Einfluss der Fallzahl auf verschiedene Varianzkomponenten und Gütemaße bei der Anwendung ausgewählter Verfahren des Maschinellen Lernens

Bachelorarbeit

Fachhochschule Stralsund
Fachbereich Wirtschaft
Studiengang Betriebswirtschaftslehre

vorgelegt von: Bjarne Seen
Matrikelnummer: 16156
Bjarne.Seen@fh-stralsund.de

Gutachter: Prof. Dr. Lieven Kennes
Dr. Paul Wolf

eingereicht am: 13.03.2020

Inhaltsverzeichnis

1 Einleitung **1**

2 Methoden **4**

2.1 Entscheidungsbaum 4

2.2 Random Forest 6

2.3 Durchführung 7

2.3.1 Selektion kleinerer Datensätze 7

2.3.2 Untersuchung verschiedener Varianzkomponenten 8

2.4 Gütemaße 10

2.5 Datensatz 13

3 Ergebnisse **22**

3.1 Gütemaße der verschiedenen Fallzahlen 22

3.2 Varianzkomponenten der Modelle 27

3.2.1 Standardabweichung des Mittelwerts 27

3.2.2 Mittelwert der Standardabweichung 27

4 Diskussion und Interpretation der Ergebnisse **30**

4.1 Gütemaße der verschiedenen Fallzahlen 30

4.2 Varianzkomponenten der Modelle 33

5 Fazit **37**

A Anhang **41**

A.1 Datensatz 41
A.2 R Studio Code 42
A.3 Abbildungen 43
A.4 Erklärung Speicherung Daten 51
A.5 Eidesstaatliche Erklärung 53

Abbildungsverzeichnis

2.1 Beispiel Entscheidungsbaum ($n = 100$) 6

2.2 Flowchart Selektion kleinerer Datensätze 9

2.3 Flowchart Untersuchung verschiedener Varianzkomponenten 10

2.4 ROC Kurve Beispiel 14

2.5 Windrichtungen der stärksten Böe (Sydney) 19

2.6 Boxplot der Variable MinTemp 20

2.7 Boxplot der Variable MaxTemp 20

2.8 Boxplot der Variable Humidity9am 20

2.9 Boxplot der Variable Humidity3pm 20

2.10 Boxplot der Variable WindSpeed9am 21

2.11 Boxplot der Variable WindSpeed3pm 21

3.1 Korrektklassifikationsrate 24

3.2 Sensitivität 24

3.3 positiver Vorhersagewert 25

3.4 Spezifität 25

3.5 Negativer Vorhersagewert 25

3.6 Area under the curve 25

3.7 Entscheidunsbaum $n = 1000$ 29

3.8 Random Forest $n = 1000$ 29

3.9 Entscheidunsbaum $n = 400$ 29

3.10 Random Forest $n = 400$ 29

3.11 Entscheidunsbaum $n = 100$ 29

3.12 Random Forest $n = 100$ 29

A.1 Weitere Diagramme qualitiative Variablen Datensatz Sydney ($n = 1701$) . . 43

A.2 (1) Boxplots zu weiteren Variablen des Datensatzes Sydney ($n = 1701$) . . 44

A.3 (2) Boxplots zu weiteren Variablen des Datensatzes Sydney ($n = 1701$) . . 45

A.4 Boxplots zu Varianzkomponenten des Gütemaßes Sensitivität 46

A.5 Boxplots zu Varianzkomponenten des Gütemaßes positiver Vorhersagewert 47

A.6 Boxplots zu Varianzkomponenten des Gütemaßes Spezifität 48

A.7 Boxplots zu Varianzkomponenten des Gütemaßes negativer Vorhersagewert 49

A.8 Boxplots zu Varianzkomponenten des Gütemaßes Area under the curve . . 50

Tabellenverzeichnis

2.1 Konfusionsmatrix 11

2.2 Legende Datensatz Rain in Australia 18

2.3 Zusammenfassung deskriptiver Statistik Datensatz Sydney 20

3.1 Gütemaße Mittelwert 26

3.2 Gütemaße Standardabweichung 26

3.3 Gütemaße Standardabweichung des Mittelwerts 28

3.4 Gütemaße Mittelwert der Standardabweichung 28

4.1 Gütemaße Mittelwerte Prozentuale Veränderung 34

4.2 Gütemaße Standardabweichungen Prozentuale Veränderung 34

4.3 Gütemaße Standardabweichungen des Mittelwerts Prozentuale Veränderung 36

4.4 Gütemaße Mittelwerte der Standardabweichung Prozentuale Veränderung 36

1. Einleitung

In den letzten Jahren hat Maschinelles Lernen die Welt in vielen Bereichen revolutioniert. Trotz eines kleinen Datensatzes ein geeignetes Modell des Maschinellen Lernens zu bauen bleibt jedoch eine Herausforderung. Das Problem kleiner Datensätze ist eines, welches besonders häufig in der Medizin auftaucht. Dies liegt oft an dem Aufwand und den Kosten, welche mit der Erhebung der Daten verbunden sind. Seien es Daten von Patienten mit Herzkrankheiten oder von Patienten, welche eine Organtransplantat erhalten haben. Da auch die Qualität der Daten hierbei eine große Rolle spielt, ist es wichtig so viele Werte wie möglich, so genau wie möglich zu erheben. Somit bleiben oft aus finanziellen und zeitlichen Gründen Datensätze beschränkt auf 50 bis 100 Einträge.

Mit Algorithmen des Maschinellen Lernens ist es möglich unterschiedlichste Probleme anzugehen, sei es Klassifikation, Clustering oder Regression. In der Regel jedoch funktioniert dies besser bei Vorliegen von *Big Data*. Mehr Daten sorgen für einen größeren Trainingsdatensatz, auf dem die Algorithmen trainiert werden können und somit oft sehr gute Ergebnisse erzielen. In der Medizin wiederum ergibt sich die Herausforderung diese Algorithmen auf kleine Datensätze anzuwenden und trotz eines kleineren Trainingsdatensatzes gute Ergebnisse zu erreichen. Denn gerade bei medizinischen Problemen können diese Algorithmen große Hilfe leisten, in dem Herzkrankheiten oder das Abstoßen eines Transplantats frühzeitig vorhergesagt werden kann.

Ziel dieser Arbeit ist es aufzuzeigen, wie viel ungenauer ein Algorithmus auf einem kleineren Datensatz ist als auf einem weitaus größeren, sofern dies überhaupt der Fall ist. Wie viel mehr bringt es die Größe des kleinen Datensatzes zu erweitern? Reicht es, statt Daten von 100 Patienten, Daten von 400 zu erheben oder müssen es 1000 oder mehr sein, um signifikant bessere Modelle zu bauen? Welches der Modelle funktioniert eventuell besser bei kleinen Datensätzen?

Bisherige Paper behandelten zum einen die Verbesserung der Gütemaße von den Anwendungen des Maschinellen Lernens Entscheidungsbaum und neuronales Netz bei

kleinen Datensätzen mithilfe der sogenannten *Method of multiple runs* (Shaikhina et al., 2015). Hierbei wurden zwei kleine Datensätze ($n = 80$ und $n = 35$) genutzt, wobei die Algorithmen mehrfach mit unterschiedlichen anfänglichen Parametern angewendet worden sind und das beste Modell gewählt wurde. In einem weiteren Paper wurden mehrere neuronale Netze auf einem kleinen Datensatz ($n = 56$) mithilfe der *Method of multiple runs* gebaut (Shaikhina and Khovanova, 2017). Zusätzlich analysierte man die Gütemaße des Modells bei größeren Datensätzen ($n = 100$ und $n = 1000$), um zu zeigen, dass trotz des deutlich kleineren Datensatzes von $n = 56$ vergleichbare Ergebnisse erzielt werden konnten. Dies verstärkte man des Weiteren durch das Bauen von neuronalen Netzen auf Ersatzdaten (engl.: *surrogate data*), um zu zeigen, dass die guten Modelle der vielen neuronalen Netze nicht durch Zufall zustande kamen. Die Ersatzdaten sind Daten, welche zwar die gleichen Eigenschaften bezüglich der Lagemaße der einzelnen Variablen besitzen, jedoch voneinander unabhängig generiert werden. Shaikhina et al. (2017) nach ist der Unterschied in der Performance auf einem Datensatz mit einer Fallzahl von $n = 80$ zwischen den Anwendungen des Maschinellen Lernes Random Forest und Decision Tree nicht signifikant. So konnten beiden Modelle die Daten gleich gut klassifizieren. In der Studie von Ali et al. (2012) wurden 20 Datensätze zwischen $n = 148$ und $n = 20000$ untersucht. Dafür nutzte man zum einen den J48 Algorithmus für die Erstellung eines Entscheidungsbaums und zum anderen die Random Forest Ensemble-Methode. Es konnte gezeigt werden, dass die Korrektklassifikationsrate mit höherer Fallzahl für beide Anwendungen stieg, wobei der Random Forest bessere Werte bei höherer Fallzahl und J48 bei niedrigerer Fallzahl erzielte. Ob dieser Unterschied signifikant war, konnte jedoch nicht gezeigt werden. Prajwala (2015) konnten zeigen, dass bei der Durchführung beider Anwendungen des Maschinellen Lernens Random Forest besser klassifizierte als ein Entscheidungsbaum. Als Algorithmus für den Entscheidungsbaum wurde der ID3 (Iterative Dichotomiser 3) genutzt. In der Studie wurde der gleiche Datensatz wie in dieser Arbeit genutzt, wobei es sich um ein Teildatensatz mit $n = 256$ Fällen handelt.

Im Folgenden wird sowohl ein Entscheidungsbaum als auch ein Random Forest Modell mehrfach auf unterschiedlich großen Datensätzen gebaut. Dabei werden diese Modelle anhand unterschiedlicher Gütemaße miteinander verglichen, um festzustellen, wie sehr diese voneinander abweichen. Zudem werden auch die Varianzkomponenten untersucht, welchen durch das zufällige Erstellen kleinerer Datensätze entstehen, um zu zeigen, wie viel Ungenauigkeit darauf zurückzuführen ist. Es soll gezeigt werden wie viel Varianz durch unterschiedliche Datensätze, auch gleicher Größe, entstehen kann und wie groß diese ist.

Die Arbeit ist dafür in drei größere Kapitel aufgeteilt. Im Kapitel 2 der Arbeit wird die Methodik beschrieben, wobei dazu vor allem die Vorstellung der beiden Anwendung des Maschinellen Lernens Entscheidungsbaum und Random Forest zählt. Des Weiteren wird vorgestellt, mit welchen unterschiedlichen Fallzahlen gearbeitet wird und inwiefern Varianzkomponenten berücksichtigt werden. Letztlich werden die Gütemaße aufgezeigt mit dessen Hilfe die Modelle analysiert und bewertet werden können. Auch wird der Datensatz, welcher als Grundlage der Arbeit genutzt wird, ausführlich dargestellt und erläutert.

Kapitel 3 setzt sich mit den erzielten Ergebnissen auseinander, welche sowohl in Tabellen als auch grafisch vorgestellt werden. Diese Ergebnisse werden im Kapitel 4 diskutiert und interpretiert.

Die wichtigsten Fragestellungen der Arbeit, welche im Verlauf als Grundlage dienen, sind die folgende:

- Vergleich der Anwendungen Entscheidungsbaum und Random Forest bei unterschiedlichen Fallzahlen, insbesondere kleinen Datensätzen
- Untersuchung der unterschiedlichen Fallzahlen und der Einfluss dieser anhand von Varianzkomponenten

2. Methoden

2.1 Entscheidungsbaum

Ein klassisches Verfahren zum Klassifizieren von Datensätzen ist das Entscheidungsbaumverfahren. Es bietet die Möglichkeit sowohl eine Regression als auch eine Klassifikation durchzuführen. Ziel bei der Klassifikation mit dem Entscheidungsbaumverfahren ist es die Daten mithilfe der Einflussvariablen in möglichst heterogene Gruppen aufzuteilen. Dies geschieht in der Regel über mehrere Knoten und am Ende des Baumes sind Blätter vorzufinden, welchen unterschiedliche Ergebnisvariablen zugeschrieben werden. Der von Breiman et al. (1984) publizierte CART(*Classification and Regression Trees*)-Algorithmus wird ausschließlich bei Binärbäumen eingesetzt. Binärbaum bedeutet, dass die Knoten des Baumes nur zwei Äste besitzen, welche zu weiteren Knoten oder den Blättern führen. Bei dem CART-Algorithmus werden die Rohdaten verwendet, diese müssen also nicht vor dem Trainieren des Modells angeglichen werden. Es kann sowohl mit quantitativen als auch mit qualitativen Einfluss- und Ergebnisvariablen gearbeitet werden. Bei qualitativen Einflussvariablen wird an Knoten geprüft auf das Vorkommen in einer Menge, während bei quantitativen Variablen geschaut wird, ob der Wert kleiner, größer, kleiner gleich oder größer gleich eines bestimmten Schwellenwertes ist. Von dem anfänglichen Wurzelknoten werden mithilfe des Gini Index als Aufteilungskriterium die Daten in zwei neue Knoten geteilt.

Bei dem Gini Index handelt es sich um einen Wert der wie folgt berechnet wird:

$$Gini(B) = 1 - \sum_{i=1}^{I} p(b_i)^2$$

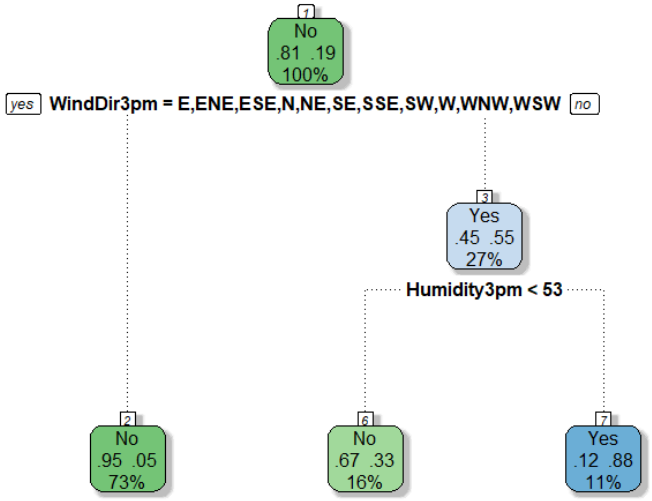
Hierbei steht $p(bi)$ für die relativen Häufigkeiten der einzelnen Klassen in einem Knoten und I für die Anzahl der unterschiedlichen Klassen, bei binären Problemen ist demnach $I = 2$. In dem binären Fall liegt der Wert des Gini Index immer zwischen $0 <=$

$Gini(B) <= 0.5(1 - 1/I)$. Beträgt der berechnete Wert $gini(B) = 0$, spricht man von völliger Reinheit eines Knotens bzw. Blattes, da nur eine der beiden Klassen dort vorkommt. Sofern wiederum $gini(B) = 0.5$ beträgt ist der Knoten bzw. das Blatt völlig unrein, da genau gleich viele beider Klassen vorkommen. Demnach ist der Gini Index ein Maß der Unreinheit eines Knoten bzw. Blattes. Sinnvoll genutzt wird dieser Index, um zu entscheiden welche Einflussvariable als Knoten genutzt werden kann um die besten beiden nächsten Knoten zu erzielen. Dafür wird der Gini Index des aktuellen Knoten berechnet und alle gemäß Datenlage möglichen Splits durchprobiert. Dabei werden jeweils die Gini Indizes der beiden neuen Knoten berechnet und gewichtet addiert und mit dem ursprünglichen Gini Index des Ausgangsknotens verglichen. Der bestmögliche Split ist dann jener, bei dem die größte Differenz zu dem ursprünglichen Gini Index vorliegt, da diese genau die Verminderung der Unreinheit angibt.

Dieses Vorgehen wiederholt sich solange bis keine Aufteilungen mehr möglich sind, da die einzelnen Blätter homogene Daten enthalten. Dies führt zu einem Problem, welches sich Overfitting nennt (Dietterich, 1995). Das bedeutet, dass der gebaute Entscheidungsbaum sehr gut auf den Trainingsdaten, wenn nicht sogar ohne einen Fehler funktioniert. Zurückzuführen ist das darauf, dass Einflussvariablen, die statistisch unabhängig von der Zielvariablen sind, genutzt werden, um die Daten zu splitten. Die Berechnung des Gini Index erklärt dieses zwar als sinnvollen Split, jedoch lässt sich das nicht auf ungesehene Daten übertragen. Würden man also das Modell auf den Trainingsdaten ansetzen, wäre die erzielte Genauigkeit sehr hoch oder sogar 100%. Das Ziel eines solchen Modells ist es aber auf neuen ungesehen Daten eine hohe Menge an richtigen Vorhersagen zu treffen. Wird ein Modell auf ungesehenen Daten angewendet, welches von Overfitting betroffen ist, führt dies zum Erhöhen des Generalisierungsfehlers, welcher angibt wie gut ein Modell auf neuen Daten funktioniert. Bei dem R-Paket *rpart*, welches für diese Arbeit genutzt wurde, gibt es hierfür einen sogenannten Parameter cp. cp steht für *complexity parameter* und liegt standardmäßig bei dem Wert $cp = 0.01$. Bei dem Parameter handelt es sich um einen Stopp-Parameter, welcher angibt, wann das weitere Aufteilen der Knoten nicht mehr den minimalen Wert der Verbesserung des Entscheidungsbaumes erreicht. Dafür werden die Summe der Fehlklassifikationen in allen Blättern addiert und die Anzahl der Knoten mit dem Wert α multipliziert. Diese beiden Werte werden aufaddiert und bilden die Kostenkomplexität des Modells.

$$R_\alpha(T) = R(T) + \alpha|T|$$

Abbildung 2.1: Beispiel Entscheidungsbaum ($n = 100$)



cp entspricht α und kann in der rpart Funktion als Schwellenwert angegeben werden, um Overfitting des Baumes zu vermeiden. Dabei sollte darauf geachtet werden, dass der Wert nicht zu groß gewählt wird, denn das führt zu einem zu kleinen Entscheidungsbaum, welcher nicht alle vorhandenen Informationen sinnvoll verarbeitet und daher auch zu einem erhöhten Generalisierungsfehler beiträgt. Somit führt ein höheres α zu höheren Kosten, während ein kleineres α zu kleineren Kosten führt. Dies wird bei dem Entscheidungsbaum dadurch erkennbar, das bei Wahl eines kleinen Komplexitätsparameters ($cp = 0.0001$) als Stopp Parameter, dieser Baum sehr groß wird und dementsprechend der Baum sehr früh stoppt weiter Knoten zu bilden, wenn dieser Parameter groß ($cp = 0.1$) gewählt wird. In Abbildung 2.1 ist ein kleiner Entscheidungsbaum, welcher mithilfe von *fancyRPartPlot* erstellt wurde. Er besteht aus einem Wurzelknoten, einem weiteren Knoten und drei Blättern.

2.2 Random Forest

Random Forest ist ein Verfahren, welches sowohl für Klassifikation als auch Regression genutzt wird (Hayes et al., 2015). Es handelt sich um eine Ensemble Methode, das heißt, dass mehrere Modelle genutzt werden, welche jeweils auf unterschiedlichen Teildatensätzen erzeugt werden. Es ähnelt dem Entscheidungsbaumverfahren in dem Maße, dass es mehrere Entscheidungsbäume baut, und am Ende eine Mehrheitswahl bei

der Klassifikation oder bei der Bildung des Durchschnitts der Vorhersagen der einzelnen Regressionen herbeizieht. Random Forest nutzt *Bootstrap Aggregating* (*Bagging*) um mehrere Entscheidungsbäume auf unterschiedlichen Trainingsdatensätzen zu trainieren. Bagging beinhaltet zum einen ein *bootstrap sample*, also die Erstellung von Teildatensätzen. Die gesamte Strichprobe N wird dafür in Teildatensätze der Größe $n < N$ eingeteilt. Das geschieht in der Regel $100 - 500$ mal, je nach Algorithmus unterschiedlich bzw. frei wählbar. Dabei geschieht die Auswahl zufällig und mit Zurücklegen. Das *Aggregating* in dem Wort *Bagging* steht für das Aufaddieren der Ergebnisse. Das heißt, dass für jeder der trainierten Entscheidungsbäume eine Vorhersage zur Klassenzuordnung getroffen wird. Der Random Forest berücksichtigt nun die einzelnen Vorhersagen der Bäume in einer Mehrheitswahl , um so zu einer Vorhersage für die ungesehenen Daten zu kommen.

Zusätzlich geschieht die Auswahl der Einflussvariablen bei einem Random Forest zufällig. Dies führt zu einer erhöhten Varianz der Bäume und soll Overfitting vermeiden. Nach Hastie et al. (2009) wählt man dabei von den gesamten Einflussvariablen m für den Fall eines Klassifikationsproblems in der Regel $\sqrt{m}$ oder $\log_2 m$ Einflussvariablen für die einzelnen Bäume. Als weitere Besonderheit beim Random Forest gegenüber den üblichen Entscheidungsbäumen werden die einzelnen Bäume nicht gestutzt. Das bedeutet, das vorher erklärte Verfahren mit dem cp Schwellenwert wird nicht berücksichtigt und die Splits finden statt, bis homogene Blätter vorhanden sind. Zusätzlich berechnet ein Random Forest den sogenannten *Out-Of-Bag Error*. Dafür werden die beim *Bagging* übrig gebliebenen Daten genutzt und stellen den *Out-Of-Bag* Datensatz dar. Damit kann der durchschnittliche Vorhersagefehler von jedem Trainingseintrag x_i berechnet werden, wobei nur die Bäume genutzt werden welche x_i nicht in ihrem Trainingssatz hatten. Es ist demnach eine Art integriertes Kreuzvalidierungsverfahren, welches bereits nach dem Training ein Überblick über die Güte des Modells bietet. Im Rahmen dieser Arbeit wurde zur Durchführung das R-Paket *randomforest*, basierend auf CARTs, genutzt.

2.3 Durchführung

2.3.1 Selektion kleinerer Datensätze

Da es ein Ziel der Arbeit ist, zu zeigen wie gut die Klassifikationsverfahren Random Forest und Entscheidungsbaum auf unterschiedlich kleinen Datensätzen funktionieren, werden Teildatensätze von dem ursprünglichen Datensatz gebildet. Die Auswahl dieser

$n = 1000, 400$ und 100 großen Datensätze erfolgt zufällig. Die Wahl dieser Größen ist unter anderem auf die Studie von Shaikhina and Khovanova (2017), in welcher neuronale Netze auf den Stichproben $n = 1030, 100$ und 56 untersucht wurden zurückzuführen. Die Unterteilung der Teildatensätze in ein Trainings- und ein Testdatensatz geschah ebenso zufällig. Der Trainingsdatensatz dient dazu, dass Modell zu trainieren, während mithilfe des Testdatensatzes evaluiert wird, wie gut das Modell gelernt hat und auf zuvor ungesehenen Daten funktioniert. Hierbei ist es üblich, 75% der Daten zum Training des Modells zu nutzen und die übrigen 25% zum Testen des Modells. Um jedoch zu vermeiden, dass durch eine Zufallsauswahl ein nicht repräsentatives Ergebnis erzielt wird, werden die Teildatensätze jeweils einhundert mal von dem $n = 1701$ großen Ausgangsdatensatz gebildet. Daraufhin trainiert und testet man die Klassifikationsverfahren einhundert mal. Die Bildung eines Durchschnitts der Gütemaße und der Standardabweichung dieser enthält die Informationen, wie gut die Verfahren auf unterschiedlich großen Stichproben funktionieren. In Abbildung 2.2 ist ein Flowchart dargestellt, welcher einen kleinen Überblick über den Ablauf gibt.

2.3.2 Untersuchung verschiedener Varianzkomponenten

Bei dem erläuterten Verfahren in Unterkapitel 2.3.1 erfolgt zweimal eine zufällige Auswahl der Daten bei der Bildung des Trainingsdatensatzes für das Training des Modells. Zum einen bei der Auswahl des Teildatensatzes aus dem ursprünglichen Datensatz ($n = 1701$), zum anderen beim Aufteilen in ein Trainings- und Testdatensatz, bei dem man die Daten zufällig splittet. Da in der Praxis oft nur ein Datensatz vorliegt und demnach nur bei der Wahl von Trainings- und Testdatensatz Spielraum vorhanden ist, ist es sinnvoll darzustellen, welche zusätzlich Varianz aufgrund dessen entsteht. Um dies aufzuzeigen werden in einem weiteren Experiment zehnmal Teildatensätze gebildet auf jeweils den Größen $n = 1000, 400$ und 100. Zu jedem dieser insgesamt 30 Teildatensätze werden nun wieder einhundertmal zufällig die Trainings- und Testdatensätze gebildet. Jede dieser einhundert Trainingsdatensätze wird genutzt, um ein Entscheidungsbaum und ein Random Forest Modell zu trainieren. Mithilfe der dazugehörigen Testdatensätze berechnet man die Gütemaße der einzelnen Modelle. Somit liegen insgesamt 3000 trainierte Entscheidungsbäume und ebenso viele Random Forest Modelle und deren zugehörige Gütemaße vor. Um diese Daten sinnvoll zu interpretieren, wird das arithmetische Mittel und die Standardabweichung der 100 Gütemaße der Modelle gebildet, welche auf demselben Teildatensatz trainiert wurden. In Abbildung 2.3 ist ein Flowchart dargestellt,

Abbildung 2.2: Flowchart Selektion kleinerer Datensätze

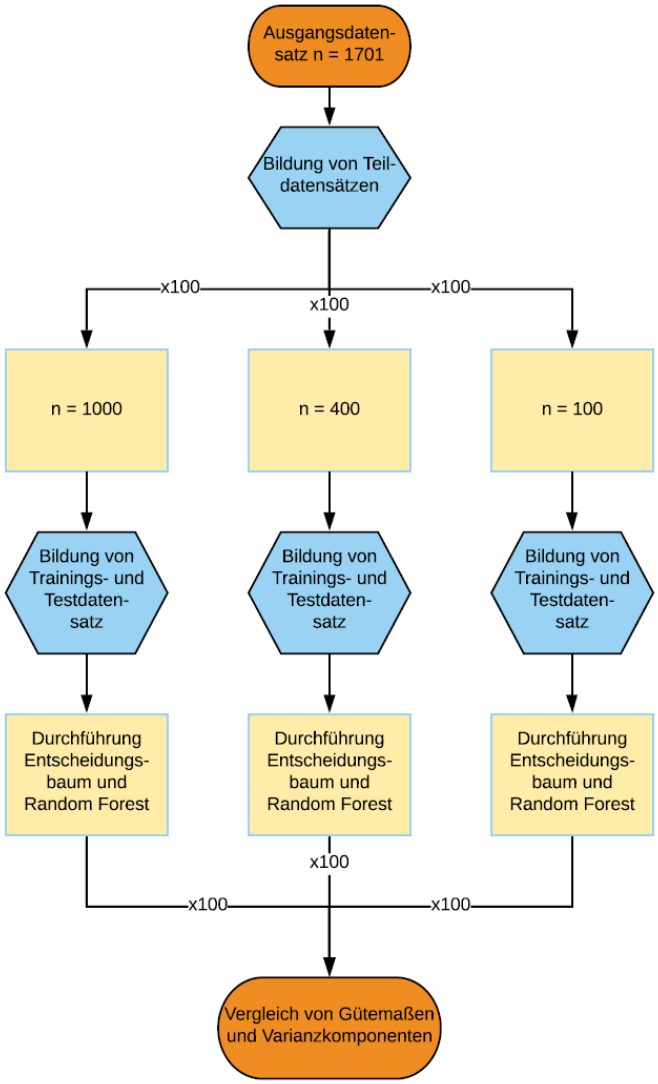
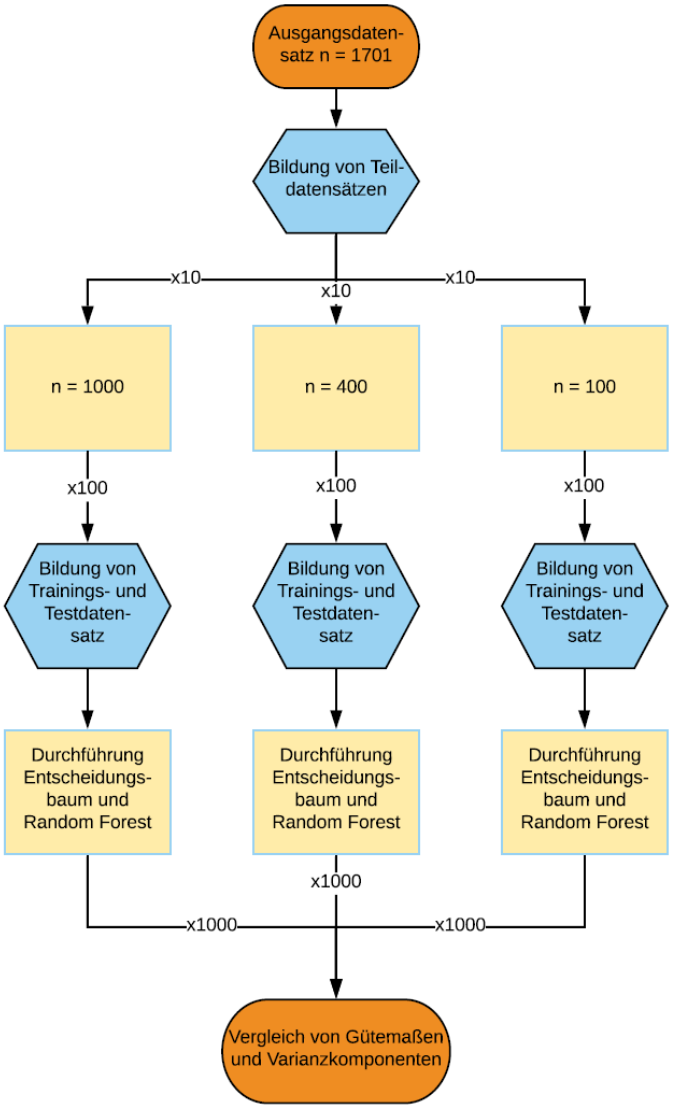


Abbildung 2.3: Flowchart Untersuchung verschiedener Varianzkomponenten



welcher einen kleinen Überblick über den Ablauf gibt. Unterschied zum obigen Verfahren ist, dass nun zuerst die zehn Teildatensätze gebildet werden, und auf diesen fixen Datensätzen zufällig Trainings- und Testdatensatz gebildet, trainert und getestet werden.

2.4 Gütemaße

Um die Performance der Verfahren zu klassifizieren, gibt es eine Anzahl unterschiedlicher Gütemaße. Bei der Klassifikation durch ein Verfahren unterscheidet man zwischen der direkten Zuordnung zu einer Klasse oder alternativ die Berechnung einer Klassenwahrscheinlichkeit. Bei der Berechnung der Klassenwahrscheinlichkeit werden die Beobachtungen durch Festlegen eines Schwellenwertes c in die jeweiligen Klassen einsortiert.

	Tatsache positiv	Tatsache negativ
Test positiv	True Positive (TP)	False Positive(FP)
Test negativ	False Negative (FN)	True Negative (TN)

Tabelle 2.1: Konfusionsmatrix

In der Regel wird dabei der Schwellenwert $c = 0.5$ genutzt. Demnach werden Beobachtungen mit einer Wahrscheinlichkeit über 0.5 der positiven Klasse zugeordnet und Beobachtungen mit Wahrscheinlichkeiten kleiner oder gleich 0.5 der negativen Klasse.

Als Grundlage der meisten binären Klassifikationsprobleme dient eine Konfusionsmatrix. Dabei stellt man eine 2x2 Matrix auf, in der die vier unterschiedlichen Fälle der Klassifikation dargestellt sind. Beim Klassifizieren der Bonität in Bezug auf die Rückzahlung eines Kredites eines Bankkunden können also folgende Fälle auftreten:

- **Richtig positiv** (engl.: *True positive*): Der Kunde kann zahlen und der Test bestätigt dies.
- **Falsch positiv** (engl.: *False positive*): Der Kunde kann nicht zahlen, der Test meine aber er sei in der Lage.
- **Falsch negativ** (engl.: *False negative*): Der Kunde kann zahlen, der Test widerspricht dem.
- **Richtig negativ** (engl.: *True positive*): Der Kunde kann nicht zahlen und der Test bestätigt dies.

Mithilfe der Matrix, welche in Tabelle 2.1 dargestellt ist, lassen sich viele Gütemaße berechnen die zwar alle die Güte des Verfahrens, jedoch mit Fokus auf unterschiedlichen Aspekte, beschreiben.

- Korrektklassifikationsrate:

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

  Die wohl bekannteste ist die Korrektklassifikationsrate (engl.: *Accuracy*) oder auch Genauigkeit, bei welcher die Summe der richtig klassifizierten, also die richtig positiven und richtig negativen, durch die Summer aller Fälle geteilt wird. Dies entspricht dem Anteil der Datensätze, welche durch das Verfahren richtig klassifiziert werden.

- Sensitivität:

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

  Die Sensitivität (engl.: *Sensitivity, True Postive Rate, Recall*) oder auch Trefferquote oder Richtig-Positiv-Rate ist ein weiteres Gütemaß. Sie beschreibt eine bedingte Wahrscheinlichkeit bei welcher nur die positiv klassifizierten Objekte betrachtet werden und der Anteil der davon tatsächlich positiven Fälle berechnet wird.

- Spezifität:

$$Spec = \frac{TN}{TN + FP}$$

  Demgegenüber steht die Spezifität (engl.: *Specifity, True Negative Rate*). Auch diese ist eine bedingte Wahrscheinlichkeit, welche den Anteil aller richtig als negativ klassifizierten Objekte an den gesamten negativen Fällen beschreibt.

- positiver Vorhersagewert:

$$PPV = \frac{TP}{TP + FP}$$

  Zusätzlich interessant sind noch der positive und negative Vorhersagewert. Auch bei diesen beiden Gütemaßen handelt es sich um bedingte Wahrscheinlichkeiten, wobei bei dem positiven Vorhersagewert (engl.: *positive predictive value or Precision*) die Gesamtheit der als positiv klassifizierten Ergebnisse betrachtet wird und der Anteil der davon richtig positiv klassifizierten Ergebnisse das Gütemaß beschreibt.

- negativer Vorhersagewert:

$$NPV = \frac{TN}{TN + FN}$$

  Demnach erklärt der negative Vorhersagewert (engl.: *negative predictive value*) den Anteil der richtig negativ klassifizierten Objekte an allen negativ klassifizierten Objekten.

Bei einem binären Klassifikationsproblem bietet sich zudem die Betrachtung der Fläche unter der *Receiver-Operator-Characteristic*(ROC)-Kurve an (Centor, 1991). Die ROC-Kurve wird in einem Diagramm dargestellt, bei welchem die Sensitivität (Richtig-Positiv-Rate) auf der Abszisse und die Falsch-Positiv-Rate auf der Ordinate abgebildet wird. Die Kurve wird nun beschrieben durch die jeweilige Wahl des Schwellenwertes der Klassenwahrscheinlichkeit. Es werden also die Sensitivität und die Falsch-Positiv-Rate für alle möglichen Schwellenwerte der Klassenwahrscheinlichkeit abgebildet, ohne dass der

Schwellenwert selbst in dem Diagramm auftaucht. In Abbildung 2.4 ist ein Beispiel einer ROC-Kurve dargestellt. Bildet die Kurve eine lineare Gerade zwischen den Punkten $P1(0|0)$ und $P2(1|1)$, so ist davon auszugehen, dass es sich um einen Verfahren handelt, welches die Beobachtungen zufällig klassifiziert. Ist die Kurve anfangs stark steigend und verläuft danach nahe der oberen Grenze des Diagramms nach rechts, spricht dies für ein Verfahren mit sehr guter Genauigkeit. Verläuft demnach die Kurve unterhalb der Geraden, handelt es sich um ein Verfahren, welches schlechter als ein Zufallsverfahren prognostiziert bzw. den Zusammenhang zwischen Einflussvariablen und der Zielvariablen falsch interpretiert.

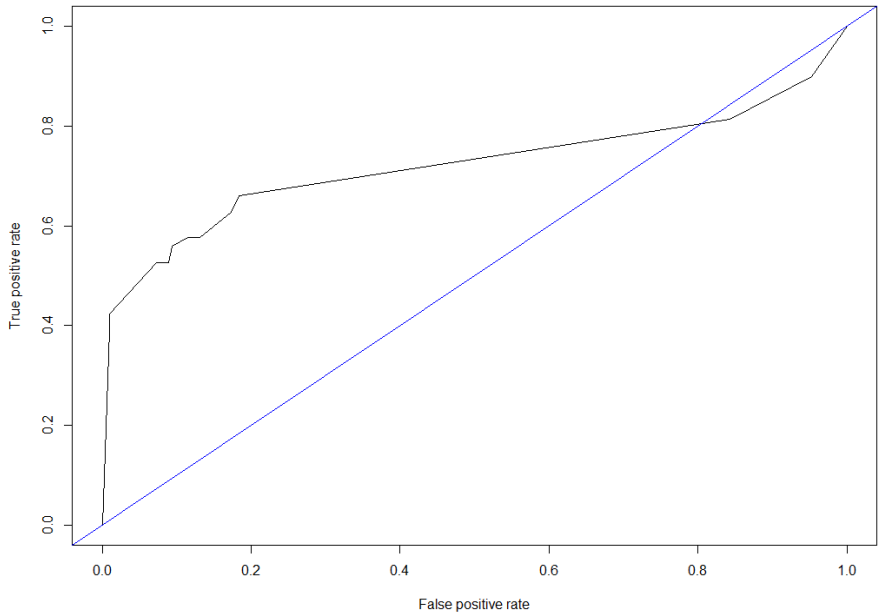
Die sogenannte Area under the Curve (AUC), also Fläche unter der Kurve, beschreibt den Flächeninhalt unter der ROC-Kurve. Dies ist ein weiteres Gütemaß, welches für die Bewertung der Modelle berücksichtigt wird. Auch bei der AUC ist ein Wert nahe 1 sehr gut, da dies bedeutet, dass das Modell die Klassen genau auseinander halten kann. Ein Wert von 0.5 wiederum entspricht dem schlechtesten Wert, da dies bei einer Geraden zwischen den Punkten $P1(0|0)$ und $P2(1|1)$ der Fall ist. Dies bedeutet, dass das Modell keinen Zusammenhang erkennt und zufällig klassifiziert. Ein AUC von unter 0.5 bedeutet, dass das Modell den Zusammenhang in die falsche Richtung erkennt. Negiert man dies, erzielt man einen Wert über 0.5.

Wichtig ist zu erwähnen, dass es bei sehr kleinen Stichproben dazu kommen kann, dass die Bildung der ROC-Kurve nicht möglich ist. Denn bei einem kleinen Testdatensatz, welcher bei einem gesamten Stichprobenumfang $n = 100$ lediglich $n_{test} = 25$ Werte enthält, kann es vorkommen, dass diese alle dieselbe Ausprägung haben. Sofern also alle Testdatensatzeinträge entweder der einen oder anderen Klasse angehören kann entweder keine Sensitivität oder keine Falsch-Positiv-Rate berechnet werden. Dies führt zu dem Problem, dass keine ROC-Kurve für diesen Testdatensatz gebaut werden kann. Im Verlauf dieser Arbeit wurde für diesen Eintrag keine AUC berechnet und dementsprechend nicht berücksichtigt.

2.5 Datensatz

Bei der Auswahl des Datensatzes für diese Arbeit war insbesondere die Größe ($n >= 1000$) und die Verteilung der Klassen wichtig. Eine unausgewogene Klassenverteilung bringt weitaus mehr Herausforderung mit sich über welche schon viele andere Arbeiten geschrieben wurden. Daher spielt auch diese eine wichtige Rolle, so kann der Fokus

Abbildung 2.4: ROC Kurve Beispiel



der Arbeit auf den unterschiedlich großen Stichprobengrößen liegen und die dadurch entstehenden Unterschiede lassen sich besser differenzieren. Der für diese Arbeit genutzte Datensatz *Rain in Australia* stammt von Kaggle. Er beinhaltet insgesamt 142193 Einträge mit 24 unterschiedlichen Variablen. Bei den Daten handelt es sich um Wetterdaten, welche täglich an 49 unterschiedlichen Wetterstationen in Australien im Zeitraum vom 01.11.2007 bis 25.06.2017 gemessen wurden. Bei den 24 Variablen handelt es sich um 23 Einflussvariablen. Diese lassen sich unterteilen in fünf qualitative Variablen und 19 quantitative Variablen. Im folgenden werden diese kurz vorgestellt:

qualitative Variablen:

- *Location*: Ort der Messstation. 49 unterschiedlich Wetterstationen aus ganz Australien. Die wenigsten Einträge aus Nhil ($min = 1569$) und die meisten aus Canberra ($max = 3418$). Durchschnittlich liegen $\bar{x} = 2902$ Einträge pro Station vor.
- *WindGustDir, WindGustDir9am, WindGustDir3pm*: Windrichtungen aus welche zum einen die stärkste Böe des Tages (*WindGustDir*), aber auch die Windrichtungen zu den Zeitpunkten 9 Uhr (*WindGustDir9am*) und 15 Uhr (WindGustDir3pm). Berücksichtigt wurden dabei 16 Kompassrichtungen. Die stärkste WindBöe konnte mit $max = 9780$ aus Westen gemessen werden, während dies um 9 Uhr mit $max = 11393$ aus Norden und um 15 Uhr mit $max = 10663$ aus Südosten der Fall war. Die Richtungen aus denen der Wind am seltensten kam war Ostsüdosten (*WindGust-*

Dir, $min = 7305$), Westsüdwesten (*WindDir9am*, $min = 6843$) und Nordnordosten (*WindDir3pm*, $min = 6444$). Die Anzahl fehlender Messungen belaufen sich auf 9330, 10013 und 3778.

- *RainToday*: binäre Variable, welche entweder *Yes* oder *No* annimmt und lediglich aussagt ob es an dem Tag regnete oder nicht. Die Ausprägung *Yes* liegt dabei 109332 mal vor während *No* 31455 mal vorkommt. Es liegen 1406 fehlende Einträge vor.

quantitative Variablen:

- *Date*: Das Datum der jeweiligen Messung, wobei der Gesamtzeitraum zwischen dem 01.11.2007 und dem 25.06.2017 liegt.

- *minTemp*: die niedrigste Temperatur des Tages. Bei dieser Variablen liegen 637 fehlende Werte vor. Die durschnittliche gemessene Niedgristtemperatur beträgt $\bar{x} = 12.19$ Grad Celsius, während die maximale Tiefsttemperatur bei $max = 33.9$ Grad und die minimale bei $min = -8.5$ liegt.

- *maxTemp*: die höchste gemessene Temperatur des Tages. Es liegen dort 322 fehlende Werte vor. Die durchschnittliche Höchstemperatur liegt bei $\bar{x} = 23.23$ Grad, die minimale bei $min = -4.8$ und die maximal bei $max = 48.1$ Grad.

- *Rainfall*: Der Niederschlag an einem Tag. Es fehlen hier Werte bei 1406 Einträgen. Der durchschnittlich gemessene Niederschlag beläuft sich auf $\bar{x} = 2.35$ mm, der minimale auf $min = 0.0$ mm und der maximale auf $max = 371.0$ mm.

- *Evaporation*: Die Evaporation an einem Tag. Die Anzahl der fehleden Werte liegt bei 60843. Der durschnittliche Wert ist $\bar{x} = 5.47$ mm, der minimale $min = 0.0$ mm und der maximale $max = 145.00$ mm.

- *Sunshine*: Die Anzahl der Stunden, in welche die Sonne an dem Tag geschienen hat. Werte fehlen bei 67816 Einträgen. Die durchschnittliche Zeit liegt bei $\bar{x} = 7.62$ Stunden, die minimale bei $min = 0.0$ Stunden und die maximale bei $max = 14.5$.

- *WindGustSpeed*: Die Geschwindigkeit der schnellsten Böe des Tages. Es fehlen Werte in Höhe von 9270. Das arithmetische Mittel der Geschwindikeit der schnellsten Böe liegt bei $\bar{x} = 39.98$ km/h. Die kleinste gemessene Geschwindikgeit beträgt $min = 6.0$ km/h und die höchste $max = 135$ km/h.

- *WindSpeed9am*: Die Windgeschwindigkeit um 9 Uhr am Vormittag. Es liegen fehlende Werte bei 1348 Einträgen vor. Am Vormittag konnte die durchschnittliche Windgeschwindigkeit auf $\bar{x} = 14.00$ km/h gemessen werden. Der langsamste Wind wehte mit $min = 0$ km/h und der schnellste mit $max = 130$ km/h.

- *WindSpeed3pm*: Die Wingeschwindigkeit um 3 Uhr am Nachmittag. Fehlende Werte liegen in Höhe von 2630 vor. Die durschnittliche Geschwindigkeit zu dieser Zeit beträgt $\bar{x} = 18.64$ km/h. Der kleinste Wert lag auch hier bei $min = 0$ km/h und der größte bei $max = 87.0$ km/h.

- *Humidity9am*: Die Luftfeuchtigkeit um 9 Uhr am Vormittag. Es gibt 1774 fehlende Werte. Die durschnittle Luftfeuchtigkeit um diese Zeit wurde auf $\bar{x} = 68.84$% gemessen. Der kleinste gemessene Wert der Luftfeuchtigkeit liegt bei $min = 0$% und der höchste bei $max = 100$%.

- *Humidity3pm*: Die Luftfeuchtigkeit um 3 Uhr am Nachmittag. Hier beträgt die Anzahl der fehlenden Werte 3610. Am nachmittag wurde der durchschnittliche Wert der Luftfeuchtigkeit auf $\bar{x} = 51.48$% gemessen. $min = 0$% ist auch hier der kleinste Wert und dementsprechend auch $max = 100$% der höchste Wert.

- *Pressure9am*: Der Luftdruck um 9 Uhr am Vormittag. Die Fehlwerte belaufen sich auf 14014. Durschnittlich lag der Luftdruck um 9 Uhr bei $\bar{x} = 1017.65$ Hektopascal. Der Minimalwert beträgt $min = 980.5$ und der Maximalwert $max = 1041.0$ Hektopascal.

- *Pressure3pm*: Der Luftdruck um 3 Uhr am Nachmittag. Es liegen 13981 fehlende Werte vor. Der durschnittliche Wert für den Luftdruck liegt bei 3 Uhr am nachmittag bei $\bar{x} = 1015.26$ Hektopascal. Der niedrigste gemessene Luftdruck beläuft sich auf $min = 977.1$ Hektopascal und der höchste auf $max = 1039.6$.

- *Cloud9am*: Die Bewölkung um 9 Uhr am Vormittag. Die Werte liegen dabei auf einer Skala von 0 bis 9, wobei 0 gar nicht bewölkt bedeutet und 9 voll bewölkt. Die Anzahl der fehlenden Werte liegt bei 53657. Durchschnittlich lag die Bewölkung bei $\bar{x} = 4.44$ auf der Skala. Um 9 Uhr liegt der Wert für die niedrigste Bewölkung bei $min = 0$ und für die höchste bei $max = 9$.

- *Cloud3pm*: Die Bewölkung um 3 Uhr am Nachmittag. Die Skala ist dabei die gleich wie bei Cloud9am. Es fehlen 57094 Werte. $\bar{x} = 4.50$ ist durchschnittle Bewölkung zu dieser Zeit. Die geringste Bewölkung liegt bei $min = 0$ und die größte bei $max = 9$.

- *Temp9am*: Die Temperatur um 9 Uhr am Vormittag. Es gibt fehlende Werte in Höhe von 904. Der Mittelwert der um diese Zeit gemessenen Temperatur ist $\bar{x} = 16.99$ Grad. Der Minimalwert liegt bei $min = -7.2$ Grad und der Höchstwert bei $max = 40.2$ Grad Celsius.

- *Temp3pm*: Die Temperatur um 3 Uhr am Nachmittag. Hier fehlen Werte in Höhe von 2726. Die durchschnittliche Temperatur welche zu dieser Zeit gemessen wurde, liegt bei $\bar{x} = 21.69$. Die minimale gemessene Temperatur beträgt $min = -5.4$ Grad und die maximale $max = 46.7$ Grad Celsius.

- *RISK_MM*: Der Niederschlag des Folgetags. Diese Variable ist lediglich der *Rainfall* des nächsten Tages und wurde genutzt um die Zielvariable zu bestimmen.

Bei der **Zielvariablen** handelt es sich um *RainTomorrow*, welche wie *RainToday* ebenfalls binär nur *Yes* oder *No* annehmen kann. Es liegen dabei 110316 Fälle mit der Ausprägung *No* vor und 31877 Fälle mit der Ausprägung *Yes*. Dabei ist es das Ziel der Anwendungen des Maschinellen Lernens Entscheidungsbaum und Random Forest diese nach dem Training bei dem Testdatensatz richtig zu bestimmen. Die gesamten Variablen wurde auch nochmal in Tabelle 2.2 übersichtlich dargestellt.

Da die Betrachtung von unterschiedlichen Wetterstationen zusammen wenig sinnvoll erscheint, werden in der Arbeit nur die Daten des Ortes Sydney berücksichtigt. Da die Einflussvariablen *WindGustSpeed* und *WindGustDir* erst ab dem 01.07.2012 in Sydney gemessen wurden, werden die Daten davor nicht weiter berücksichtigt. Übrige fehlende Werte liegen nur bei metrischen Variablen vor und können somit durch den arithmetischen Mittelwert der Spalte ersetzt werden. Die Spalten *Location, Date* und *RISK_MM* werden für das Training der Algorithmen nicht herbeigezogen. *Location* und *Date* haben keinen Einfluss auf die Zielvariable, während *RISK_MM* zu stark mit der Zielvariable korreliert, da diese 0 annimmt, wenn es am Folgetag nicht regnet. So bleibt ein Datensatz der Größe $n = 1701$ mit 20 Einflussvariablen. Es liegen keine fehlenden Werte in dem Datensatz Sydney vor. Im Folgenden werden noch einmal die **qualitiativen Variablen** dieses Datensatzes kurz beschrieben:

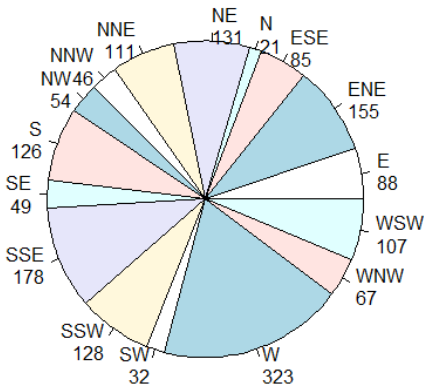
- *WindGust*: Die häufigste Richtung aus welcher die stärkste Böe in Sydney kam ist Westen mit $max = 323$ Einträgen. Nur $min = 21$ mal kam der Wind aus dem Norden. In Abbildung 2.5 sind die Daten grafisch dargestellt.

- *WindDir9am*: Die Windrichtung um 9 Uhr vormittags war ebenfalls am häufigsten Westen mit $max = 599$. $min = 31$ mal kam der Wind aus dem Südwesten.

Spalte		Bedeutung	Einheit
Location		Ort der Messstation	ausgeschrieben
Date		Tag des Monats	dd.mm.jjjj
Temp	Min	Minimale Temperatur in den letzten 24 Std. bis 09:00 Uhr	Grad Celsius
	Max	Maximale Temperatur in den letzten 24 Std. bis 09:00 Uhr	Grad Celsius
Rainfall		Niederschlag in den letzten 24 Std. bis 09:00 Uhr	Millimeter
Evaporation		Class A Pan Evaporation in den letzten 24 Stunden bis 09:00 Uhr	Millimeter
Sunshine		Heller Sonnenschein in den letzten 24 Stunden bis Mitternacht	Stunden
WindGust	Dir	Richtung der stärksten Windböe der letzten 24 Stunden bis Mitternacht	Kompass Richtungen
	Speed	Geschwindigkeit der stärksten Böe in den letzten 24 Stunden bis Mitternacht	Kilometer pro Stunde
9am	Temp	Temperatur um 9 Uhr	Grad Celsius
	humidity	Relative Feuchtigkeit um 9 Uhr	Prozent
	Cloud	Bruchteil des Himmels, welcher um 9 Uhr bedeckt ist	Achtel
	Dir	durschnittliche Windrichtung in den 10 Minuten vor 9 Uhr	Kompass Richtungen
	Speed	durchschnittliche Windgeschwindigkeit in den 10 Minuten vor 9 Uhr	Kilometer pro Stunde
	Pressure	Luftdruck reduziert auf durchschnittliche Meereshöhe um 9 Uhr	Hektopascal
15 Uhr	Temp	Temperatur um 15 Uhr	Grad Celsius
	humidity	Relative Feuchtigkeit um 15 Uhr	Prozent
	Cloud	Bruchteil des Himmels, welcher um 15 Uhr bedeckt ist	Achtel
	Dir	durschnittliche Windrichtung in den 10 Minuten vor 15 Uhr	Kompass Richtungen
	Speed	durchschnittliche Windgeschwindigkeit in den 10 Minuten vor 15 Uhr	Kilometer pro Stunde
	Pressure	Luftdruck reduziert auf durchschnittliche Meereshöhe um 15 Uhr	Hektopascal
RainToday		Fall ob es heute geregnet hat oder nicht	binär ja/nein
RISK_MM		Niederschlag des Folgetages	mm

Tabelle 2.2: Legende Datensatz Rain in Australia

Abbildung 2.5: Windrichtungen der stärksten Böe (Sydney)



- *WindDir3pm*: Zum Nachmittag hin scheint der Wind sich zu drehen, denn hier kommt er am häufigsten aus der Richtung Osten ($max = 282$). Am seltensten weht er zu dieser Zeit aus Richung Nordnordwest ($min = 19$).

- *RainToday*: Der Regen heute besitzt 1271 Ausprägungen No und 430 Yes.

- *RainTomorrow*: Die Zielvariable, ob es morgen regnet oder nicht, hat 1273 Ausprägungen Yes und 428 mal No. Die Diagramme bzw. Abbildungen für diese Variable und die drei vorherigen sind im Anhang zu finden.

Die quantitativen Werte sind übersichtlich in der Tabelle 2.3 zusammengefasst. Es sind das arithmetische Mittel (*mean*), der kleinste Wert (*min*), der größte Wert (*max*) und die Standardabweichung (*sd*) aufgeführt. Zusätzlich sind noch die Boxplots einiger Variablen in den Abbildungen 2.6 bis 2.11 dargestellt. Abbildungen der anderen Variablen sind im Anhang vorzufinden.

Name	mean	min	max	sd
MinTemp	14.96	5.00	27.10	4.54
MaxTemp	23.40	11.70	45.80	4.53
Rainfall	3.31	0.00	119.40	10.09
Evaporation	5.37	0.00	18.40	2.83
Sunshine	7.34	0.00	13.60	3.74
WindGustSpeed	41.79	17.00	96.00	12.98
WindSpeed9am	15.21	2.00	54.00	6.89
WindSpeed3pm	19.67	2.00	57.00	7.45
Humidity9am	66.38	19.00	100.00	15.19
Humidity3pm	52.98	10.00	96.00	16.08
Pressure9am	1018.53	996.50	1039.00	7.06
Pressure3pm	1016.11	994.00	1036.00	7.09
Cloud9am	4.21	0.00	8.00	2.71
Cloud3pm	4.16	0.00	8.00	2.62
Temp9am	18.04	6.70	36.50	4.93
Temp3pm	21.84	11.00	44.70	4.30

Tabelle 2.3: Zusammenfassung deskriptiver Statistik Datensatz Sydney

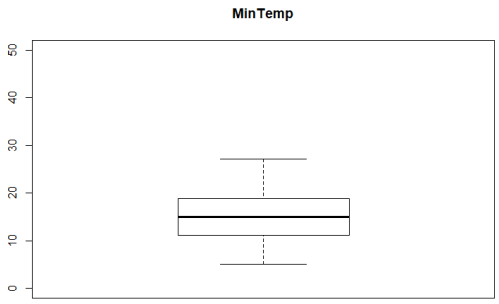


Abbildung 2.6: Boxplot der Variable MinTemp

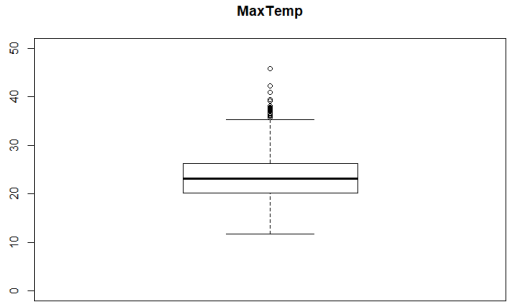


Abbildung 2.7: Boxplot der Variable MaxTemp

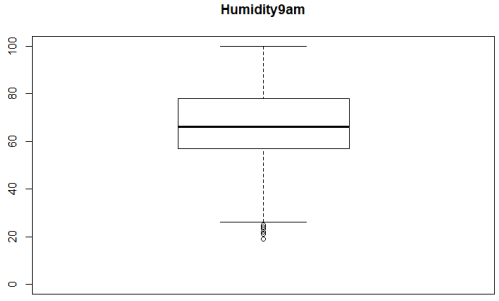


Abbildung 2.8: Boxplot der Variable Humidity9am

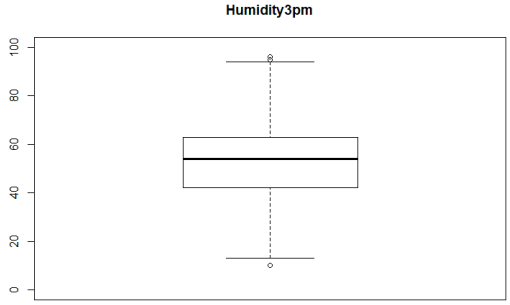


Abbildung 2.9: Boxplot der Variable Humidity3pm

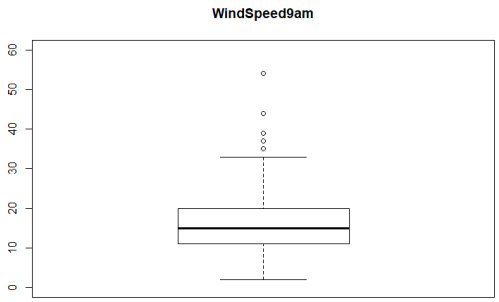


Abbildung 2.10: Boxplot der Variable WindSpeed9am

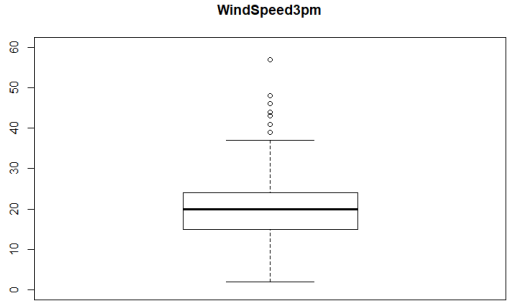


Abbildung 2.11: Boxplot der Variable WindSpeed3pm

3. Ergebnisse

3.1 Gütemaße der verschiedenen Fallzahlen

Die in Tabelle 3.1 und in den Abbildungen darstellten Ergebnisse, sind die Werte der 100-fach durchgeführten Experimente. Dabei wurden diese jeweils auf drei unterschiedlichen Fallzahlen ($n = 1000, 400$ und 100) durchgeführt. Bei den Boxplots in den Abbildungen 3.1 bis 3.6 handelt es sich bei den linken drei Boxen jeweils um die Ergebnisse der Gütemaße der unterschiedlichen Fallzahlen des Entscheidungsbaums, bei den rechten um die des Random Forests. In Tabelle 3.1 sind die jeweiligen Mittelwerte der Gütemaße abgedruckt, wobei der höchste Wert dabei fett markiert ist. Dabei fällt auf, dass zum einen alle Höchstwerte des jeweiligen Gütemaßes von dem Random Forest erzielt wurden und zum anderen entgegen der Erwartung die besten Ergebnisse bei vier von den sechs Gütemaßen bei einer Fallzahl von $n = 400$ vorliegen. Lediglich die Sensitivität weist den höchsten Wert auch bei der höchsten Fallzahl $n = 1000$ auf. Bei der Area under the Curve, dem Flächeninhalt unter der ROC Kurve, ist der höchste Wert sogar bei $n = 100$ aufzufinden.

Genauer erkennt man die Differenzen und Unterschiede in den Boxplots. So ist in dem Boxplot in Abbildung 3.1 zu erkennen, dass sie sowohl bei den Entscheidungsbäumen als auch bei den Random Forests von links nacht rechts größer werden. Dies spricht für eine größere Streuung, welche auch durch die Länge der Whiskers, den Linien über und unter den Boxen, und den Ausreißern, den schwarzen Punkten, zu erkennen ist. Besonders auffällig ist allerdings, dass für beide Anwendungen der Boxplot für $n = 400$ höher liegt und somit auf eine bessere Korrektklassifikationsrate hindeutet als für $n = 1000$. Zwar ist die Box etwas größer, aber der Median liegt jeweils deutlich über dem der dazugehörigen größeren Fallzahl. Vergleicht man die Boxplots des Entscheidungsbaumes mit dem des Random Forests, ist zu erkennen, dass der Random Forest ein wenig besser klassifizieren konnte als der Entscheidungsbaum. Die Mediane und auch das arithmetische Mittel

aus Tabelle 3.1 liegen alle etwas höher.

In der Abbildung 3.2 ist wiederum ein ganz anderes Bild zu erkennen. Vor allem auffällig ist, dass die Werte der Mediane alle um $Sens = 0.5$ liegen, welches deutlich niedriger als die Werte der Korrektklassifikationsrate ist. Die Streuung und damit die Größe der Box steigt stark mit Verlust der Fallzahl, ebenso der Median. Für $n = 100$ gehen die Whiskers sogar von 0 bis 1. Das bedeutet, dass das Gütemaß Sensitivität stark bei geringer Fallzahl variiert. Auch bei der Sensitivität ist festzustellen, dass sowohl die Boxen als auch die arithmetischen Mittelwerte aus 3.1 bei dem Random Forest besser als bei dem Entscheidungsbaum sind.

In Abbildung 3.3 sind die unterschiedlichen Boxplots für den positiven Vorhersagewert dargestellt. Es sind ziemlich deutliche Unterschiede zwischen den Fallzahlen aber auch den unterschiedlichen Anwendungen Entscheidungsbaum und Random Forest zu erkennen. Bei Betrachtung der drei Boxplots des Entscheidungsbaums ist sehr gut zu erkennen, dass zum einen die Streuung größer wird, insbesondere von $n = 1000$ auf $n = 400$. Der Unterschied zwischen $n = 400$ und $n = 100$ besteht weniger in der Größe der Varianz als in dem Median. Fast die ganze Box für $n = 100$ liegt unter der für $n = 400$, was schließen lässt, dass der positive Vorhersagewert stark nachlässt bei einem sehr kleinen Datensatz. Auch bei diesem Gütemaß liegen sowohl die Boxen des Random Forest höher als auch das arithmetische Mittel, wie in Tabelle 3.1 zu erkennen. Bei den Boxplots fällt zudem auf, dass der Median für $n = 400$ bei Random Forests weiter oben liegt als bei $n = 1000$. Bei $n = 100$ ist er deutlich niedriger und streut sehr stark.

Wie in Abbildung 3.4 zu erkennen, sind die Werte durchweg sehr hoch. Sowohl bei den Entscheidungsbäumen als auch bei den Random Forests nimmt die Streuung mit kleinerer Fallzahl zu. Die Mediane verhalten sich nicht ganz nach den Erwartungen, denn in beiden Fällen liegen diese höher bei $n = 400$ als bei $n = 1000$. Bei den Random Forest liegt sogar der bei $n = 100$ über dem bei $n = 1000$. Zwischen Entscheidungsbaum und Random Forest ist auch wieder erkennbar, dass der Random Forest bessere Spezifität über alle Fallzahlen erzielt. Auch die Streuung ist weniger stark als bei den Entscheidungsbäumen.

Ein sehr unübliches Bild ist in Abbildung 3.5 zu erkennen. Denn mit Ausnahme für den Entscheidungsbaum $n = 100$ Boxplot steigen die Mediane mit kleiner werdenden Datensätzen. Allerdings wird die Streuung größer und die Ausreißer mehr, was auch dazu führt, dass das arithmetische Mittel bei Random Forest $n = 100$ kleiner ist als bei $n = 400$, obwohl der Median größer ist. Dies ist auf die Resistenz gegen Ausreißer des

Medians zurückzuführen. Beim Vergleich zwischen Entscheidungsbaum und Random Forest ist der Random Forest etwas besser, jedoch nicht so eindeutig wie bei vorherigen Gütemaßen.

Als letztes Gütemaß in Abbildung 3.6 lässt die Area under the Curve einen deutlichen Unterschied zwischen den Boxplots der Entscheidungsbäume und Random Forests erkennen. Bei den Boxplots der Entscheidungsbäume ist eine klare Tendenz zu erkennen, welche den Erwartungen entspricht, dass mit abnehmender Fallzahl die Varianz und der Median bzw. auch das arithmetische Mittel sinken. Die Boxplots des Random Forests wiederum unterscheiden sich in dem Maße, dass die Streuung bei weitem nicht so stark steigt und die Area under the Curve bei $n = 100$ deutlich besser ist als bei $n = 1000$ und $n = 400$.

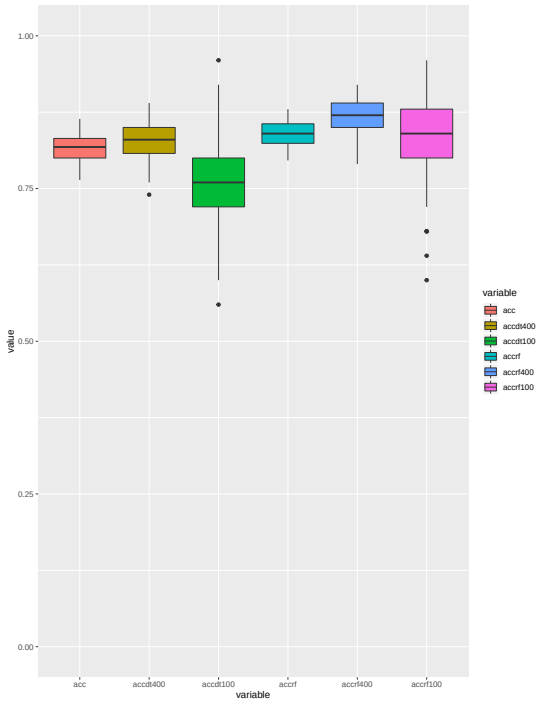


Abbildung 3.1: Korrektklassifikationsrate

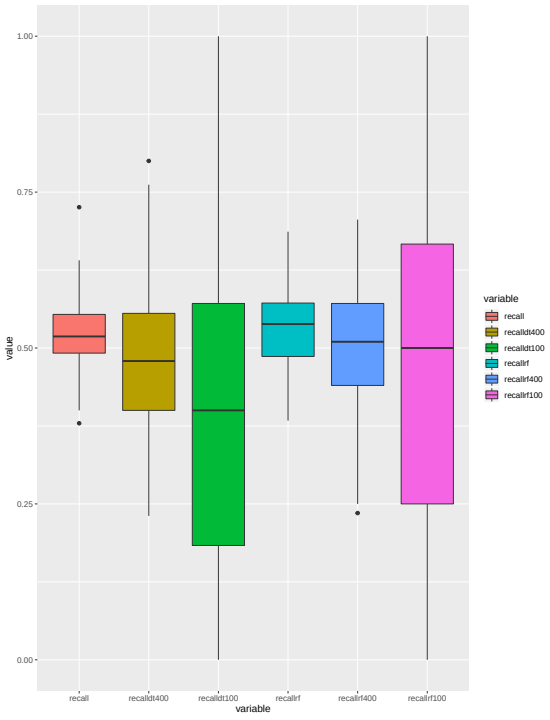


Abbildung 3.2: Sensitivität

Boxplots der Gütemaße für Entscheidungsbäume und Random Forests ($n = 1000, 400$ und 100)

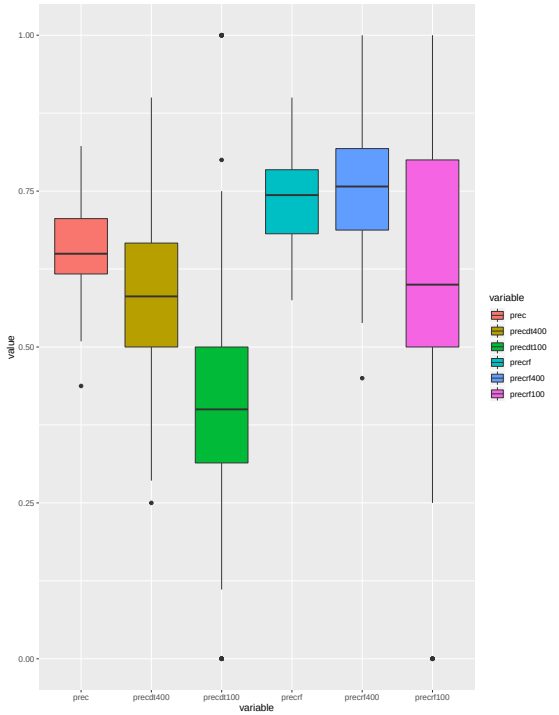


Abbildung 3.3: positiver Vorhersagewert

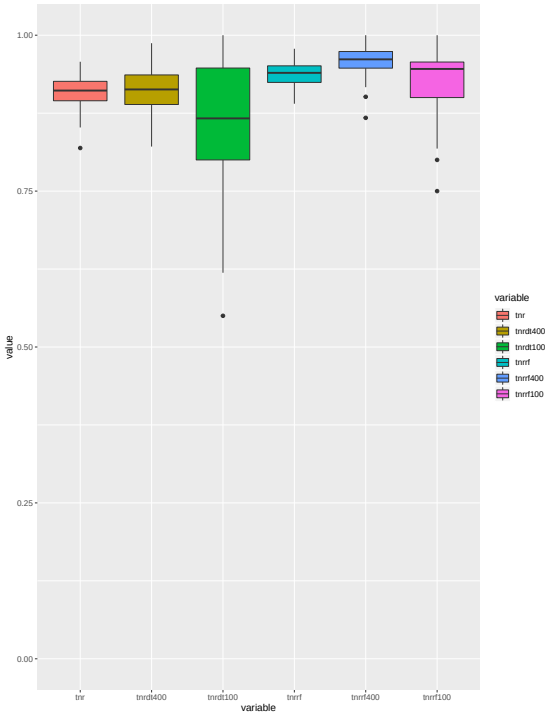


Abbildung 3.4: Spezifität

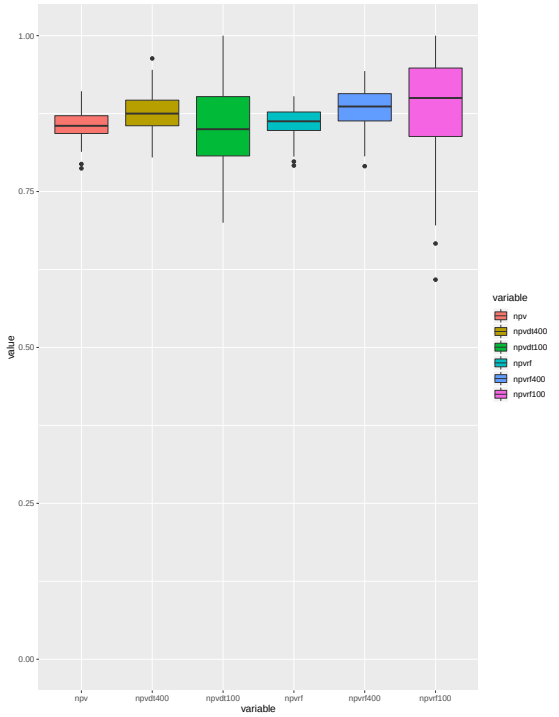


Abbildung 3.5: Negativer Vorhersagewert

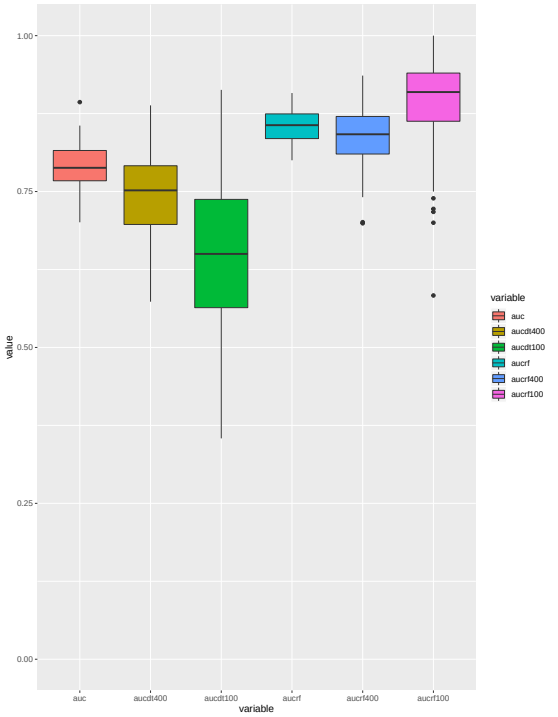


Abbildung 3.6: Area under the curve

Boxplots der Gütemaße für Entscheidungsbäume und Random Forests ($n = 1000, 400$ und 100)

In Tabelle 3.2 sind die Standardabweichung der Gütemaße für die unterschiedlichen Fallzahlen und die beiden Anwendungen Entscheidungsbaum und Random Forest dargestellt. Die kleinsten Werte, was bei der Standardabweichung die geringste Streuung der Daten bedeutet, sind fett markiert. Auffällig ist, dass diese Werte nur bei der Fall-

Gütemaß	n = 1000		n = 400		n = 100	
	dt	rf	dt	rf	dt	rf
Acc	0.8159	0.8391	0.8253	**0.8643**	0.7672	0.8400
Sens	0.5219	**0.5330**	0.4815	0.5041	0.4026	0.4633
PPV	0.6549	0.7365	0.5800	**0.7502**	0.4324	0.6234
Spec	0.9105	0.9382	0.9112	**0.9568**	0.8611	0.9335
NPV	0.8558	0.8616	0.8766	**0.8831**	0.8574	0.8796
AUC	0.7883	0.8569	0.7453	0.8410	0.6518	**0.8904**

Tabelle 3.1: Gütemaße Mittelwert

Gütemaß	n = 1000		n = 400		n = 100	
	dt	rf	dt	rf	dt	rf
Acc	0.0224	**0.0210**	0.0321	0.0283	0.0681	0.0715
Sens	**0.0555**	0.0566	0.1051	0.0897	0.2932	0.2542
PPV	0.0702	**0.0692**	0.1276	0.1061	0.2651	0.2801
Spec	0.0251	**0.0187**	0.0366	0.0220	0.1024	0.0534
NPV	**0.0224**	0.0228	0.0319	0.0316	0.0741	0.0774
AUC	0.0372	**0.0253**	0.0661	0.0473	0.1228	0.0765

Tabelle 3.2: Gütemaße Standardabweichung

zahl $n = 1000$ liegen, was darauf zurückschließen lässt, dass der größere Datensatz für ein besseres Training der Modelle und damit einhergehende konstantere Testergebnisse sorgt. Vier der kleinsten Standardabweichungen der Gütemaße sind dem Random Forest zuzuordnen und bei den zwei Ausnahmen, wo der Entscheidungsbaum bessere Werte erzielt, liegen diese nur knapp unter denen des Random Forests. Vergleicht man die Werte zwischen Random Forest und Entscheidungsbaum für die anderen Fallzahlen $n = 400$ und $n = 100$ ist dort ein sehr ähnliches Muster zu erkennen. Bis auf positiver Vorhersagewert (PPV) und negativer Vorhersagewert (NVP) bei jeweils $n = 100$ erzielte der Random Forest durchweg bessere Werte bei der Standardabweichung der Gütemaße. Bei dem negativen Vorhersagewert ist der Unterschied minimal, während die Standardabweichung bei dem positiven Vorhersagewert für den Entscheidungsbaum deutlich besser ist.

Vergleicht man die unterschiedlichen Fallzahlen miteinander ist klar zu erkennen, dass bei sinkender Fallzahl auch die Standardabweichung der Gütemaße steigt. Dabei gibt es keine Ausnahme.

3.2 Varianzkomponenten der Modelle

3.2.1 Standardabweichung des Mittelwerts

In der Tabelle 3.3 sind die Standardabweichungen der Mittelwerte der Gütemaße dargestellt. Hierbei handelt es sich bei der Standardabweichung um die Standardabweichung von den zehn arithmetischen Mitteln der jeweils 100-fach ausgeführten Anwendungen Entscheidungsbaum und Random Forest. Dabei unterscheiden sich bei den 100 Modellen lediglich die Aufteilung des Trainings- und Testdatensatzes. Der Ursprungsdatensatz wurde nur zwischen den zehn Durchführungen zufällig gewählt, um so zu überprüfen, wie sehr die Ergebnisse sich differenzieren. Dafür wurde von diesen Mittelwerten die Standardabweichung berücksichtigt.

In Tabelle 3.3 sind die niedrigsten Werte fett markiert, da diese die geringste Streuung und somit besten Werte darstellen. Zu erkennen ist, dass dies eindeutig bei dem Random Forest Algorithmus auf der $n = 1000$ großen Stichprobe der Fall ist. Bei allen sechs Gütemaßen liefert dies den besten Wert. Vergleicht man die Anwendungen Random Forest und Entscheidungsbaum bei den anderen Fallzahlen ist es nicht mehr so eindeutig. Bei $n = 400$ liefert der Random Forest bessere Werte bei der Korrektklassifikationsrate, dem positiven Vorhersagewert, der Spezifität und der Area under the Curve, während der Entscheidungsbaum für die Gütemaße Sensitivität und negativer Vorhersagewert bessere Ergebnisse erzielt. Für die Fallzahl $n = 100$ liegt der Entscheidungsbaum Algorithmus sogar bei vier Gütemaßen besser: Korrektklassifikationsrate, Sensitivität, positiver und negativer Vorhersagewert. Dementsprechend besser Standardabweichung der Mittelwerte liegen dem Random Forest bei Spezifität und Area under the Curve vor.

Bei Betrachtung der unterschiedlichen Fallgrößen ist ein eindeutiger Trend zu erkennen. Je kleiner die Stichprobe ist, desto größer wird die Standardabweichung des Mittelwerts. Einzige Ausnahme ist der Wert des Entscheidungsbaumes bei der Fallzahl $n = 400$ und dem Gütemaß negativer Vorhersagewert. Hier liegt die Standardabweichung des Mittelwertes mit 0.21% niedriger als 0.26% bei $n = 1000$.

3.2.2 Mittelwert der Standardabweichung

Die Mittelwerte der Standardabweichung sind in Tabelle 3.4 zu erkennen. Wie in dem Unterkapitel 3.2.1 handelt es sich bei dem Mittelwert um das arithmetische Mittel der zehn Standardabweichungen, welche bei dem Verfahren gebildet wurden, welches in

Gütemaß	n = 1000		n = 400		n = 100	
	dt	rf	dt	rf	dt	rf
Acc	0.0022	**0.0013**	0.0039	0.0027	0.0064	0.0075
Sens	0.0075	**0.0073**	0.0084	0.0089	0.0224	0.0296
PPV	0.0082	**0.0053**	0.0142	0.0141	0.0209	0.0275
Spec	0.0033	**0.0018**	0.0043	0.0021	0.0083	0.0036
NPV	0.0026	**0.0018**	0.0021	0.0030	0.0067	0.0091
AUC	0.0045	**0.0021**	0.0076	0.0028	0.0082	0.0075

Tabelle 3.3: Gütemaße Standardabweichung des Mittelwerts

Gütemaß	n = 1000		n = 400		n = 100	
	dt	rf	dt	rf	dt	rf
Acc	0.0241	**0.0206**	0.0353	0.0291	0.0691	0.0653
Sens	0.0653	**0.0590**	0.1115	0.1026	0.2598	0.2452
PPV	0.0757	**0.0670**	0.1408	0.1187	0.2263	0.2592
Spec	0.0276	**0.0191**	0.0405	0.0234	0.0900	0.0531
NPV	0.0237	**0.0230**	0.0337	0.0304	0.0730	0.0728
AUC	0.0378	**0.0240**	0.0715	0.0478	0.1140	0.0726

Tabelle 3.4: Gütemaße Mittelwert der Standardabweichung

Kapitel 2.3.2 erläutert wird.

In Tabelle 3.4 sind die niedrigsten Werte fett markiert, da diese für die besten Mittelwerte der Standardabweichung stehen. Je geringer dieser Wert ist, desto weniger streuen die Gütemaße im Durchschnitt. Dabei liegen auch hier die besten Werte bei dem Random Forest und einer Stichprobe im Umfang von $n = 1000$. Deutlich ist auch, dass Random Forest über alle Gütemaße und Stichprobengrößen besser funktioniert, mit Ausnahme des positiven Vorhersagewertes bei der Fallzahl $n = 100$.

Bei Berücksichtigung der Fallzahlen und das Verhalten der Mittelwerte der Standardabweichung im Hinblick auf diese, ist eindeutig festzustellen, dass diese durchweg steigen. Dabei erhöhen sich die Werte der Gütemaße Sensitivität und Positiver Vorhersagewert besonders stark bis auf 25.98% und 22.63% für den Entscheidungsbaum bei einer Stichprobengröße von $n = 100$. Bei dem Random Forest steigen die Werte auf 24.52% für die Sensitivität und 25.92% für den positiven Vorhersagewert.

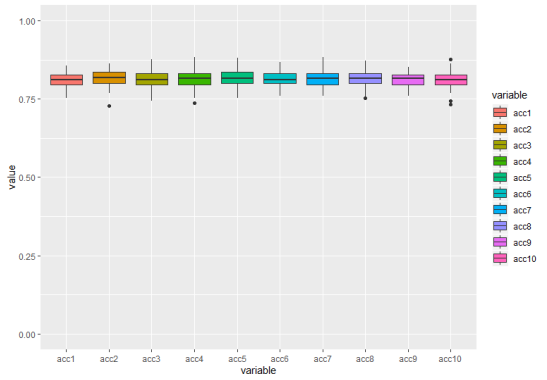


Abbildung 3.7: Entscheidunsbaum $n = 1000$

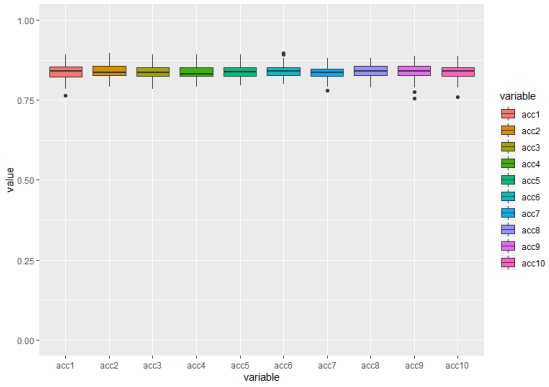


Abbildung 3.8: Random Forest $n = 1000$

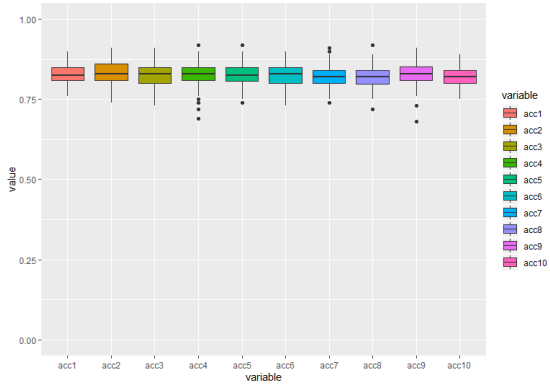


Abbildung 3.9: Entscheidunsbaum $n = 400$

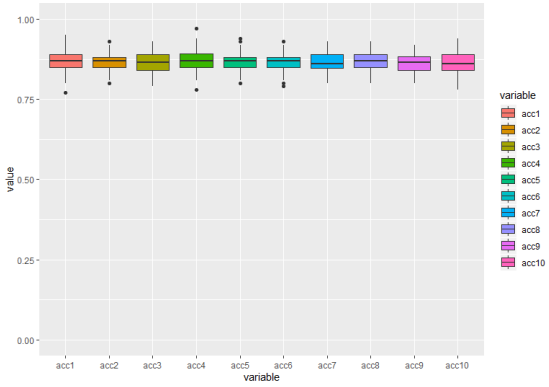


Abbildung 3.10: Random Forest $n = 400$

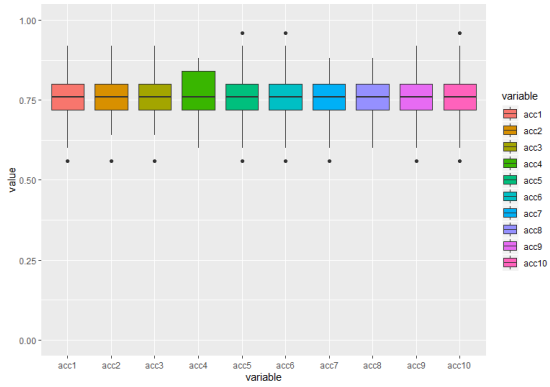


Abbildung 3.11: Entscheidunsbaum $n = 100$

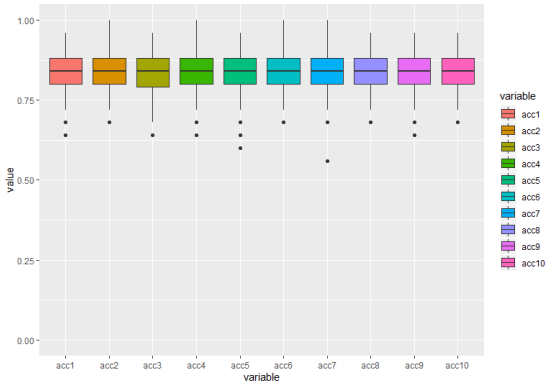


Abbildung 3.12: Random Forest $n = 100$

Boxplots des Gütemaßes Korrektklassifikationsrate bei den Anwendungen Entscheidungsbaum und Random Forests bei unterschiedlichen Fallzahlen ($n = 1000, 400, 100$) zur Untersuchung der Varianzkomponenten

4. Diskussion und Interpretation der Ergebnisse

4.1 Gütemaße der verschiedenen Fallzahlen

In Kapitel 3.1 konnten viele unterschiedlich Ergebnisse in Hinsicht auf die sechs untersuchten Gütemaße gefunden werden. Im Folgenden werde dieser genauer untersucht und diskutiert.

Die Korrektklassifikationsrate verhält sich entgegen den Erwartungen. Die höchsten Werte werden bei $n = 400$ gemessen. Ein größerer Datensatz scheint also nicht immer einherzugehen mit einer besseren Korrektklassifikationsrate bei den Anwendungen Entscheidungsbaum und Random Forest. Eindeutig ist die Standardabweichung bzw. Varianz jedoch zu deuten, denn diese steigt bei abnehmender Fallzahl. Dies bedeutet, dass je kleiner der Datensatz ist, desto mehr streut die Korrektklassifikationsrate der Modelle. Demnach gelten die berechneten Mittelwerte bei kleineren Stichproben als weniger bedeutsam, da oft eine große Streuung der Daten vorliegt.

Bei der Sensitivität ist festzustellen, dass diese den Erwartungen entspricht. Das soll heißen, bei fallender Stichprobengröße sinkt der Median und das arithmetische Mittel, während die Standardabweichung steigt. Es ist allerdings sehr auffällig, dass die Werte der Sensitivität sehr gering ausfallen. Dies lässt darauf schließen, dass beide Anwendungen sich schwertun, bezogen auf den Datensatz, den Regen vorherzusagen. Denn die Sensitivität ist eben genau das Gütemaß, welches angibt mit welcher Rate der tatsächlich stattgefundene Regen am Folgetag richtig durch das Modell vorhergesagt werden konnte (Lalkhen and McCluskey, 2008). Bei einer Sensitivität von $53,30\%$ (Random Forest, $n = 1000$) bedeutet dies also, dass $53,30\%$ der gesamten Folgetage, an denen es regnet, richtig vorhergesagt werden.

Ein weiteres Bild, welches in etwa den Erwartungen entspricht, liefern die Ergebnisse des positiven Vorhersagewertes. Die Standardabweichung steigt mit kleiner werdendem Datensatz und der Median und das arithmetische Mittel, bis auf Ausnahme bei Random Forest $n = 400$, nimmt ab.

Die Spezifität, das Gegenstück zur Sensitivität, ist das Gütemaß mit den höchsten Werten. Dies lässt sich besonders auf die niedrigen Werte der Spezifität zurückführen. Denn die Spezifität ist die Rate mit welcher ein Folgetag, an dem es nicht regnet, richtig klassifiziert wird. Ist also die Sensitivität gering, ist davon auszugehen, dass die Spezifität der Test besser ist und vice versa.

Der negative Vorhersagewert widerspricht den Erwartungen des Verlaufs des Medians und des arithmetischen Mittels, da diese bei fallender Stichprobengröße steigen, mit Ausnahme von Entscheidungsbaum $n = 100$. Aber auch hier steigt die Standardabweichung entsprechend der Erwartungen. Das sowohl die Median und als auch das arithmetische Mittel steigen, bedeutet, dass die Modelle bei kleineren Datensätze die negativen klassifizierten Objekte zu höheren Anteil richtig klassifizieren. Also sind ein größerer Anteil der negativ klassifizierten Objekte richtig negativ klassifiziert.

Als letztes Gütemaß wird die Area under the curve betrachtet. Sie zeigt die stärkste Differenz zwischen den Ergebnissen des Entscheidungsbaumes und des Random Forests. Bei den drei unterschiedlichen Fallzahlen des Entscheidungsbaumes verhält sich das Gütemaß gemäß den Erwartungen. Die Standardabweichung steigt und der Median und das arithmetische Mittel sinken. Die Werte des Random Forests jedoch liegen zum einen deutlich höher und die besten sogar bei $n = 100$. Dies ist ein sehr unerwartetes Verhalten des Gütemaßes.

Durchweg ist festzustellen, obwohl sich bei wenigen Gütemaßen die Ergebnisse wie die Erwartungen verhalten, einige diesen auch widersprechen. So liegen vier der sechs maximalen arithmetischen Mittelwerten bei der Stichprobengröße $n = 400$ und einer sogar bei $n = 100$. Dieses lässt vermuten, dass nicht immer ein größerer Datensatz auch bessere Ergebnisse mit sich bringt. Jedoch ist es wichtig zu berücksichtigen, dass sowohl die Boxen und Whiskers der Boxplots als auch die Standardabweichungen der Gütemaße durchweg größer werden, bei kleinerem Datensatz. Damit wird den Mittelwerten weniger Gewicht zugesprochen und sie sollten mit größerer Vorsicht betrachtet werden. Ein etwas besseres Gütemaß bei der Anwendung von Algorithmen des Maschinellen Lernens mag zwar verlockend klingen, aber ob bei dem einen Test der Wert nun gut oder schlecht ist, lässt sich bei kleinen Datensätzen oft schwer sagen.

Auch muss man sich vorab Gedanken machen, welche Gütemaße einem besonders wichtig sind. Betrachtet man bei diesen Ergebnissen zum Beispiel die Gütemaße Sensitivität und Spezifität sind klare Unterschiede festzustellen. Demnach geben die Modelle sehr gut darüber Auskunft, wenn es morgen nicht regnet (hohe Spezifität). Dies ist zwar eine gute Eigenschaft, wenn es jedoch von größerer Bedeutung ist, eine genauere Auskunft zu erhalten, wenn es regnet, sind die Modelle nicht besonders gut geeignet (geringe Sensitivität).

Bei Untersuchung der Ergebnisse zwischen Entscheidungsbaum und Random Forest ist klar festzustellen, dass der Random Forest Algorithmus bessere Ergebnisse liefert. Das arithmetische Mittel ist bei dem Random Forest für jedes Gütemaß und Fallzahl höher. Dies liegt unter anderem daran, dass es sich bei diesem um eine Ensemble-Methode handelt, welche mehrere unterschiedliche Entscheidungsbäume nutzt, und somit Vorteile gegenüber einen einzigen Entscheidungsbaum mit sich bringt. Auf der einen Seite führt das Bilden des Durchschnitts bzw. die Mehrheitswahl, welche bei dem Random Forest durchgeführt wird, um für ein Testobjekt eine Vorhersage zu treffen, zur Reduzierung von Overfitting. Entscheidungsbäume dagegen müssen sorgfältig trainiert und gekürzt werden, um dieser Gefahr aus dem Weg zu gehen. Auch lässt sich an den Werten der Standardabweichung erkennen, welche zwar nicht immer, aber in den meisten Fällen kleiner ist, dass das Verfahren des Random Forest zur Minderung der Varianz bzw. der Standardabweichung beiträgt. Sofern bei dem Entscheidungsbaum ein Knoten, welcher auf dem Trainingsdatensatz gut gepasst hat, auf dem Testdatensatz aufgrund kleinerer Verschiedenheiten weniger gut passt und falsch klassifiziert, führt dies zu erheblicher Verschlechterung der Ergebnisse.

Es ist allerdings wichtig festzuhalten, dass Random Forest auch Nachteile gegenüber dem Entscheidungsbaum mit sich bringt. Zum einen sind diese komplexer und oft schwerer zu interpretieren, ein Entscheidungsbaum ist auch für Unwissende mit einer kurzen Erklärung zu verstehen. Demnach ist auch die Veranschaulichung deutlich schwerer, da in der Regel bis zu 500 Bäume erstellt werden. Auch dauert das Bauen des Modells aus diesem Grund deutlich länger, welches gerade bei größeren Datensätzen ein guter Grund sein kann zu einem Entscheidungsbaum oder einer anderen Methode des Maschinellen Lernens zu greifen. Trotz dieser Nachteile überwiegt oft der Vorteil, dass die Genauigkeit und Ergebnisse dem Entscheidungsbaum überliegen, sodass Random Forest bevorzugt verwendet wird.

In den Tabellen 4.1 und 4.2 sind die prozentualen Veränderungen der Werte der Tabellen

3.2 und 3.3 dargestellt. Zu sehen sind die relativen Veränderungen der einzelnen Modelle, zum einen von der Fallzahl $n = 1000$ auf $n = 400$ und zum anderen von $n = 400$ auf $n = 100$. Es handelt sich dabei immer um die Veränderung von der größeren zur kleineren Fallzahl der jeweiligen Anwendung. In der Reihe *MAvg* ist die durchschnittliche prozentuale Veränderung des Modells bei der jeweilgen Fallzahländerung zu finde. *Avg* steht wiederum für den gesamten Durchschnitt für die Änderung der Fallzahl, unabhängig von dem Modell.

Zu erkennen ist in Tabelle 4.1, dass bei den Mittelwerten der Random Forest deutlich besser funktioniert als der Entscheidungsbaum. Somit nehmen die Mittelwerte im Durchschnitt sogar um 0.34% zu, bei der Veränderung der Fallzahl von $n = 1000$ auf $n = 400$. Auch bei der Fallzahländerung von $n = 400$ auf $n = 100$ liefert der Random Forest bessere Werte, so fällt dieser lediglich um -4.13% während bei dem Entscheidungsbaum die Mittelwerte um -11.52% sinken. Bei Betrachtung von *Avg*, also unabhängig von den Modellen, ist festzustellen, dass eine Verkleinerung des Datensatzes von $n = 1000$ auf $n = 400$ nur zu einer durchschnittlichen. Verminderung der Mittelwerte der Gütemaße um -1.58% führt. Nutzt man einen Datensatz der Größe $n = 100$ fällt dieser Wert um weitere -7.82%.

In Tabelle 4.2, in welcher die prozentualen Änderungen der Standardabweichung dargestellt sind, ist ein ähnliches Bild zu erkennen. Der Random Forest scheint bei der Veränderung von $n = 1000$ auf $n = 400$ noch resistenter bezogen auf die Standardabweichung zu sein, sodass diese um 48.29% steigt. Der Wert des Entscheidungsbaumes liegt bei dieser Veränderung bei 63.39% und somit höher. Für beide Anwendungen gilt, dass die Standardabweichung bei noch kleinerem Datensatz ($n = 100$) stark weiter ansteigt, bei dem Entscheidungsbaum um 132.79% und bei dem Random Forest um 141.57%. Dies zeigt erneut, dass ein kleiner Datensatz große Unsicherheiten bezüglich der Ergebnisse mit sich bringt und daher die Wahl des Modells und der Parameter des Modells eine umso wichtigere Rolle spielt, um konstant gute Ergebnisse zu erzielen. Unabhängig der Modelle liegen die durchschnittlichen Veränderungen bei 55.84% ($n = 1000 \rightarrow 400$) und 137.18% ($n = 400 \rightarrow 100$).

4.2 Varianzkomponenten der Modelle

Im Folgenden werden die Ergebnisse des Kapitels 3.2 genauer betrachtet und interpretiert. Dabei wird auf die Frage zurückgegriffen, inwiefern das Verkleinern eines Datensat-

Gütemaß	n = 1000 → 400		n = 400 → 100	
	dt	rf	dt	rf
Acc	1.15%	3.00%	-7.04%	-2.81%
Sens	-7.74%	-5.42%	-16.39%	-8.09%
PPV	-11.44%	1.86%	-25.45%	-16.90%
Spec	0.08%	1.98%	-5.50%	-2.44%
NPV	2.43%	2.50%	-2.19%	-0.40%
AUC	-5.45%	-1.86%	-12.55%	5.87%
MAvg	-3.50%	0.34%	-11.52%	-4.13%
Avg	**-1.58%**		**-7.82%**	

Tabelle 4.1: Gütemaße Mittelwerte Prozentuale Veränderung

Gütemaß	n = 1000 → 400		n = 400 → 100	
	dt	rf	dt	rf
Acc	43.40%	34.76%	112.15%	152.65%
Sens	89.37%	58.48%	178.97%	183.39%
PPV	81.77%	53.32%	107.76%	164.00%
Spec	45.82%	17.65%	179.78%	142.73%
NPV	42.41%	38.60%	132.29%	144.94%
AUC	77.69%	86.96%	85.78%	61.73%
MAvg	63.39%	48.29%	132.79%	141.57%
Avg	**55.84%**		**137.18%**	

Tabelle 4.2: Gütemaße Standardabweichungen Prozentuale Veränderung

zes schlechtere Ergebnisse liefert bzw. wie viel unzuverlässiger diese werden.

Die Standardabweichung des Mittelwertes, welche in Tabelle 3.3 dargestellt sind, berücksichtigt lediglich die Standardabweichung der Mittelwerte der zehn unterschiedlichen Datensätze. Die Varianzen, welche innerhalb dieses Datensatzes auftreten haben kein Einfluss auf diesen Wert, nur die arithmetischen Mittelwerte dieser 100 Iterationen werden genutzt, um die Varianz zu untersuchen, welche durch das Auswählen unterschiedlicher Datensätze entsteht. Für die Fallzahl $n = 1000$ liegen die Werte für alle Gütemaße unter 1% für den Random Forest sogar bei vier von sechs Gütemaßen unter $0,21\%$. Dieses Ergebnis zeigt, dass die zufällige Wahl des Datensatzes bei einer Stichprobengröße von $n = 1000$ durchschnittlich zu einer Abweichung meist deutlich kleiner als 1% der arithmetischen Mittelwerte führt.

In Tabelle 4.3 sind die prozentualen Veränderungen der Werte in Tabelle 3.3 dargestellt. Die Spalten und Reihen lesen sich dabei wie in den Tabellen des vorherigen Unterkapitels 4.1. Besonders auffällig ist dabei, dass die einzig negative Veränderung bei dem negativen Vorhersagewert auftritt, welcher bei der Fallzahländerung von $n = 1000$ auf $n = 400$ um -19.23% abnimmt. Ansonsten liefern die prozentualen Veränderungen der Standardabweichung des Mittelwerts einen eindeutigen Trend. Dabei liegen die Änderungen des Random Forest, jedoch höher als die des Entscheidungsbaumes. Dies kann durchaus daran liegen, dass die Ursprungswerte des Random Forest für die Standardabweichungen des Mittelwertes bei $n = 1000$ sehr gering sind und somit relativ stärker steigen. Bei Betrachtung des *Avg* ist festzustellen, dass der Wert für die erste Fallzahländerung mit 54.56% geringer ausfällt als für die zweite Fallzahländerung (128.83%). Die lässt schließen, dass bei kleineren Datensätzen ($n = 100$) die Mittelwerte der Gütemaße im Durchschnitt deutlich stärker streuen, um mehr als das doppelte als bei $n = 400$.

In Tabelle 4.4 sind die prozentualen Veränderungen der Mittelwerte der Standardabweichung dargestellt. Die Werte in *MAvg* liegen mit 63.55% und 57.70% sowohl in den linken Spalten als auch in den rechten Spalten mit 97.96% und 116.67% näher beieinander als in der vorherigen Tabelle 4.3. Dies deutet darauf hin, dass die Mittelwerte der Standardabweichung bei den Anwendungen Entscheidungsbaum und Random Forest sich ähnlich verhalten, aber auch stärker ansteigen, desto kleiner der Datensatz wird. Die Werte in der Reihe *Avg* liegen bei 60.62% und 107.32%. Auch dies entspricht den bisherigen Werten insofern, dass der Anstieg stärker wird, wenn der Datensatz kleiner wird.

Bei den drei untersuchten Fallzahlen konnte also mithilfe der Varianzkomponenten Standardabweichung des Mittelwertes und Mittelwerte der Standardabweichung gezeigt wer-

Gütemaß	n = 1000 → 400		n = 400 → 100	
	dt	rf	dt	rf
Acc	77.27%	107.69%	64.10%	177.78%
Sens	12.00%	21.92%	166.67%	232.58%
PPV	73.17%	166.04%	47.18%	95.04%
Spec	30.30%	16.67%	93.02%	71.43%
NPV	-19.23%	66.67%	219.05%	203.33%
AUC	68.89%	33.33%	7.89%	167.86%
MAvg	40.40%	68.72%	99.65%	158.00%
Avg	**54.56%**		**128.83%**	

Tabelle 4.3: Gütemaße Standardabweichungen des Mittelwerts Prozentuale Veränderung

Gütemaß	n = 1000 → 400		n = 400 → 100	
	dt	rf	dt	rf
Acc	46.47%	41.26%	95.75%	124.40%
Sens	70.75%	73.90%	133.00%	138.99%
PPV	86.00%	77.16%	60.72%	118.37%
Spec	46.74%	22.51%	122.22%	126.92%
NPV	42.19%	32.17%	116.62%	139.47%
AUC	89.15%	99.17%	59.44%	51.88%
MAvg	63.55%	57.70%	97.96%	116.67%
Avg	**60.62%**		**107.32%**	

Tabelle 4.4: Gütemaße Mittelwerte der Standardabweichung Prozentuale Veränderung

den, dass eine Varianz durch das Wählen unterschiedlicher Datensätze entsteht. Diese ist zwar gering, aber steht in starkem Zusammenhang mit der Größe des Datensatzes. So steigen die Varianzkomponenten bei kleineren Datensätzen. Schon eine Wahl von $n = 400$ statt $n = 100$ sorgt dafür, dass diese mehr als halbiert werden. Der Unterschied zwischen den Fallzahlen $n = 1000$ und $n = 400$ ist zwar nicht so groß aber auch bemerkbar.

5. Fazit

Ziel der vorliegenden Arbeit war es zu untersuchen, inwiefern die Anwendungen Entscheidungsbaum und Random Forest durch die Veränderungen von Fallzahlen beeinflusst werden. Dabei wurde die Untersuchung zum einen durch das Analysieren von ausgewählten Gütemaßen und zum anderen der Betrachtungen weiterer Varianzkomponenten unterstützt.

Bei den Ergebnissen konnte mit ein paar Ausnahmen gezeigt werden, dass ein kleinerer Datensatz zu schlechteren Gütemaßen und größerer Varianz dieser führt. Die Arbeit konnte bestätigen, dass das Verfahren Random Forest bessere arithmetische Mittel der Gütemaße als das Verfahren Entscheidungsbaum erzielen konnte. Andererseits musste festgestellt werden, dass bei Untersuchung der Standardabweichung kein eindeutiges Ergebnis in Bezug auf die Verfahren Entscheidungsbaum und Random Forest festgestellt werden konnte.

Bei der Untersuchung der Varianzkomponenten bestätigt sich erneut, dass eine niedrigere Fallzahl zu höherer Streuung der Ergebnisse führt. Dabei ist der Unterschied zwischen den untersuchten Fallzahlen $n = 1000$ und $n = 400$ nicht so stark wie $n = 400$ und $n = 100$. Einen Datensatz von $n = 100$ auf $n = 400$ zu erweitern, kann also schon weitaus bessere Ergebnisse mit sich bringen, während die weiteren Daten bis $n = 1000$ diese nur minimal aufbessern.

So konnte die Arbeit zeigen, dass bei kleinen Datensätzen die Ergebnisse mit Vorsicht behandelt werden sollten und die Wahl eines guten Modells mit passenden Parametern von Wichtigkeit ist. Eine Erweiterung des Datensatzes kann dabei schon viel bewirken. Inwiefern es in der Praxis umsetzbar ist, ob es sich lohnt, statt Daten von 100 Patienten Daten von 400 Patienten aufzunehmen, ist individuell zu erwägen und nicht mit dieser Arbeit beantwortet.

Es ist zu berücksichtigen, dass die gesamte Untersuchung lediglich auf einem Datensatz und mit nur zwei unterschiedlichen Verfahren durchgeführt wurde. Es bietet so zwar

einen guten Überblick über das Thema, aber weitere Arbeiten in diesem Bereich können sich mit anderen Verfahren, Datensätzen und unter Umständen auch anderen Fallzahlen beschäftigen.

Literaturverzeichnis

Ali, J., Khan, R., Ahmad, N., Maqsood, I., 2012. Random forests and decision trees. International Journal of Computer Science Issues (IJCSI) 9, 272.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. wadsworth int. Group 37, 237–251.

Centor, R.M., 1991. Signal detectability: the use of roc curves and their analyses. Medical decision making 11, 102–106.

Dietterich, T., 1995. Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR) 27, 326–327.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hayes, T., Usami, S., Jacobucci, R., McArdle, J.J., 2015. Using classification and regression trees (cart) and random forests to analyze attrition: Results from two simulations. Psychology and aging 30, 911.

Lalkhen, A.G., McCluskey, A., 2008. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care & Pain 8, 221–223.

Prajwala, T., 2015. A comparative study on decision tree and random forest using r tool. International journal of advanced research in computer and communication engineering 4, 196–199.

Shaikhina, T., Khovanova, N.A., 2017. Handling limited datasets with neural networks in medical applications: A small-data approach. Artificial intelligence in medicine 75, 51–63.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., 2015. Machine learning for predictive modelling based on small data in biomedical engineering. IFAC-PapersOnLine 48, 469–474.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., 2017. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control .

A. Anhang

A.1 Datensatz

Im folgenden ist ein Auszug des zu Grunde liegenden Datensatzes Rain in Australia, welcher von Kaggle stammt. Der vollständige Datensatz wurde per E-Mail am 11.03.2020 an den Erstgutachter Prof. Dr. Kennes geschickt.

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RISK_MM	RainTomorrow
1	2008-12-01	Albury	13.4	22.9	0.6	NA	NA	W	44	W	WNW	20	24	71	22	1007.7	1007.1	8	NA	16.9	21.8	No	0.0	No
2	2008-12-02	Albury	7.4	25.1	0.0	NA	NA	WNW	44	NNW	WSW	4	22	44	25	1010.6	1007.8	NA	NA	17.2	24.3	No	0.0	No
3	2008-12-03	Albury	12.9	25.7	0.0	NA	NA	WSW	46	W	WSW	19	26	38	30	1007.6	1008.7	NA	2	21.0	23.2	No	0.0	No
4	2008-12-04	Albury	9.2	28.0	0.0	NA	NA	NE	24	SE	E	11	9	45	16	1017.6	1012.8	NA	NA	18.1	26.5	No	1.0	No
5	2008-12-05	Albury	17.5	32.3	1.0	NA	NA	W	41	ENE	NW	7	20	82	33	1010.8	1006.0	7	8	17.8	29.7	No	0.2	No
6	2008-12-06	Albury	14.6	29.7	0.2	NA	NA	WNW	56	W	W	19	24	55	23	1009.2	1005.4	NA	NA	20.6	28.9	No	0.0	No
7	2008-12-07	Albury	14.3	25.0	0.0	NA	NA	W	50	SW	W	20	24	49	19	1009.6	1008.2	1	NA	18.1	24.6	No	0.0	No
8	2008-12-08	Albury	7.7	26.7	0.0	NA	NA	W	35	SSE	W	6	17	48	19	1013.4	1010.1	NA	NA	16.3	25.5	No	0.0	No
9	2008-12-09	Albury	9.7	31.9	0.0	NA	NA	NNW	80	SE	NW	7	28	42	9	1008.9	1003.6	NA	NA	18.3	30.2	No	1.4	Yes
10	2008-12-10	Albury	13.1	30.1	1.4	NA	NA	W	28	S	SSE	15	11	58	27	1007.0	1005.7	NA	NA	20.1	28.2	Yes	0.0	No
11	2008-12-11	Albury	13.4	30.4	0.0	NA	NA	N	30	SSE	ESE	17	6	46	22	1011.8	1008.7	NA	NA	20.4	28.8	No	2.2	Yes
12	2008-12-12	Albury	15.9	21.7	2.2	NA	NA	NNE	31	NE	ENE	15	13	89	91	1010.5	1004.2	8	8	15.9	17.0	Yes	15.6	Yes
13	2008-12-13	Albury	15.9	18.6	15.6	NA	NA	W	61	NNW	NNW	28	28	76	93	994.3	993.0	8	8	17.4	15.8	Yes	3.6	Yes
14	2008-12-14	Albury	12.6	21.0	3.6	NA	NA	SW	44	W	SSW	24	20	65	43	1001.2	1001.8	NA	7	15.8	19.8	Yes	0.0	No
15	2008-12-15	Albury	9.8	27.7	NA	NA	NA	WNW	50	NA	WNW	NA	22	50	28	1013.4	1010.3	0	NA	17.3	26.2	NA	0.0	No
16	2008-12-17	Albury	14.1	20.9	0.0	NA	NA	ENE	22	SSW	E	11	9	69	82	1012.2	1010.4	8	1	17.2	18.1	No	16.8	Yes
17	2008-12-18	Albury	13.5	22.9	16.8	NA	NA	W	63	N	WNW	6	20	80	65	1005.8	1002.2	8	1	18.0	21.5	Yes	10.6	Yes
18	2008-12-19	Albury	11.2	22.5	10.6	NA	NA	SSE	43	WSW	SW	24	17	47	32	1009.4	1009.7	NA	2	15.5	21.0	Yes	0.0	No
19	2008-12-20	Albury	9.8	25.6	0.0	NA	NA	SSE	26	SE	NNW	17	6	45	26	1019.2	1017.1	NA	NA	15.8	23.2	No	0.0	No
20	2008-12-21	Albury	11.5	29.3	0.0	NA	NA	S	24	SE	SE	9	9	56	28	1019.3	1014.8	NA	NA	19.1	27.3	No	0.0	No
21	2008-12-22	Albury	17.1	33.0	0.0	NA	NA	NE	43	NE	N	17	22	38	28	1013.6	1008.1	NA	1	24.5	31.6	No	0.0	No
22	2008-12-23	Albury	20.5	31.8	0.0	NA	NA	WNW	41	W	W	19	20	54	24	1007.8	1005.7	NA	NA	23.8	30.8	No	0.0	No
23	2008-12-24	Albury	15.3	30.9	0.0	NA	NA	N	33	ESE	NW	6	13	55	23	1011.0	1008.2	5	NA	20.9	29.0	No	0.0	No
24	2008-12-25	Albury	12.6	32.4	0.0	NA	NA	W	43	E	W	4	19	49	17	1012.9	1010.1	NA	NA	21.5	31.2	No	0.0	No
25	2008-12-26	Albury	16.2	33.9	0.0	NA	NA	WSW	35	SE	WSW	9	13	45	19	1010.9	1007.6	NA	1	23.2	33.0	No	0.0	No
26	2008-12-27	Albury	16.9	33.0	0.0	NA	NA	WSW	57	NA	W	0	26	41	28	1006.8	1003.6	NA	1	26.6	31.2	No	0.0	No
27	2008-12-28	Albury	20.1	32.7	0.0	NA	NA	WNW	48	N	WNW	13	30	56	15	1005.2	1001.7	NA	NA	24.6	32.1	No	0.0	No
28	2008-12-29	Albury	19.7	27.2	0.0	NA	NA	WNW	46	NW	WSW	19	30	49	22	1004.8	1004.2	NA	NA	21.6	26.1	No	1.2	Yes
29	2008-12-30	Albury	12.5	24.2	1.2	NA	NA	WNW	50	WSW	SW	11	22	76	70	1005.6	1003.4	8	8	12.5	18.2	Yes	0.8	No
30	2008-12-31	Albury	12.0	24.4	0.8	NA	NA	W	39	WNW	WNW	17	17	48	28	1006.1	1005.1	1	NA	16.9	22.7	No	0.0	No
31	2009-01-01	Albury	11.3	26.5	0.0	NA	NA	WNW	56	W	WNW	19	31	46	26	1004.5	1003.2	NA	NA	19.7	25.7	No	0.0	No
32	2009-01-02	Albury	9.6	23.9	0.0	NA	NA	W	41	WSW	SSW	19	11	44	22	1014.4	1013.1	NA	NA	14.9	22.1	No	0.0	No
33	2009-01-03	Albury	10.5	26.8	0.0	NA	NA	SSE	26	SSE	E	11	7	43	22	1018.7	1014.8	NA	NA	17.1	26.5	No	0.0	No
34	2009-01-04	Albury	12.3	34.6	0.0	NA	NA	WNW	37	SSE	NW	6	17	41	12	1015.1	1010.3	NA	NA	20.7	33.9	No	0.0	No
35	2009-01-05	Albury	12.9	35.8	0.0	NA	NA	WNW	41	ENE	NW	6	26	41	9	1012.6	1009.2	NA	NA	22.4	34.4	No	0.0	No
36	2009-01-06	Albury	13.7	37.9	0.0	NA	NA	W	52	SE	WNW	4	28	33	8	1010.9	1006.7	NA	NA	23.1	36.8	No	0.0	No
37	2009-01-07	Albury	16.1	38.9	0.0	NA	NA	W	57	E	W	6	30	34	12	1007.0	1002.7	NA	NA	25.2	38.4	No	0.0	No
38	2009-01-08	Albury	14.0	28.3	0.0	NA	NA	W	48	W	WSW	17	24	43	15	1011.9	1010.9	NA	NA	17.9	27.6	No	0.0	No
39	2009-01-09	Albury	12.5	28.4	0.0	NA	NA	NE	37	SSE	S	20	9	38	16	1017.8	1013.7	NA	NA	17.2	26.6	No	0.0	No
40	2009-01-10	Albury	17.0	30.8	0.0	NA	NA	NE	37	NNE	E	15	11	36	24	1013.4	1008.1	NA	NA	20.2	29.3	No	0.0	No
41	2009-01-11	Albury	16.9	32.0	0.0	NA	NA	S	31	SSE	N	13	17	52	31	1009.9	1006.8	NA	NA	22.8	30.0	No	0.0	No
42	2009-01-12	Albury	17.3	34.7	0.0	NA	NA	SW	35	SE	WSW	7	15	48	16	1014.1	1012.1	NA	NA	24.2	33.2	No	0.0	No
43	2009-01-13	Albury	17.2	37.7	0.0	NA	NA	NNW	35	SE	NW	7	17	51	19	1015.7	1010.9	NA	NA	24.3	35.7	No	0.0	No
44	2009-01-14	Albury	17.4	43.0	0.0	NA	NA	NW	39	SSE	SSW	7	17	40	8	1011.6	1006.9	NA	NA	25.6	41.5	No	0.0	No
45	2009-01-15	Albury	19.8	32.7	0.0	NA	NA	WNW	44	W	W	20	28	34	28	1008.4	1009.2	NA	NA	27.6	27.1	No	0.0	No
46	2009-01-16	Albury	14.9	26.7	0.0	NA	NA	SW	56	WSW	SW	20	31	46	20	1014.1	1012.7	NA	NA	18.0	25.5	No	0.0	No
47	2009-01-17	Albury	10.5	28.4	0.0	NA	NA	SE	33	SE	SW	19	11	35	16	1018.7	1017.4	NA	NA	16.0	25.8	No	0.0	No

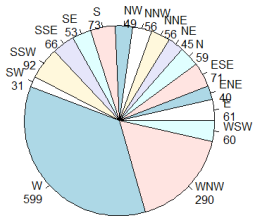
A.2 R Studio Code

Im folgenden ist ein Auszug des zu dieser Arbeit erstellten R Studio Codes. Der vollständige Code wurde per E-Mail am 11.03.2020 an den Erstgutachter Prof. Dr. Kennes geschickt.

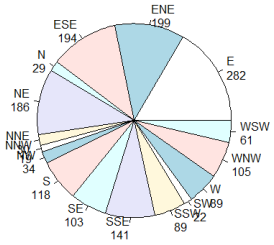
```

##Entscheidungsbaum (DT)
#n=1000

k <- 100
acc <- c(rep(0, k))
recall <- c(rep(0,k))
prec <- c(rep(0,k))
tnr <- c(rep(0,k))
npv <- c(rep(0,k))
auc <- c(rep(0,k))


DT1000 <- data.frame("acc" = acc, "recall" = recall, "prec" = prec, "tnr" = tnr, "npv" = npv, "auc" = auc)


set.seed(123)
for (i in seq_len(k)) {
  n <- 1000
  shuff <- dat5[sample(n),]
  train_ind <- sample(1:n, size = round(0.75*n))
  train <- shuff[train_ind,]
  test<- shuff[-train_ind,]
  tree <- rpart(RainTomorrow ~ ., train, method = "class")
  pred <- predict(tree,test,type="class")
  conf <- table(test$RainTomorrow, pred)
  DT1000$acc[i] <- (sum(diag(conf)))/sum(conf)
  DT1000$prec[i] <- conf[2,2]/sum(conf[,2])
  DT1000$recall[i] <- conf[2,2]/sum(conf[2,])
  DT1000$tnr[i] <- conf[1,1]/sum(conf[1,])
  DT1000$npv[i] <- conf[1,1]/sum(conf[,1])
  if (sum(as.numeric(test$RainTomorrow)) != sum(conf)) {
    rocpred <- performance(prediction(predict(tree, test, type="prob")[,2], test$RainTomorrow), measure = "auc")
    DT1000$auc[i] <- rocpred@y.values[[1]]
  } else {
    DT1000$auc[i] <- NaN
  }
}



sapply(DT1000, mean, na.rm = TRUE)
sapply(DT1000, sd, na.rm = TRUE)




#n=400

DT400 <- data.frame("accdt400" = acc, "recalldt400" = recall, "precdt400" = prec, "tnrdt400" = tnr, "npvdt400" = npv, "aucdt400" = auc)

set.seed(123)
for (i in seq_len(k)) {
  n <- 400
  shuff <- dat5[sample(n),]
  train_ind <- sample(1:n, size = round(0.75*n))
  train <- shuff[train_ind,]
  test<- shuff[-train_ind,]
  tree <- rpart(RainTomorrow ~ ., train, method = "class")
  pred <- predict(tree,test,type="class")
  conf <- table(test$RainTomorrow, pred)
  DT400$accdt400[i] <- (sum(diag(conf)))/sum(conf)
  DT400$precdt400[i] <- conf[2,2]/sum(conf[,2])
  DT400$recalldt400[i] <- conf[2,2]/sum(conf[2,])
  DT400$tnrdt400[i] <- conf[1,1]/sum(conf[1,])
```

A.3 Abbildungen

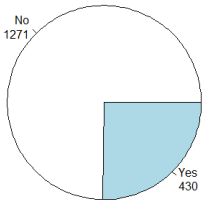
In der Abbildung A.1 sind weitere Diagramme, welche die qualitativen Variablen des Sydney Datensatzes ($n = 1701$) beschreiben.



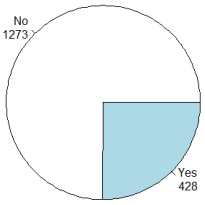
Windrichtung 9 Uhr Vormittag (Sydney)



Windrichtung 3 Uhr Nachmittag (Sydney)



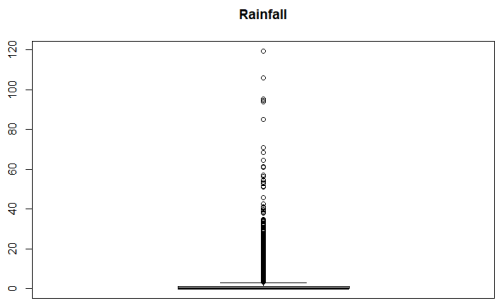
Regen Heute (Sydney)



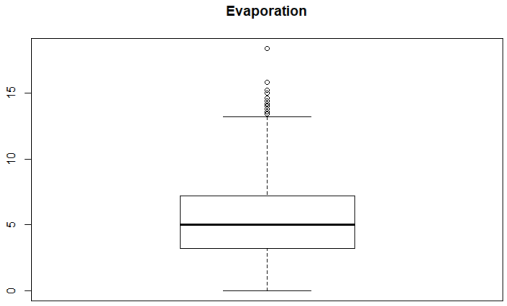
Regen Morgen (Sydney)

Abbildung A.1: Weitere Diagramme qualitiative Variablen Datensatz Sydney ($n = 1701$)

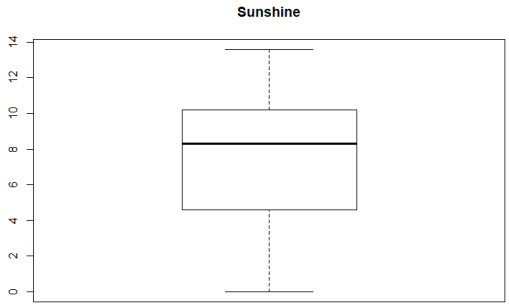
In den Abbildungen A.2 und A.3 sind die weiteren Boxplots zu dem Datensatz Sydney ($n = 1701$) zu finden.



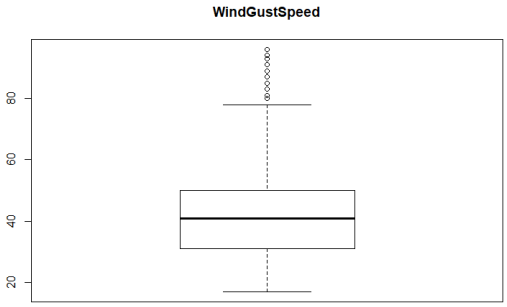
Boxplot der Variable Rainfall



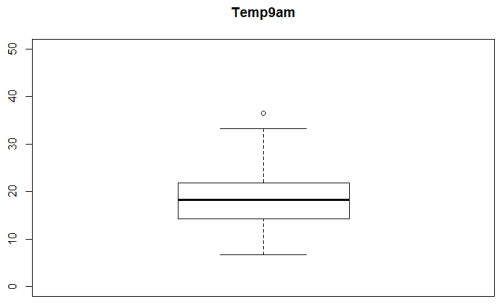
Boxplot der Variable Evaporation



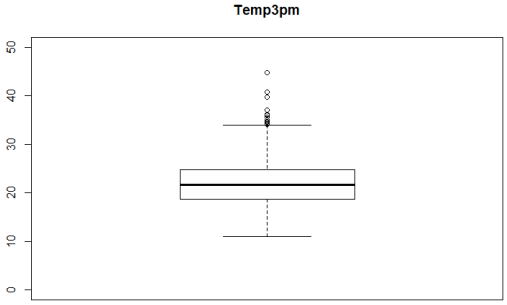
Boxplot der Variable Sunshine



Boxplot der Variable WindGustSpeed

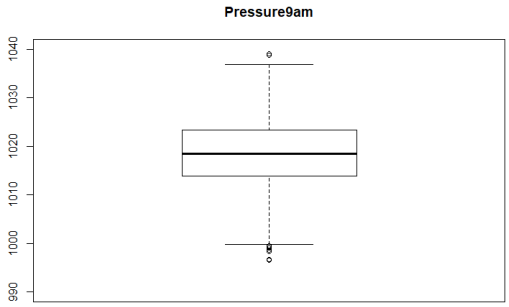


Boxplot der Variable Temp9am

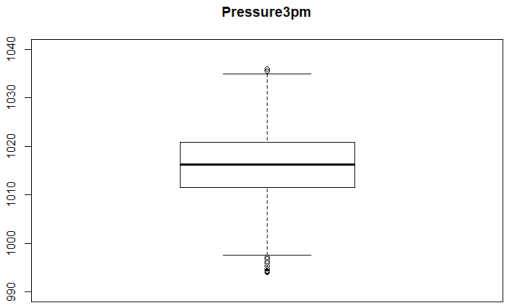


Boxplot der Variable Temp3pm

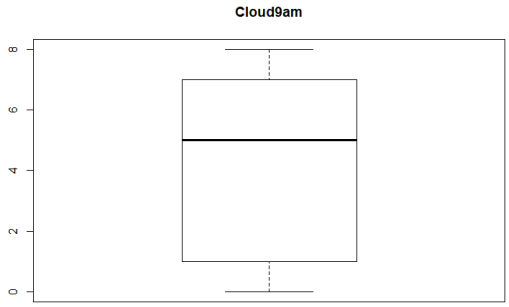
Abbildung A.2: (1) Boxplots zu weiteren Variablen des Datensatzes Sydney ($n = 1701$)



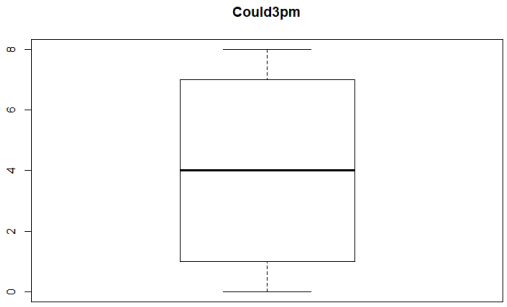
Boxplot der Variable Pressure9am



Boxplot der Variable Pressure3pm



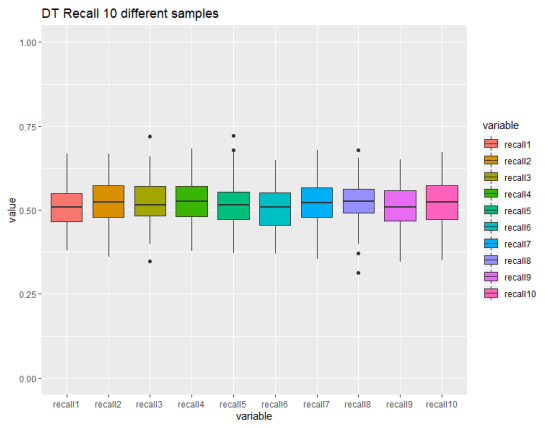
Boxplot der Variable Cloud9am



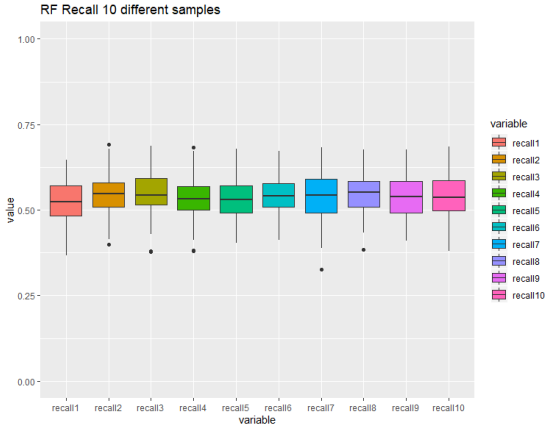
Boxplot der Variable Cloud3pm

Abbildung A.3: (2) Boxplots zu weiteren Variablen des Datensatzes Sydney ($n = 1701$)

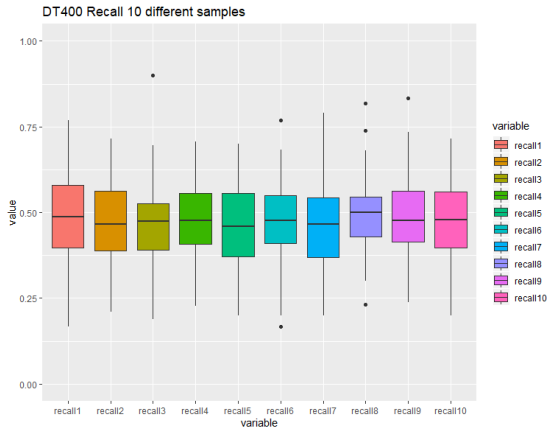
In den folgenden Abbildungen A.4 bis A.8 sind weitere Boxplots der Gütemaße Sensitivität, positiver Vorhersagewert, Spezifitität, negativer Vorhersagewert und Area under the Curve. Dabei sind links die Plots des Entscheidungsbaums und rechts die Plots des Random Forests. Von oben nach unten Folgen sie dem Schema $n = 1000, 400$ und 100.



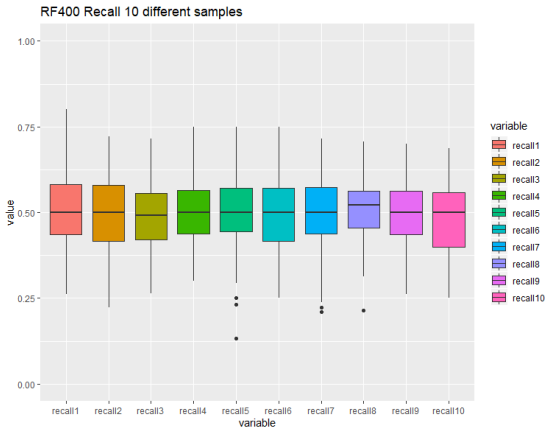
Entscheidungsbaum $n = 1000$



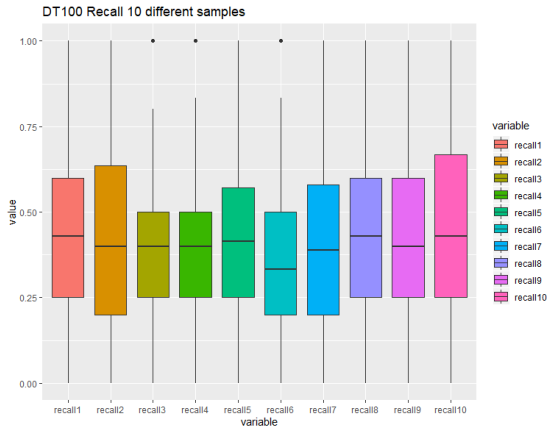
Random Forest $n = 1000$



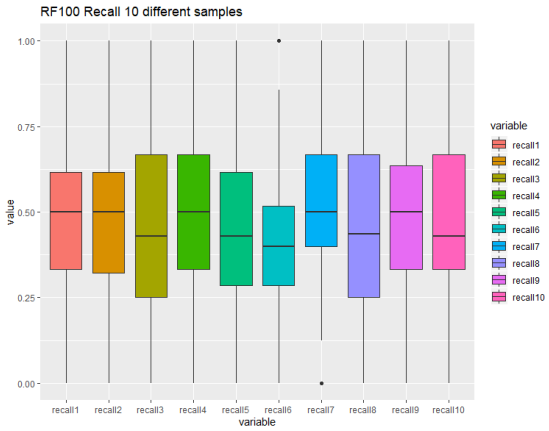
Entscheidungsbaum $n = 400$



Random Forest $n = 400$

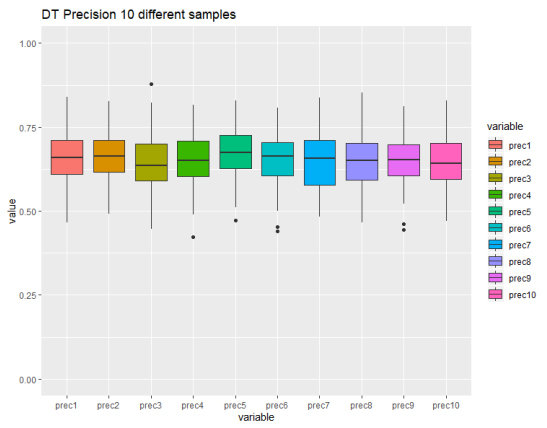


Entscheidungsbaum $n = 100$

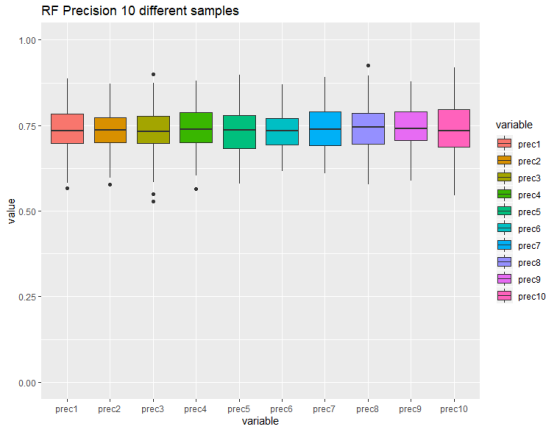


Random Forest $n = 100$

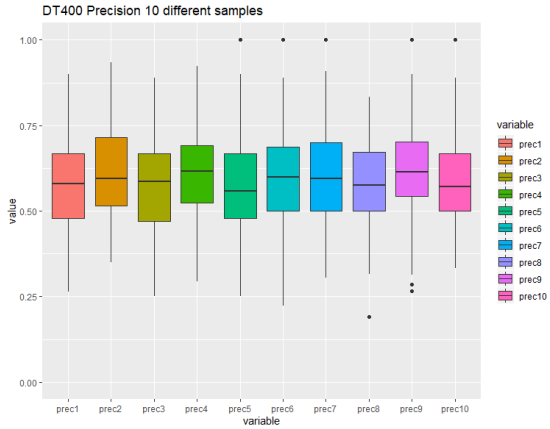
Abbildung A.4: Boxplots zu Varianzkomponenten des Gütemaßes Sensitivität



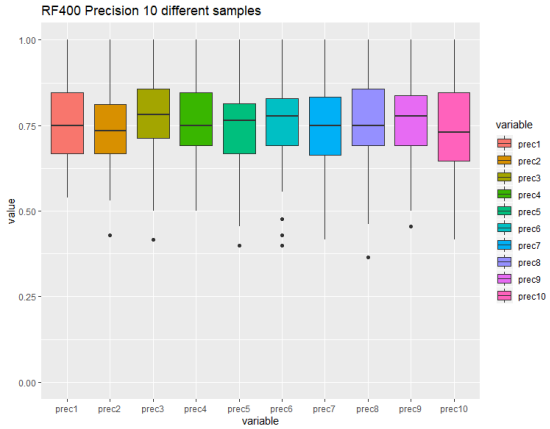
Entscheidungsbaum $n = 1000$



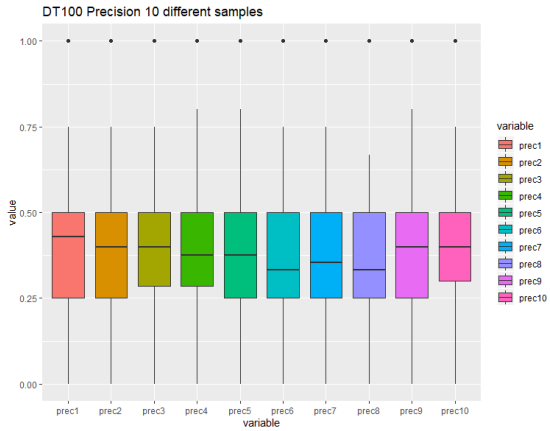
Random Forest $n = 1000$



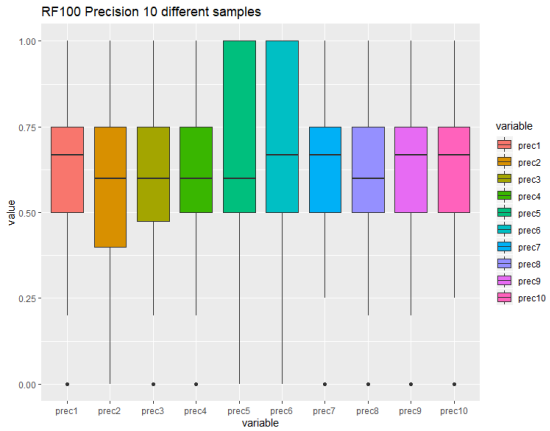
Entscheidungsbaum $n = 400$



Random Forest $n = 400$

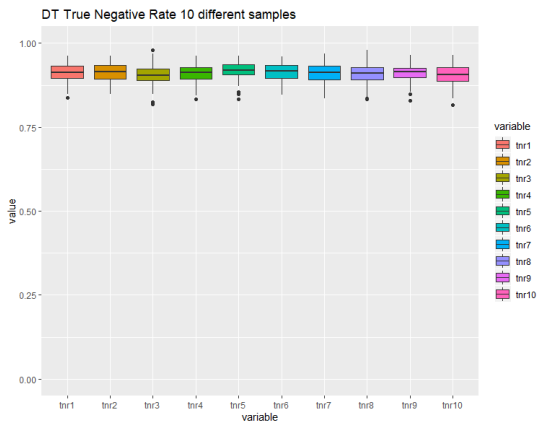


Entscheidungsbaum $n = 100$

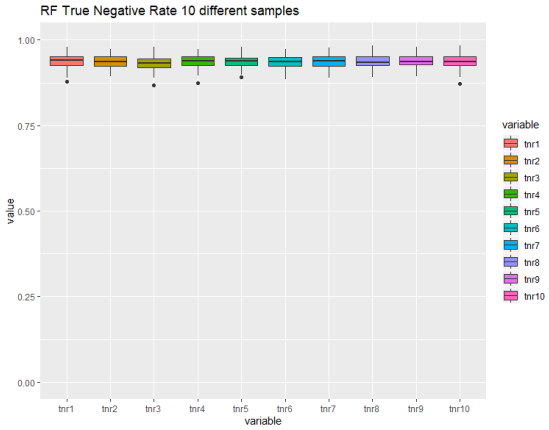


Random Forest $n = 100$

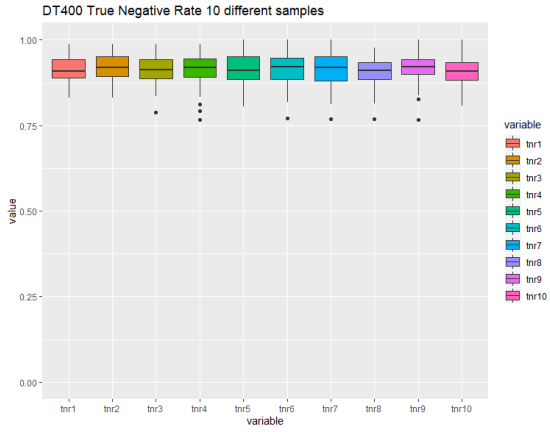
Abbildung A.5: Boxplots zu Varianzkomponenten des Gütemaßes positiver Vorhersagewert



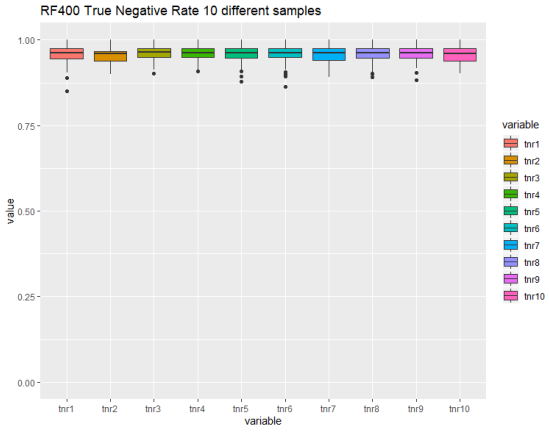
Entscheidungsbaum $n = 1000$



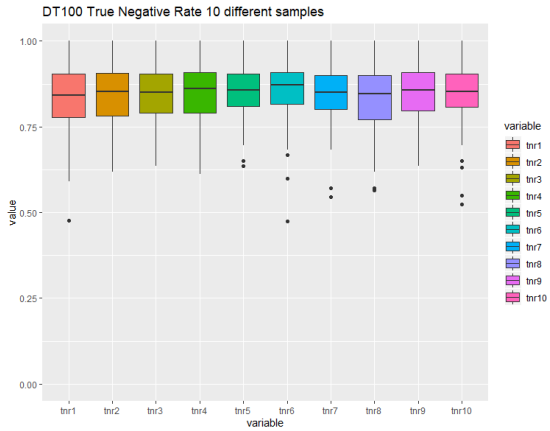
Random Forest $n = 1000$



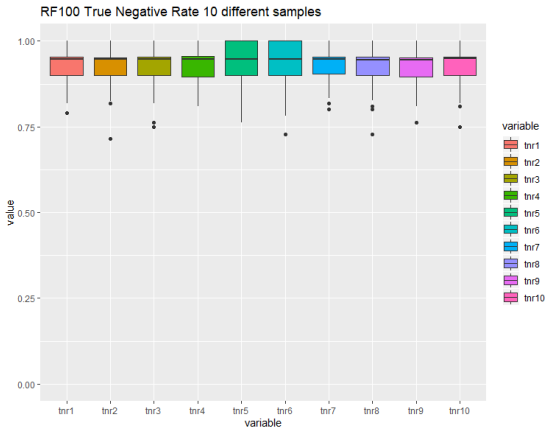
Entscheidungsbaum $n = 400$



Random Forest $n = 400$

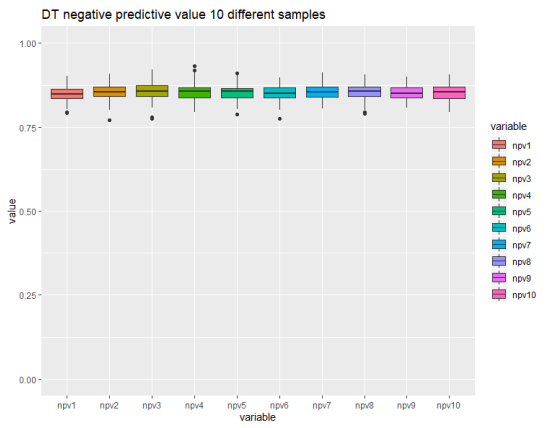


Entscheidungsbaum $n = 100$

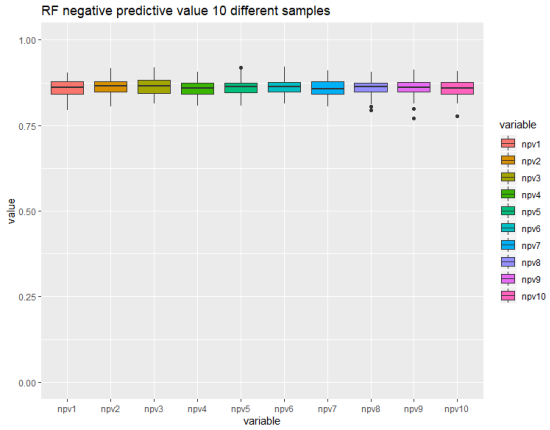


Random Forest $n = 100$

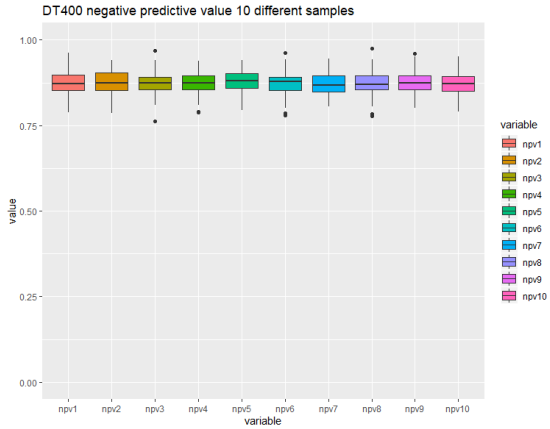
Abbildung A.6: Boxplots zu Varianzkomponenten des Gütemaßes Spezifität



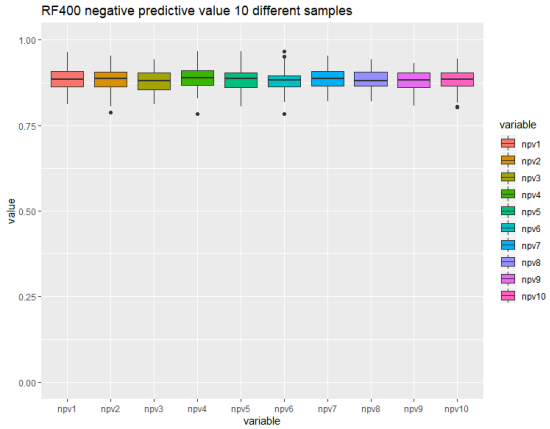
Entscheidungsbaum $n = 1000$



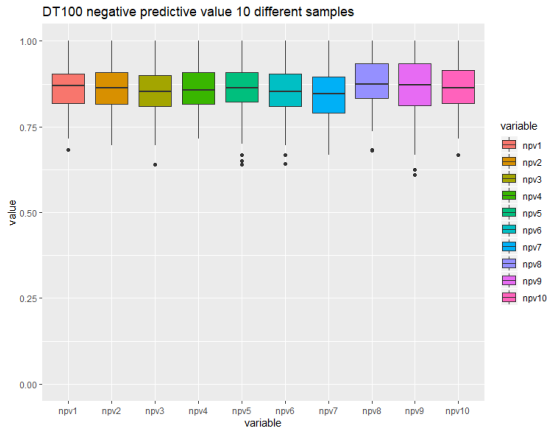
Random Forest $n = 1000$



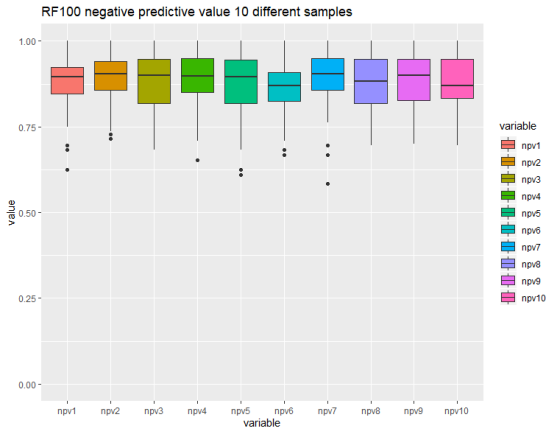
Entscheidungsbaum $n = 400$



Random Forest $n = 400$

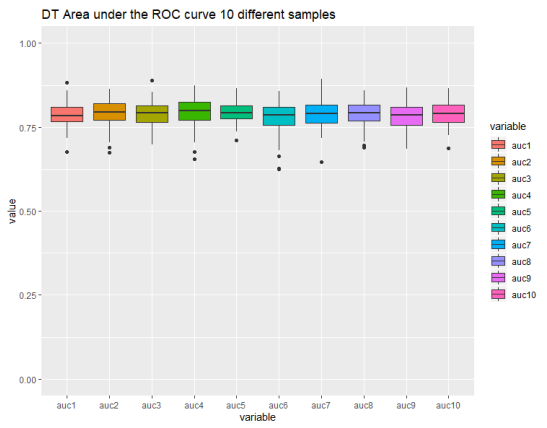


Entscheidungsbaum $n = 100$

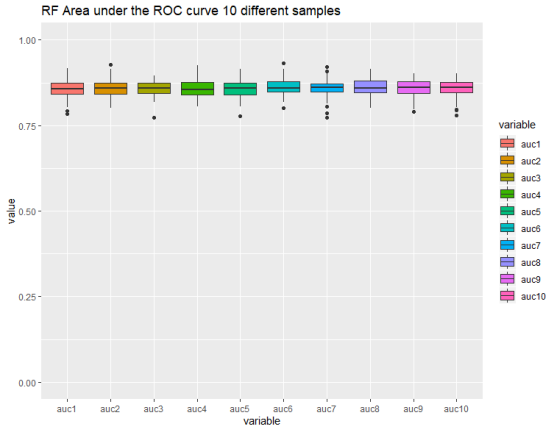


Random Forest $n = 100$

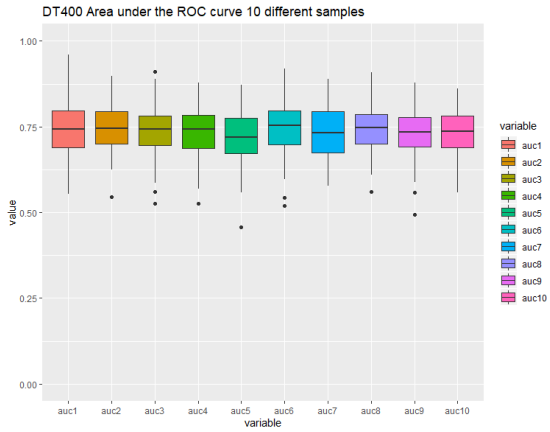
Abbildung A.7: Boxplots zu Varianzkomponenten des Gütemaßes negativer Vorhersagewert



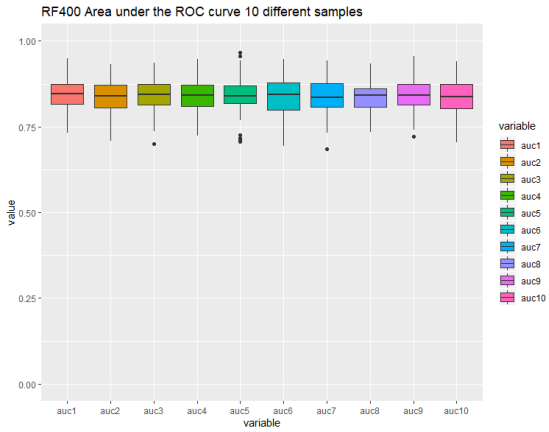
Entscheidungsbaum $n = 1000$



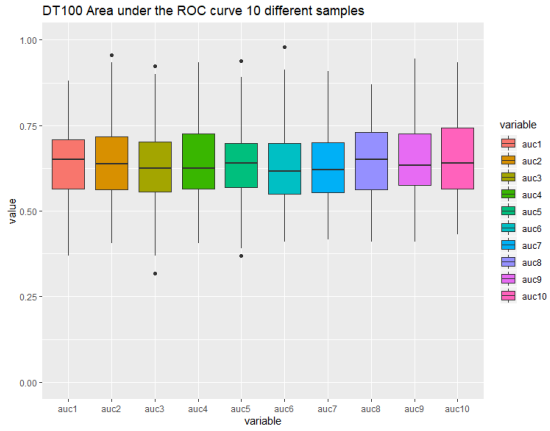
Random Forest $n = 1000$



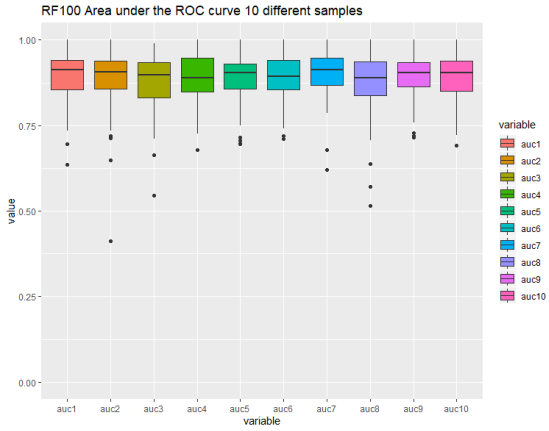
Entscheidungsbaum $n = 400$



Random Forest $n = 400$



Entscheidungsbaum $n = 100$



Random Forest $n = 100$

Abbildung A.8: Boxplots zu Varianzkomponenten des Gütemaßes Area under the curve

A.4 Erklärung Speicherung Daten

Hiermit erkläre ich, dass der beiliegende externe Datenträger kopiert und gespeichert werden darf.

Ort und Datum	Bjarne Seen

A.5 Eidesstaatliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort und Datum	Bjarne Seen