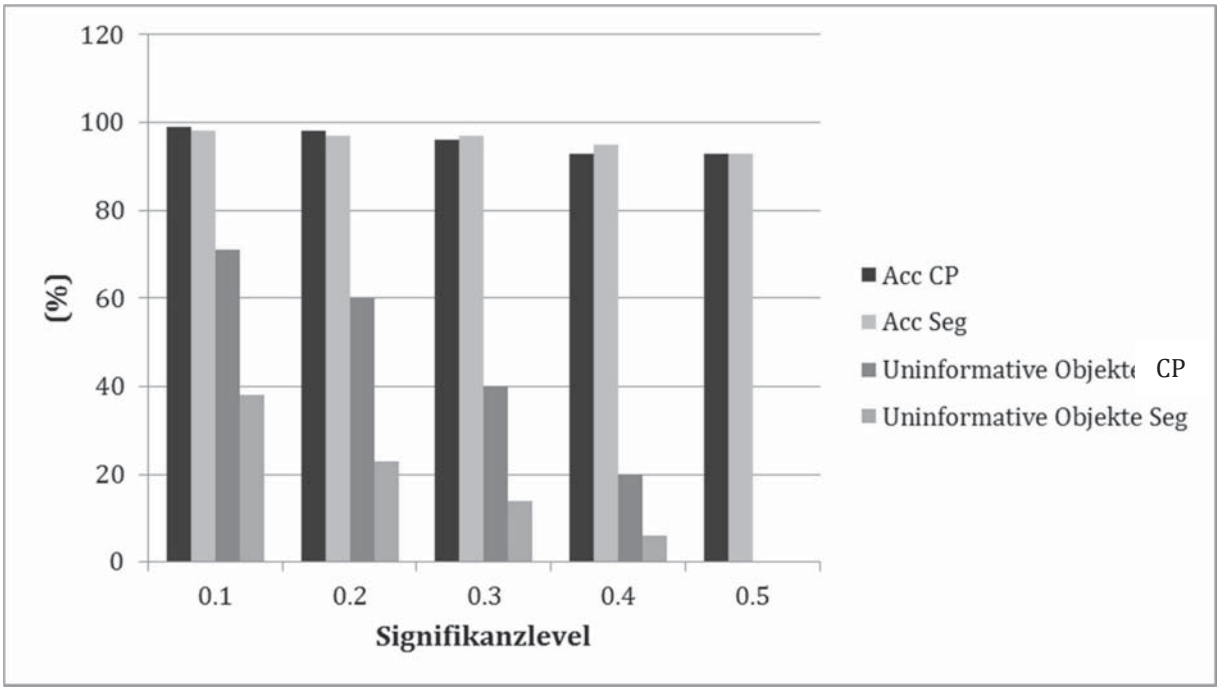


Abbildung 88: Die Korrektklassifizierungsrate (Acc) und die uninformativen Objekte des CP und der Methode, welche Segmente entfernt auf Basis des Faktor Xa Datensatzes. Beide Methoden sind in der Lage die vorgegebenen Signifikanzlevel zu halten. Der Ansatz, welcher Segmente aus der Mitte der sortierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer entfernt, bezeichnet weniger Objekte als uninformativ.

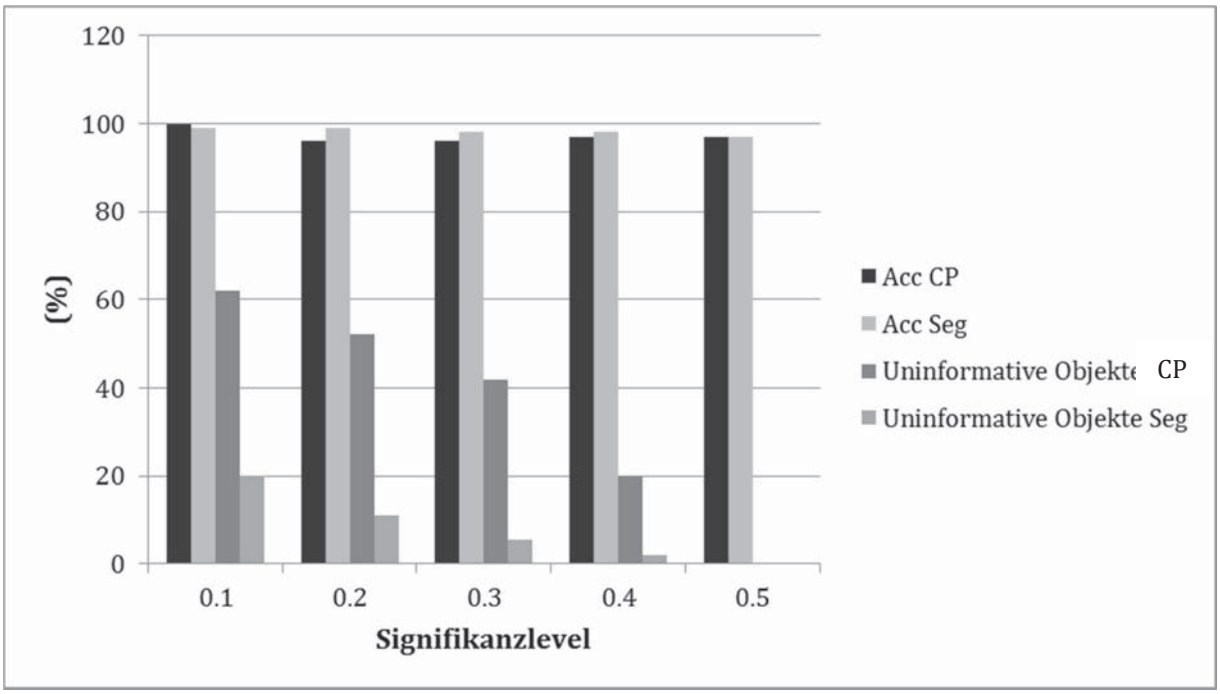


Abbildung 89: Die Korrektklassifizierungsrate (Acc) und die uninformativen Objekte des CP und der Methode, welche Segmente entfernt auf Basis des Musk2 Datensatzes. Beide Methoden sind in der Lage die vorgegebenen Signifikanzlevel zu halten. Der Ansatz, welcher Segmente aus der Mitte der sortierten Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer entfernt, bezeichnet weniger Objekte als uninformativ.