

Lehrstuhl für Finanzwissenschaft und Public Management

Prof. Dr. Berthold U. Wigger

Seminararbeit

Fach: Finanzwissenschaft

Wintersemester 2019/2020

MASCHINELLES LERNEN AUF PISA DATEN MIT HILFE VON ENTSCHEIDUNGSBÄUMEN

Prüfer: Prof. Dr. Berthold U. Wigger

Themensteller Lars Herberholz

Ausgabetermin:

Abgabetermin:

Nina Graves, 1801989

Von: Marie Rahlmeyer, 2271029

Jendrik von Wardenburg, 2247650

INHALTSVERZEICHNIS

Inhaltsverzeichnis	i
Abkürzungsverzeichnis	ii
Abbildungsverzeichnis	iii
Tabellenverzeichnis	iv
1 Einleitung	1
2 Literaturüberblick	3
3 Theorie	5
3.1 Entscheidungsbäume	5
3.2 Pruning.....	8
3.3 Ensemble Methoden	9
4 Daten	10
4.1 PISA Daten	10
4.2 Featuredaten.....	12
5 Modellanalyse	14
5.1 Entscheidungsbäume	15
5.1.1 Math Datensatz	17
5.1.2 Read Datensatz	19
5.1.3 Science Datensatz	21
5.2 Pruning.....	23
5.3 Random Forest.....	24
6 Featureanalyse	26
6.1 Math Datensatz	27
6.2 Read Datensatz	29
6.3 Science Datensatz	30
7 Fazit	33
Literaturverzeichnis	35
Anhang	37
Eidesstattliche Erklärung	42

ABKÜRZUNGSVERZEICHNIS

2SLS Two-Stage Least Squares

RF Random Forest

ABBILDUNGSVERZEICHNIS

Abb. 1:	Beispiel für einen Entscheidungsbaum mit Tiefe 2	6
Abb. 2:	Bias Variance TradeOff (Hastie, 2009)	8
Abb. 3:	Kumulierte Anzahl an Teilnahmen eines Landes an einem PISA Test der Kategorien Read (blau), Math (rot) und Science (grün).	11
Abb. 4:	Verteilung der Gesamtanzahl der erreichten PISA Scores für alle teilnehmenden Länder kumuliert.	12
Abb. 5:	Vorhersage des optimalen Modells MATH_65.....	18
Abb. 6:	Durchschnittlicher Fehler nach Jahren der Testdaten in Relation zu deren Auftreten.	19
Abb. 7:	Vorhersage des optimalen Modells READ_aggregated-years_55.....	20
Abb. 8:	Durchschnittlicher Fehler nach Jahren der Testdaten in Relation zu deren Auftreten.	21
Abb. 9:	Vorhersage des optimalen Modells SCIENCE _50.	22
Abb. 10:	Durchschnittlicher Fehler nach Jahren der Testdaten in Relation zu deren Auftreten.	22
Abb. 11:	Abnahme des R^2 je Alpha bei Pisa-Math.....	24
Abb. 12:	Ausschnitt eines Baumes für MATH_aggregated_years_75.....	26
Abb. 13:	Ausschnitt des Baumes für MATH_aggregated_years_75 mit Pruning bei einem Alpha=0.....	26
Abb. 14:	Zusammenhang zwischen der Varianz innerhalb eines Landes und der Korrelation des Features mit den Pisa-Daten am Beispiel des Datensatzes Math.....	27
Abb. 15:	Häufigkeit des Auftretens verschiedener Spaltennamen in den OECD Datensätzen.....	37

TABELLENVERZEICHNIS

Tab. 1:	Analyse der Datensätze Reading, Math und Science.....	10
Tab. 2:	Vergleich der besten Varianten der Single Trees für alle drei Datensätze.	16
Tab. 3:	Vergleich der Single Tree Modelle für den Datensatz Math mit Einbezug der Features des PISA Jahres.....	17
Tab. 4:	Vergleich der Single Tree Modelle für den Datensatz Read mit Einbezug der aggregierten PISA Jahre.	19
Tab. 5:	Vergleich der Single Tree Modelle für den Datensatz Science mit Einbezug der aggregierten PISA Jahre.	21
Tab. 6:	Vergleich der Performance der verschiedenen Alphas für das Pruning des optimalen Modells des Datensatzes Math.	23
Tab. 7:	Vergleich der Performance der verschiedenen Random Forest Modelle für die drei Datensätze.....	25
Tab. 8:	Vergleich der Top 3 Features des optimalen Modells für den Datensatz Math.	28
Tab. 9:	Vergleich der Top 3 Features des optimalen Modells für den Datensatz Read.	29
Tab. 10:	Vergleich der Top 3 Features des optimalen Modells für den Datensatz Science.....	31

1 EINLEITUNG

Alle drei Jahre kommentieren die großen Zeitungen Deutschlands die vermeintlich verbesserungswürdigen Ergebnisse deutscher Schüler innerhalb der international umfangreichsten (Michler, 2016) Bildungsstudie, der PISA Studie (welt.de, 2020) (br.de, 2019). Im Rahmen der Auseinandersetzung wird versucht, Einflussfaktoren auf die jeweiligen PISA Ergebnisse, anhand von für den Menschen kausal erscheinender Zusammenhänge, abzuleiten (br.de, 2019). Der Ansatz der vorliegenden Arbeit möchte sich dieser Auseinandersetzung von einer technischen Perspektive aus widmen. In diesem Rahmen soll eine Methode aus dem auftrumpfenden und sich gegenwärtig rasant entwickelnden Bereich des maschinellen Lernens verwendet wird. Die Erwartungen an diesen Ansatz bestehen aus der Möglichkeit völlig neue, unscheinbare Einflussfaktoren auf die Scores der PISA Studie zu finden und den Score im Rahmen der angewendeten Methode bestmöglich vorhersagen zu können.

Die vorliegende Arbeit beschäftigt sich somit mit der Fragestellung, ob Entscheidungsbäume mit Eingabe großer Datenmengen für die Vorhersage der PISA Scores geeignet sind. Hierfür wurden über 100.000 Features aus den OECD und Weltbank Datenbanken extrahiert, aufbereitet und zur Erstellung der Machine Learning Modelle verwendet. Das Entscheidungsbaum Modell wurde anhand verschiedener Hyperparameter optimiert, dessen Ergebnisse analysiert, diskutiert und mit den Ergebnissen von Pruning- und Random Forest Modellen verglichen. Zusätzlich wurde eine Featureanalyse durchgeführt, welche die Haupteinflussfaktoren auf die Scores der PISA Studie untersucht. Resultat der Arbeit ist eine gute Vorhersagbarkeit der PISA Scores in den drei Bereichen Math, Read und Science mittels teils geprunter Modelle eines Entscheidungsbaumes. Weiter konnte festgestellt werden, dass das Verwenden eines Ensemble Ansatzes, wie dem Random Forest, ohne spezifische Optimierung der Parameter nicht zur Optimierung des Modells beiträgt. Des Weiteren wurde beobachtet, dass die Verwendung einer maximalen Datenmenge als Featurebasis sowohl eine hohe Komplexität der Datenaufbereitung zur Folge hat, als auch zu Uneindeutigkeiten der Auswahl von Features für die Bäume führen kann. Weiter konnte festgestellt werden, dass eine Vielzahl der verwendeten Features als Proxys für die Länder und nicht die PISA Scores als solche fungieren, was Spielraum für schärfere Klassifizierungsmaßnahmen im Rahmen der Datenaufbereitung lässt.

Nach der Einführung in Kapitel 1, wird dem Leser in Kapitel 2 ein Überblick über den aktuellen Forschungsstand bezüglich PISA Daten und deren Analyse gegeben. Dabei wird die einschlägige Literatur rund um die Forschungsfrage diskutiert. Kapitel 3 stellt eine Einführung in die bei der Studie verwendete Methodik dar. Dabei werden Einsatz und Konstruktion von Entscheidungsbäumen sowie verschiedene Erweiterungen, wie Pruning und Random Forest, vorgestellt. Kapitel 4 beschreibt die verwendete

Datenbasis und deren Extraktion. In Kapitel 5 werden die erstellten Machine Learning Modelle und deren Ergebnisse beschrieben und diskutiert. Dabei wird die Frage adressiert, ob Entscheidungsbäume in der Lage sind, den PISA Score zu modellieren bzw. vorherzusagen. Kapitel 6, die Featureanalyse analysiert welche Variablen den PISA Score (im Modell) erklären können. Abschließend wird die Arbeit und deren Ergebnisse in Kapitel 7 zusammengefasst, kritisch hinterfragt und ein Ausblick für weitere Forschung gegeben.

2 LITERATURÜBERBLICK

Obwohl die Literatur eine Beantwortung der Forschungsfrage, welche dieser Arbeit zugrunde liegt, nicht hergibt, so gibt sie doch Einblick in die Arbeiten, welche sich mit sehr ähnlichen Fragestellungen beschäftigen.

Die Arbeit von Chon Ho Yu (Chong Ho Yu, 2012), „A Data Mining Approach to Comparing American and Canadian Grade 10 Students' PISA Science Test Performance“, befasst sich mit den Faktoren, die die Vorhersage der PISA Scores der amerikanischen und kanadischen Klasse 10 Schüler beeinflussen. Dabei arbeitet die Studie mit der Technik des Klassifizierungsbaums als Methode der Data Mining Ansätze. Im Unterschied zu der vorliegenden Arbeit wurden für die Vorhersage der PISA Scores und deren Erklärung der Raum der Einflussfaktoren auf die Antworten aus den Fragebögen der PISA-Studie begrenzt. Anstatt mit gesellschaftlichen oder wirtschaftlichen Einflussfaktoren befasst sich die Arbeit von Chon Ho Yu lediglich mit bildungsnahen Faktoren wie Wissenschaftsfreude oder dem Einsatz von Bildungssoftware.

Andere Arbeiten basieren zwar auf den PISA Ergebnissen und zielen darauf ab, diese zu analysieren, nutzen jedoch andere Ansätze des maschinellen Lernens als die Entscheidungsbäume. So auch die Arbeit von Arantza Gorostiagaa (Arantza Gorostiagaa, 2015), welche sich mit der Vorhersage von PISA Scores in Bezug auf spanische Schüler beschäftigt. Zu diesem Zweck werden die Anwendbarkeit der logistischen Regression, Fisher'sche Diskriminanzfunktions-Analyse und einer Support Vector Machine miteinander verglichen. Hierbei erzeugen die linearen und nichtlinearen Support Vector Machines, zulasten der Rechenkapazität, die genauesten Ergebnisse. Als Erklärungsgrundlage der Ergebnisse dienen die Antworten auf die im Rahmen der PISA Studie von Schülern und Lehrern ausgefüllten Fragebögen. Ähnlich befasst sich Fuchs (T.Fuchs, 2007) in seiner Arbeit mit der Analyse der PISA Ergebnisse anhand der aus den Fragebögen erhaltenen Hintergrundinformationen. Als Analysemethode dient hierfür eine Bildungs-Produktivitäts-Funktion.

Einige Arbeiten konzentrieren sich auf verschiedene gesellschaftliche oder soziologische Sachverhalte, welche mit der PISA Studie in Zusammenhang stehen. So untersucht Guiso (Luigi Guiso, 2008) in seiner Arbeit die PISA Scores und deren Aussage vor dem Hintergrund der Gender Gap, mit Hilfe einer Regressionsanalyse. Vandenbergh (Vincent Vandenbergh, 2016) hingegen widmet sich dem Einfluss von staatlicher im Vergleich zu privater Bildung auf die schulischen Leistungen von Kindern. Hierfür verwendet er eine Regressionsanalyse, den zweistufigen Ansatz von Heckmann, sowie ein Neigungs-Score-Matching. Auch Schnepf (S.Schnepf, 2007) nutzt eine Analyse der PISA Scores, um den Bildungsnachteil von Zuwanderern zu

untersuchen. Als Methodik kommt hierfür die OLS-Regression zum Einsatz. Die Arbeit von Jenkins (S. Jenkins, 2008) zielt in eine ähnliche Richtung. In ihr wird die Wohlstandslücke in Zusammenhang mit den PISA Scores mittels einfacher statistischer Verfahren analysiert. Auch Hanushek (E. Hanushek S. L., 2013) analysiert die PISA Scores mittels einer Panel Schätzung, um Rückschlüsse bezüglich der Sinnhaftigkeit von Schulautonomie zu erörtern. In seinem einschlägigen Werk „Economics of Education“ (E. Hanushek L. W.) werden diese Analysen weitergeführt und breiter diskutiert.

Vor einigen Jahren hat sich das Feld der Anwendung von Data Mining Ansätzen auf soziologische Fragestellungen weiter konkretisiert. Daraus ist der Begriff “Educational Datamining” hervorgegangen, welcher insbesondere in den Arbeiten von Romeo (C. Romero S. V., Educational data mining: a survey from 1995 to 2005, 2007), (C. Romero S. V., Data mining in education, Wiley Interdiscip. , 2013) konkretisiert wird. In diesen wird die aktuelle Arbeit in diesem Bereich zusammengefasst und kategorisiert.

Abseits von PISA stößt man auf Literatur, die sich auf das Erarbeiten von Modellen, darunter auch Entscheidungsbäumen in anderen bildungsnahen Kontexten befasst. So nutzt Vandamme (J.P. Vandamme, 2007) Diskriminanzanalyse, Neuronale Netzwerke, Random Forests und Entscheidungsbäume, um den akademischen Erfolg von Studenten vorher zu sagen. Auch Romeo (C. Romero S. V., 2008) vergleicht in seiner Arbeit eine Vielzahl an Data Mining und Klassifikationsmethoden auf Basis derer Studenten anhand ihrer Moodle Daten gruppiert werden sollen. Hierbei kommt auch ein Entscheidungsbaum zum Einsatz. Naik (B. Naik, 2004) hingegen beschäftigt sich in seiner Arbeit mit neuronalen Netzen, mit Hilfe derer er den Erfolg von MBA Studenten vorhersagen möchte.

Aus diesem Literaturüberblick wird eine Forschungslücke erkennbar, die durch diese Arbeit geschlossen werden kann. Eine Beurteilung der Vorhersagemöglichkeit von PISA Scores, mittels eines Entscheidungsbaums und unter Verwendung von ~30.000 Features wird einen Beitrag zu bisheriger Forschung leisten.

3 THEORIE

Das Spektrum der verschiedenen Modelle im Bereich des maschinellen Lernens ist vielfältig. Nichtsdestotrotz bietet der in dieser Arbeit verwendete Entscheidungsbaum eine gute Grundlage für weitere Optimierungen und komplexere Modelle. Für die vorliegende Arbeit werden zunächst der einfache Entscheidungsbaum, das Optimierungsverfahren Pruning und das Random Forest Modell betrachtet. Die Modelle Lernens verfügen über verschiedene Hyperparameter, die es zu optimieren gilt, um das beste Modell gem. einer vorher festgelegten Metrik auszuwählen. Für die Optimierung wird in der Praxis auf das Optimierungsverfahren Grid Search zurückgegriffen, dass alle möglichen Kombinationen der Hyperparameter miteinander auf einem vorher festgelegten Wertebereich testet. Anhand der errechneten Metrik, die als Vergleichsmaß für die verschiedenen Modelle gilt, kann dann das beste Modell ausgewählt werden. Um den Suchraum etwas einzuschränken und die Komplexität des Problems zu reduzieren, wird oft davor eine Random Search durchgeführt. Dabei werden die Intervalle für die Grid Search bestimmt, indem iterativ mit einem Approximationsverfahren ein Wertebereich gefunden wird, der die optimale Lösung enthält.

3.1 Entscheidungsbäume

Entscheidungsbäume stellen eine Unterklasse von Verfahren des maschinellen Lernens dar, welches ein Teilgebiet von Künstlicher Intelligenz ist. Sie können den Verfahren des Überwachten Lernens zugeordnet werden, bei dem die Vorhersagewerte bereits bekannt sind und während des Trainings des Modells verwendet werden (Gareth James, 2017, S. 303-307).

Entscheidungsbaum Modelle können im Kern als eine Art verschachtelter „wenn, dann“ Bedingungen gesehen werden. Das klassische Anwendungsgebiet von Entscheidungsbäumen sind Klassifikationsprobleme, also das Zuordnen einer Instanz von einer Menge an Beobachtungen zu einer bestimmten Klasse/Kategorie. Neben Klassifikationsproblemen lassen sich auch Regressionsprobleme mit Entscheidungsbaummodellen lösen. Dieser alternative Lösungsansatz für Regressionsprobleme ist i.d.R. ungenauer bzw. liefert schlechtere Vorhersagen, aber bietet durch die Einfachheit und Interpretierbarkeit dieser Modelle andere Vorteile (Gareth James, 2017, S. 303-307).

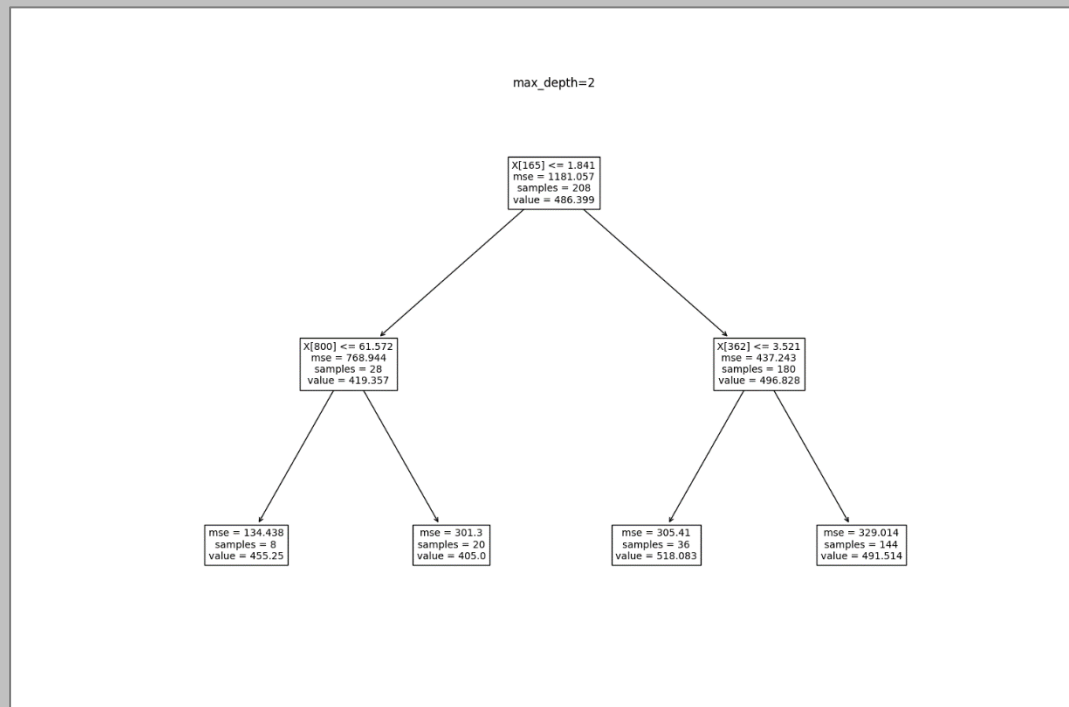


Abb. 1: Beispiel für einen Entscheidungsbaum mit Tiefe 2

Die Baumstruktur (Visualisierung) bei Entscheidungsbaum ergibt sich durch das Aufsplitten des Datensatzes anhand verschiedener Features, die der Algorithmus selbst auswählt. Die geschachtelte Struktur kommt dadurch zustande, dass nachdem der Datensatz einmal aufgesplittet wurde, die entstandenen Teilmengen sukzessiv anhand weiterer Features aufgesplittet werden, bis ein bestimmtes Abbruchkriterium erreicht wurde. Die nach dem letzten Split resultierenden Teilmengen (Nodes) werden als Leafs und alle vorigen Teilmengen als Internal Nodes bezeichnet. Dabei lassen sich letztere in Parent Nodes und Childnodes aufteilen. Als Parent Node wird immer der Node bezeichnet der über den beiden Childnodes liegt. Mit Ausnahme des ersten Nodes der den Ausgangspunkt bzw. die Wurzel (engl. „root node“) des Entscheidungsbaums darstellt. Durch das Definieren verschiedener Hyperparameter wie z.B. Splitting Metrik, Baumtiefe bzw. Anzahl der Blätter oder Abbruchkriterium, kann der Anwender den Baumbildungsprozess beeinflussen, den Lösungsraum begrenzen und Overfitting entgegenwirken. Die Entscheidung bzgl. Der Teilung der Daten trifft der Entscheidungsbaum selbst, indem er das Feature auswählt das bei vorgegebener Splitting Metrik (Hyperparameter) das beste Ergebnis liefert. Bei Regressionsproblemen und binärer Splitting Methodik haben die Features immer folgende Struktur, wenn man eine einzelne Variable betrachtet: *Wenn Wert der Variable bei Instanz X größer oder gleich Y dann wird Instanz X Teilmenge A zugeordnet, falls nicht dann wird Instanz X zu Teilmenge B zugeordnet* (Gareth James, 2017, S. 303-307).

Einer der am häufigsten für die Konstruktion von Entscheidungsbäumen verwendeten Algorithmen ist der CART (Classification and Regression Tree) Algorithmus.¹ Dieser folgt dem folgenden Prinzip:

$$\text{Minimiere} \rightarrow J(k, t_k) = \frac{m_{\text{left}}}{m} \text{MSE}_{\text{left}} + \frac{m_{\text{right}}}{m} \text{MSE}_{\text{right}}$$

$$\text{mit} \begin{cases} \text{MSE}_{\text{node}} = \sum_{i \in \text{node}} (\hat{y}_{\text{node}} - y^{(i)})^2 \\ \hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)} \end{cases}$$

$J(k, t_k)$ ist die Kostenfunktion, die minimiert werden soll. Sie besteht aus zwei Termen, von denen der erste Term ($\frac{m_{\text{left}}}{m} \text{MSE}_{\text{left}}$) den MSE innerhalb des linken Nodes und der zweite Term ($\frac{m_{\text{right}}}{m} \text{MSE}_{\text{right}}$) den MSE innerhalb des rechten Nodes nach dem Split enthält. Der MSE innerhalb der Nodes ist durch die Varianz von y bzw. die quadrierte Abweichung der Zielwerte von ihrem Mittelwert innerhalb des jeweiligen Nodes gegeben. Dies führt dazu, dass die Beobachtungen innerhalb des Nodes möglichst ähnliche y -Werte aufweisen wodurch eine Art Cluster entsteht. Der multiplikative Term vor dem MSE sorgt für eine Gewichtung der beiden Terme innerhalb der Kostenfunktion, indem er jeweils den Anteil ($\frac{m_{\text{left}}}{m}, \frac{m_{\text{right}}}{m}$) der Beobachtungen des jeweiligen Childnodes an dem Parentnode darstellt. Dadurch wird erreicht, dass der Algorithmus nicht immer die triviale Lösung (falls ohne Gewichtung) wählt und nur eine Beobachtung von dem Rest der Beobachtungen separiert (Géron, 2019)

Die Performance des Entscheidungsbaums bei Regression kann durch den RMSE oder R^2 beurteilt werden. Der R^2 wird durch folgende Formel berechnet:

$$R^2 = \frac{\text{MSE} - \text{RSS}}{\text{MSE}} = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

RSS beschreibt die Summe der Residuen der Regression. Im Gegensatz zum MSE, wird somit die Abweichung der beobachteten Werte von den prognostizierten Werten durch das Modell betrachtet. Entsprechen die prognostizierten Werte exakt den beobachteten Werten ist RSS gleich 0 und der R^2 nimmt den Wert 1 an. Dies bedeutet, dass 100% der Variation in Y durch das Modell erklärt werden kann (Gareth James, 2017, S. 79-80).

¹ Der CART Algorithmus wird von der ML Bibliothek „scikit learn“, die bei der Modellierung in Teil 5 benutzt wurde, in einer optimierten Version verwendet.

Der RMSE ist die Wurzel des MSE über alle Vorhersagen des Entscheidungsbaumes und wird durch folgende Formel beschrieben:

$$RMSE = \sqrt{\frac{1}{n - p - 1} RSS} = \sqrt{\frac{1}{n - 2} \sum (y_i - \hat{y}_i)^2}$$

n ist die Anzahl der Beobachtungen und p spiegelt die Anzahl an Freiheitsgraden wieder. Beides sind Koeffizienten des Regressionsmodells. Die Konstante des Regressionsmodells wird durch die -1 im Nenner des Bruchs abgebildet. Der RMSE ist im Vergleich zum MSE einfacher zu interpretieren, da er die Verzerrung des Ergebnisses durch die Quadrierung der Residuen aufhebt und sich das Ergebnis so auf derselben Skala wie die einzelnen Werte y_i befindet. Er kann als Maß für die durchschnittliche Abweichung vom beobachteten Wert interpretiert werden und ähnelt daher der Standardabweichung, die für die Beschreibung der Verteilung einer Datenmenge genutzt werden kann (Gareth James, 2017, S. 79-80).

3.2 Pruning

Je tiefer ein Entscheidungsbaum desto wahrscheinlicher ist es, dass dieser overfitted. Um dies zu vermeiden kann die Baumtiefe begrenzt werden. Ein anderer Ansatz ist die Anwendung des Prunings (zu Deutsch: „Stutzen“), der dem Lasso Ansatz bei der Linearen Regression ähnelt. Pruning kann den Regularisierungstechniken zugeordnet werden, die das Ziel verfolgen eine bessere Generalisierung des Modells zu erreichen. Dabei wird die Splitting Metrik bzw. Kostenfunktion noch um einen additiven Term αT erweitert. T beschreibt die Anzahl der Blätter und α ist ein damit multiplizierter Faktor, der die Kosten pro Knoten beschreibt. Je höher α gewählt wird, desto weniger Blätter soll der Baum haben bzw. desto höher ist die „Strafe“ pro Knoten (Gareth James, 2017, S. 307-308).

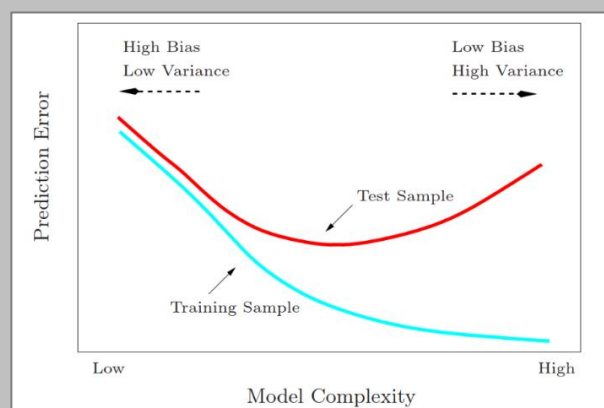


Abb. 2: Bias Variance TradeOff (Hastie, 2009)

Je höher der Parameter α gewählt wird, desto stärker wird auch die Komplexität des Baumes reduziert. Dies bringt Vorteile bezüglich der Interpretierbarkeit der Ergebnisse und reduziert den Vorhersagefehler gem. der gewählten Metrik auf den Testdaten. Ziel ist es α im Rahmen der Modellierung so zu wählen, dass wie in Abb. 2 dargestellt, ein Optimum zwischen Komplexitätsreduktion und Anpassungsfähigkeit (Varianz) des Modells erreicht wird. Mittels Kreuzvalidierung und einem Optimierungsalgorithmus wie z.B. Grid Search lässt sich dieser Hyperparameter α bestimmen (Gareth James, 2017, S. 307-308).

3.3 Ensemble Methoden

Entscheidungsbäume werden häufig in Verbindung mit Ensemble Methoden wie Bagging und Boosting eingesetzt, um Overfitting zu reduzieren und Generalisierung des Modells zu maximieren. Bagging steht für „Bootstrap Aggregating“ und beschreibt eine Sampling Methode bei der durch ziehen-mit-zurücklegen eine Reihe von Trainingsdatensätzen für verschiedene Modelle entstehen (Gareth James, 2017, S. 319-321).

Besonders verbreitet ist das Random Forest Modell, das mit einer festgelegten Anzahl von Entscheidungsbäumen arbeitet. Mittels Bagging entstandene Teilmengen bilden dabei die Basis für das Training dieser Bäume. Dabei wird für jede Teilmenge (Features und Beobachtungen) ein neuer Entscheidungsbaum konstruiert. Durch das zufällige Auswählen einer Teilmenge an Features (Unterschied zu Bagging), die bei jedem Split betrachtet werden dürfen, wird zudem noch eine Dekorrelation der Bäume erreicht. Dies wirkt dem Problem entgegen, dass verschiedene erklärende Variablen in der Praxis häufig miteinander korreliert sind. Durch zufälliges Splitting der Datenbasis soll, ähnlich wie bei Kreuzvalidierung, eine Generalisierung des Modells sichergestellt werden. Die Vorhersage des Modells wird aus den Vorhersagen der einzelnen Bäume berechnet, indem z.B. bei der Regression ein Mittelwert über alle Vorhersagen gebildet wird. Die Grundidee dahinter ist, dass viele verschiedene Vorhersagen im Mittel eine bessere bzw. robustere Vorhersage treffen können als ein einzelner Entscheidungsbaum („wisdom of the crowd effect“). Auch beim Random Forest gibt es verschiedene Hyperparameter, die, auf der Suche nach dem optimalen Modell, optimiert werden müssen. Die Anzahl der Bäume und alle Hyperparameter bzgl. eines einzelnen Baumes, wie in Abschnitt 3.2 beschrieben, gehören dazu (Gareth James, 2017, S. 319-321).

4 DATEN

Zur Vorhersage der PISA Scores mittels eines Entscheidungsbaums benötigt dieser wie in Kapitel 3 beschrieben, zwei Arten von Daten: die PISA Scores, welche als Zielvariablen bzw. Target dienen, sowie Features, welche einen besonders großen Einfluss auf die Bildung der PISA Scores haben. Hierfür wurden die Daten der Open Source Datenbanken der OECD und der Weltbank verwendet.

4.1 PISA Daten

Die PISA Studie findet seit dem Jahr 2000 im drei-jährigen Rhythmus statt. 70 OECD-Länder, Partnerländer und andere Volkswirtschaften nehmen daran teil. 500.000 15-jährige werden auf Ihre Lese-, Mathematik- und Naturwissenschaftskompetenzen mittels realer Fragestellungen getestet. Insgesamt soll eine Gesamtpopulation von 26 Millionen 15-jährigen, aus allen Schichten, mittels eines speziellen Stichprobenverfahren abgebildet werden. Dieses beinhaltet die drei Phasen „Auswahl“, „Durchführung“ und „Auswertung“. Im Rahmen der „Auswahl“ wird die Schülerpopulation im initialen Schritt nach bestimmten Kriterien in Schichten eingeteilt. Auf Basis dieser Schichten erfolgt die Auswahl der Schulen mit Hilfe von Wahrscheinlichkeiten, welche proportionale zu der Schulgröße sind. Pro Schule werden zufällig 35 Schüler ausgewählt. Die „Durchführung“ beinhaltet das zweistündige Bearbeiten von Bleistift und Papier Aufgaben sowie das Ausfüllen eines Fragebogens hinsichtlich persönlicher Hintergrundinformationen, Lerngewohnheiten, Engagement und Motivation beim Lernen durch die Schüler. Des Weiteren füllt auch die Schulleitung einen Fragebogen bezüglich der demografischen Merkmale und des Lernumfelds der Schüler aus. Im Schritt der „Auswertung“ kommen groß angelegte Bewertungsprogramme zum Einsatz, welche mittels plausiblen Werten die Schülerleistung beurteilen (OECD, 2000).

<i>Kennzahl</i>	<i>Read</i>	<i>Math</i>	<i>Science</i>
Zeitraum	2000 - 2018	2003 - 2018	2006 - 2018
Standardabweichung	34,63	41,52	36,97
Mittelwert	485,8	486,7	490,33
Median	475,75	495	497
Beobachtungen	260	231	197

Tab. 1: Analyse der Datensätze Reading, Math und Science

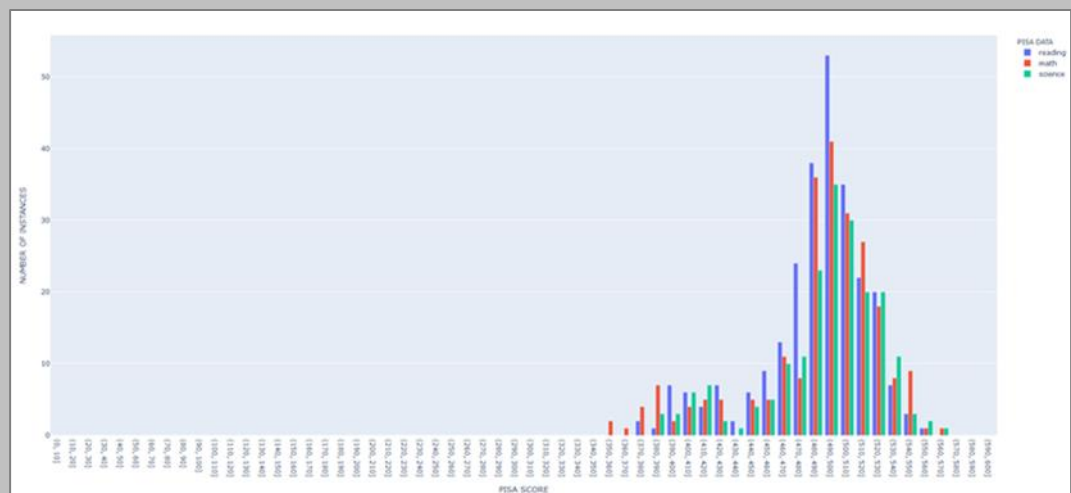


Abb. 4: Verteilung der Gesamtanzahl der erreichten PISA Scores für alle teilnehmenden Länder kumuliert.

In Abbildung 4 sind die Anzahlen der Scores der einzelnen Testkategorien zu sehen. Auffällig ist eine Ballung der Scores um das Intervall [340, 580]. Darüber hinaus kann beobachtet werden, dass die PISA Scores eines Landes in den meisten Fällen nur sehr gering variieren. Der Variationskoeffizient² beträgt im Mittel ca. 2%.

4.2 Featuredaten

Um die Features mit dem größtmöglichen Einfluss auf die PISA Scores für die Vorhersage nutzen zu können, wurden dem Entscheidungsbaum so viele Variablen wie möglich zur Modellbildung zugeführt. Da der Entscheidungsbaum die Variablen mit dem größten Informationsgehalt für die Vorhersage eigenständig auswählt, können auch Features das Modell beeinflussen, die bei einer manuellen Auswahl wegen vermeintlicher Zusammenhangslosigkeit aussortiert worden wären. Die OECD und die Weltbank gelten hierbei als große öffentlich verfügbaren Datenquellen, deren Kategorien von Wirtschafts- über Nachhaltigkeits- bis hin zu Entwicklungskennzahlen reichen.

Der Aufwand zur Extraktion der OECD Daten war hoch. So mussten für das Ausführen der API die benötigten Feature-Schlüssel aus dem Quelltext der Datenbank-Seite der OECD gewonnen werden. Mit diesen konnten mittels einer programmierten

² Der Variationskoeffizient bestimmt sich aus dem Verhältnis von Standardabweichung zum Mittelwert.

Automatisierung (siehe 0), über 100.000 Features für alle Länder der PISA Studie im Zeitraum 1998-2018 heruntergeladen und aufbereitet werden.

Die Datensätze beider Quellen wiesen eine Reihe von NA's³ auf. Um die Daten dennoch bestmöglich für die Vorhersage nutzen zu können, wurde sofern möglich, eine Interpolation der Werte innerhalb eines Landes angewendet. War dies nicht möglich, wurden fehlende Werte durch den jährlichen Durchschnitt aller Länder approximiert⁴.

³ Leerstellen

⁴ im Folgenden auch als Annual Mean bezeichnet

5 MODELLANALYSE

Um die beste Vorhersage für den PISA Score zu treffen, wurden verschiedene Modelle auf den Daten trainiert und deren Ergebnisse anschließend miteinander verglichen. Um herauszufinden welche Hyperparameter bei der Modellierung mit Entscheidungsbäumen auf den PISA Daten einen Einfluss auf die Vorhersage haben, wurden zuerst verschiedene (einzelne) Bäume trainiert. Dabei wurde zunächst für jeden Hyperparameter ein Wertebereich definiert, um die Auswirkungen der einzelnen Parameter auf die Ergebnisse zu beobachten. Da es sich bei dem PISA Scores um numerische Werte handelt, wurde für die Vorhersagen ein Entscheidungsbaum Regressor verwendet.⁵

Alle Modelle in dieser Studie wurden mit Hilfe der Python Bibliothek „Scikit Learn“ erstellt, die für die Konstruktion eine Abwandlung des CART Algorithmus verwendet. Aufbauend auf der Exploration der Hyperparameter wurde das Modell bzgl. der Hyperparameter maximale Baumtiefe, minimale Zahl der Beobachtungen pro Split und minimale Anzahl der Beobachtungen pro Leaf mittels einer Grid Search optimiert. Für die Grid Search wurde eine Generalisierung des Modells mittels 7-facher Kreuzvalidierung sichergestellt. Um die Modelle abschließend zu testen und repräsentative Ergebnisse zu erhalten wurde ein Trainings-/Testsplit durchgeführt, sodass immer 15% Testdaten nach dem Training bereitstehen. Dieser Split wurde für die Analyse aller Modelle identisch angewendet.

Ein einzelner Baum neigt i.d.R zu Overfitting und daher wird in der Praxis häufig mit Erweiterungen des einfachen Entscheidungsbaummodells gearbeitet (vgl. Kapitel 3). Im Rahmen der Studie wurde zuerst ein einzelner Entscheidungsbaum (Single Tree) anhand verschiedener Hyperparameter optimiert. Darauf aufbauend wurde dieser optimierte Baum mittels Post-Pruning weiter optimiert. Als letzten Schritt wurden eine Reihe verschiedener Random Forest Modelle auf den Daten trainiert und getestet. Bezüglich des Datensatzes sind folgende zwei Aspekte im Zusammenhang mit den Daten in der Modellierung berücksichtigt worden:

Je mehr Daten/Features dem Modell zur Verfügung stehen, desto größer ist die Wahrscheinlichkeit, dass auch einige dieser Features Erklärungskraft bzgl. des PISA Scores enthalten. Durch den zugrundeliegenden Algorithmus (vgl. Kapitel 3) ist der Entscheidungsbaum in der Lage diese Features zu erkennen und zu nutzen, um bessere Vorhersagen zu treffen. Der Einfluss der Interpolation mittels Funktion oder Mittelwert

⁵ Es wäre an dieser Stelle aber ebenso möglich gewesen die Daten in verschiedene Klassen („bins“) aufzuteilen und mit Klassifizierung zu arbeiten.

ist dabei ein wichtiger Faktor. Denn wenn z.B. 50% der Beobachtungen eines Features interpoliert oder mit dem Mittelwert ersetzt wurden und dieses zur Vorhersage des Modells einen Beitrag leistet, stellt das die Konsistenz dieses Features und auch des trainierten Modells in Frage. Um trotzdem möglichst viele Features nutzen zu können wurde ein Grenzwert bzgl. der NA's pro Feature ermittelt. Da primär die Vorhersagegenauigkeit im Fokus steht, wurde im Rahmen der Studie über ein Intervall von [50%, 100%] non-NA's pro Feature jedes Modells optimiert. Anschließend wurde mittels eines Robustheitsmaßes (% der Beobachtungen, die geschätzt wurden) jedes Feature auf seine Sinnhaftigkeit überprüft, um bei der Interpretation falsche Rückschlüsse zu vermeiden.

Der PISA Score stellt wie jede Prüfung eine Momentaufnahme (alle 3 Jahre) dar, genau wie auch die Werte der Features bezüglich der Länder für jedes einzelne Jahr. Es ist davon auszugehen, dass auch die Information über die Jahre vor der Erhebung des PISA Scores relevant für die Vorhersage sind, da anzunehmen ist, dass z.B. höhere Ausgaben für Bildung sich erst nach einer gewissen Zeit auf die Leistung der Schüler auswirken. Deshalb gibt es zwei Möglichkeit die Features zu verwenden. Eine Möglichkeit besteht darin, nur den Wert des Features für das Jahr zu nutzen, für das auch ein PISA Score vorliegt. Eine zweite Möglichkeit wäre die Werte des Features für das jeweilige und die Werte der 2 vorherigen Jahre zu aggregieren und z.B. gewichtet zu mitteln. Beide Möglichkeiten wurden für alle Modelle parallel zum Umgang mit den NA- Werten in Betracht gezogen und verglichen.

Ein Datenpunkt bzw. eine Beobachtung ist durch ein Land und ein Jahr gegeben, z.B. Deutschland 2006. Das Land wurde dem Algorithmus bewusst nicht als Feature übergeben, da die Vorhersage länderunabhängig sein soll, sondern gezielt nach anderen weniger offensichtlichen Indikatoren gesucht wird.

5.1 Entscheidungsbäume

<i>Dataset_non_NA</i> [%]	<i>Features</i> <i>in Tree</i>	<i>RMSE</i>	<i>R²_Test</i>	<i>R²_Train</i>	<i>max_</i> <i>depth</i>	<i>min_leaf</i>	<i>min_split</i>
MATH_agg_50	36	15,83	88,14%	99,29%	7	3	7
MATH_65	26	11,33	93,93%	99%	8	5	3
READ_agg_55	11	13,36	86,83%	93,11%	4	4	9
READ_90	16	16,35	80,27%	92,99%	6	10	3

SCIENCE_agg_100	7	20,19	79,86%	92,36%	8	3	3
SCIENCE_50	46	16,10	87,19%	99,53%	9	3	3

Tab. 2: Vergleich der besten Varianten der Single Trees für alle drei Datensätze.

In Tabelle 2 sind die Ergebnisse der besten Modelle bzgl. Hyperparameterkombination und maximale Schwelle in Prozent an NA Werten pro Feature angegeben. Außerdem wurden mit aggregierten Werten bzgl. Zeitpunkt der PISA Score Erhebung und nicht aggregierten Werten Modelle trainiert. Die besten Entscheidungsbaum Modelle für die jeweiligen Datensätze erreichen Werte für R^2 zwischen 86,83% und 93,93% und für RMSE zwischen 16,10 und 11,33 in den Testdaten. Dies legt nahe, dass die trainierten Modelle in der Lage sind im Schnitt mit bis zu einer Genauigkeit von $\pm 11,33$ den PISA Score vorherzusagen. Auffällig ist hier das gerade bei dem Math Datensatz das beste Ergebnis erzielt wurde, obwohl dieser mit 41,52 die höchste Standardabweichung aufweist (vgl. Kapitel 4). Mit Blick auf die Performance des Math Datensatzes mit aggregierten Jahren fällt auf, dass diese Modelle durchschnittlich deutlich schlechter abschneiden als die Modelle in denen lediglich PISA Jahre betrachtet wurden. Dies lässt auf eine höhere Variabilität der in diesen Modellen verwendeten Feature Daten innerhalb der Jahre schließen. Am Zweitbesten lassen sich die PISA Scores des Read Datensatzes vorhersagen und am schlechtesten schneiden die Vorhersagen zur schulischen Leistung im Bereich Science ab. Für die Performance der aggregierten Science Modelle fällt auf, dass hier das Modell mit einem 0%-Anteil an NA-Werten am besten abschneidet. Das ist deswegen überraschend, da diesem Modell lediglich 7 Features zur Verfügung standen, welche auch genutzt wurden. Zwischen der Anzahl an Beobachtungen und der Performance der Modelle der Datensätze (vgl. Kapitel 4) kann kein eindeutiger Zusammenhang hergestellt werden. Auch über die Hyperparameter kann über alle Datensätze hinweg keine allgemeingültige Aussage getroffen werden, so bedienen sich alle optimalen Modelle unterschiedlicher Parameter. Betrachtet man die Metrik R^2 für die Trainingsdaten, und vergleicht diese zwischen den besten Modellen als auch innerhalb der Datensätze, so kann kein Zusammenhang zwischen dem R^2 der Tests und dem R^2 der Trainings festgestellt werden. Einige Modelle weisen zwar einen besonders guten R^2 für die Trainings auf, schneiden in den Tests aber schlecht ab. Dies kann darauf hinweisen, dass die ausgewählten Trainings- und Testdaten sehr spezifisch sind und trotz der Kreuzvalidierung schlecht generalisiert werden.

5.1.1 Math Datensatz

<i>Non-NA [%]</i>	<i>Total Features</i>	<i>Features</i>	<i>RMSE</i>	<i>R²- Test</i>	<i>R²- Train</i>	<i>max_ depth</i>	<i>min_leaf</i>	<i>min_split</i>
50	34274	18	14,70	89,77%	98%	6	4	11
55	28035	23	15,60	88,48%	99%	8	5	11
60	24559	31	11,66	93,57%	99%	10	2	10
65	19652	26	11,33	93,93%	99%	8	5	3
70	15417	76	11,72	93,50%	100%	10	2	3
75	12380	49	17,80	85,00%	100%	7	2	3
80	10547	25	13,57	91,29%	99%	7	2	10
85	6257	15	13,90	90,85%	97%	5	2	12
90	4545	34	11,94	93,25%	99%	8	4	3
95	2455	27	13,59	91,33%	99%	7	2	8
100	27	10	27,51	64,19%	84%	9	8	3

Tab. 3: Vergleich der Single Tree Modelle für den Datensatz Math mit Einbezug der Features des PISA Jahres

Tabelle 3 und zeigt die Ergebnisse der trainierten Entscheidungsbaumregressoren auf den Test Daten für den Math Datensatz, welcher lediglich die PISA Jahre miteinbezieht. Bei einer Auswahl von maximal 65% non-NA Einträgen pro Feature und mit den Parametern 8, 5, 3 für die maximale Baumtiefe, minimale Beobachtungen pro Leaf sowie minimale Anzahl an Beobachtungen pro Split, konnte ein RMSE von 11,33 und ein R^2 von 93,93% erreicht werden. Bei 50% non-NA Werte, also wenn im Schnitt genau 50% der Daten künstlich hinzugefügt wurden, stehen dem Algorithmus 34.274 Features zur Verfügung von denen er aber letztendlich lediglich 18 für die Vorhersage nutzt. Auf der anderen Seite wurden im Fall der 100% non-NA Werte keine Werte künstlich hinzugefügt, was bei nur 27 Features der Fall ist. Ein proportionaler Zusammenhang zwischen dem Hinzufügen von künstlichen Werten und der Vorhersagegenauigkeit ist nicht ersichtlich. So konnten bei non-NA Werten von 60%, 65%, 70% und 90% ähnlich gute Werte für RMSE und R^2 erreicht werden. Der Prozentsatz von 12,89%, welcher auf den Anteil der geschätzten Werte für nicht verfügbare Features hinweist (siehe Anhang), liegt bei diesem, optimalen Baum, im oberen Mittelfeld aller gemessenen Anteile. Alle Regressoren sind in der Lage

ungeachtet von Hyperparametern und Datensatz die Varianz in den PISA Daten mit einer Genauigkeit von 85% bis 93,93% zu erklären. Ausgenommen davon ist nur der Regressor, der mit 100% non-NA Werten arbeitet und einen Ausreißer darstellt. Dieser Ausreißer ist möglicherweise damit zu erklären, dass diesem Regressor wenig Features zur Verfügung stehen bzw. auch nur wenige Features vom Baum genutzt werden. Alle anderen Bäume benutzen maximal 76 Features und der beste Baum benutzt genau 26 Features für die Vorhersage der PISA Scores (Math).

In Abbildung 5 ist das optimale Modell für den Math Datensatz graphisch dargestellt. So beschreibt die Linie die vorhergesagten Scores der Länder in der PISA Studie, während die Punkte die tatsächlichen Scores darstellen. Zu erkennen ist, dass viele vorhergesagten Werte nah an den tatsächlichen liegen.

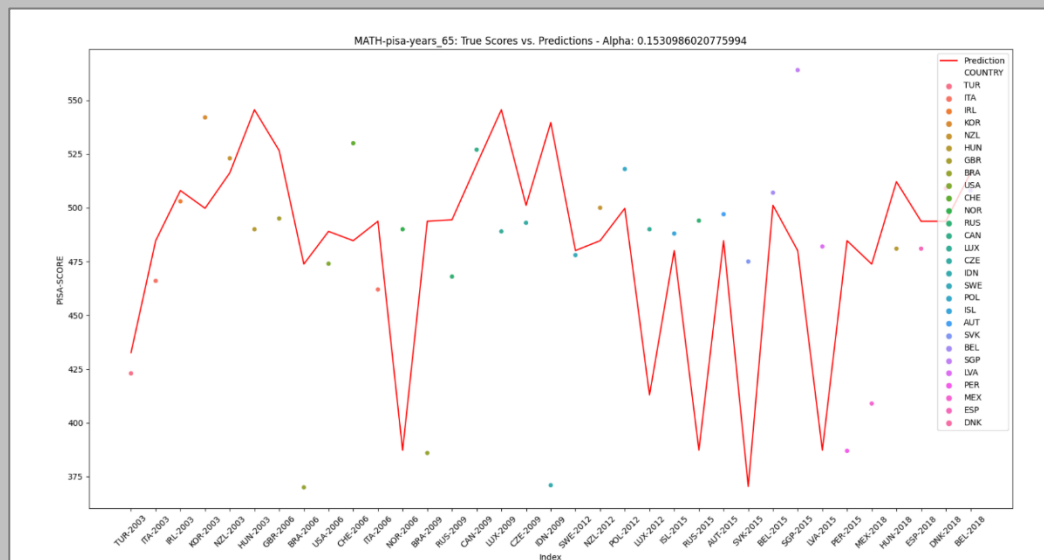


Abb. 5: Vorhersage des optimalen Modells MATH_65.

Weitere Gütemaße bezüglich der Vorhersage lassen sich aus Abbildung 6 ableiten. Aus Abbildung 6 geht hervor, dass die Scores der PISA Studie zum Jahr 2018 am besten vorhergesagt werden konnten, während der durchschnittliche Fehler im Verhältnis zur Häufigkeit der Scores aus 2012 den Rückschluss zulassen, dass diese Scores in irgendeiner Weise ungewöhnlich waren und nicht ins Modell gepasst haben.

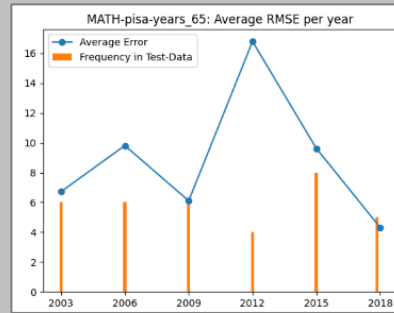


Abb. 6: Durchschnittlicher Fehler nach Jahren der Testdaten in Relation zu deren Auftreten.

5.1.2 Read Datensatz

<i>Non-NA [%]</i>	<i>Total Features</i>	<i>Features</i>	<i>RMSE</i>	<i>R²- Test</i>	<i>R²- Train</i>	<i>max_ depth</i>	<i>min_leaf</i>	<i>min_split</i>
50	24202	11	16,08	80,90%	93,11%	4	4	11
55	19414	11	13,36	86,83%	93,11%	4	4	9
60	16101	32	15,42	82,43%	98,24%	7	4	3
65	13649	41	17,08	78,45%	98,78%	7	3	3
70	11656	13	15,58	82,06%	93,11%	4	3	7
75	8432	14	21,96	64,39%	94,26%	5	7	3
80	6096	34	21,76	65,04%	98,31%	9	3	10
85	4354	40	20,68	68,43%	98,78%	10	2	9
90	3190	44	17,41	77,61%	98,78%	9	2	8
95	1856	20	19,01	73,30%	93,52%	5	3	7
100	7	7	17,89	76,36%	88,45%	6	5	3

Tab. 4: Vergleich der Single Tree Modelle für den Datensatz Read mit Einbezug der aggregierten PISA Jahre.

Führt man die gleiche Analyse für den Read Datensatz mit aggregierten Werten durch, welcher den optimalen Baum für diesen Datensatz enthält, so können bei Betrachtung der Tabelle 4 einige Unterschiede zu den Ergebnissen des Math Datensatzes beobachtet werden. So kann hier ein Zusammenhang zwischen dem Anteil der Non-NA Werte und der Performance des Modells durchaus gezogen werden. Während die Performance der

Vorhersage der Modelle mit einem Non-NA Anteil zwischen 50 und 70 im Intervall $[78,45\%; 86,83\%]$ liegt, befinden sich die Modelle mit einem Non-NA Anteil zwischen 75 und 100 in einem Intervall von $[64,39\%; 76,36\%]$. Dies lässt darauf schließen, dass die teilweise konstruierten Features, welche bei einer höheren Schätzer-Toleranz vom Modell genutzt werden können, wichtige Entscheidungsvariablen für ein gutes Modell darstellen, dieses aber gleichzeitig auch weniger robust sein kann. Anders als im Math Datensatz bildet das Modell mit einem Non-NA Anteil von 100% keinen Ausreißer. Trotz der lediglich 7 Features, die es zur Vorhersage heranziehen kann, performt es besser als vier der anderen Modelle.

Abbildung 7 zeigt, insbesondere im Vergleich mit Abbildung 6 des Datensatzes Math, dass die Abweichungen der Vorhersage zu den tatsächlichen Werten an einigen Stellen deutlicher sind.

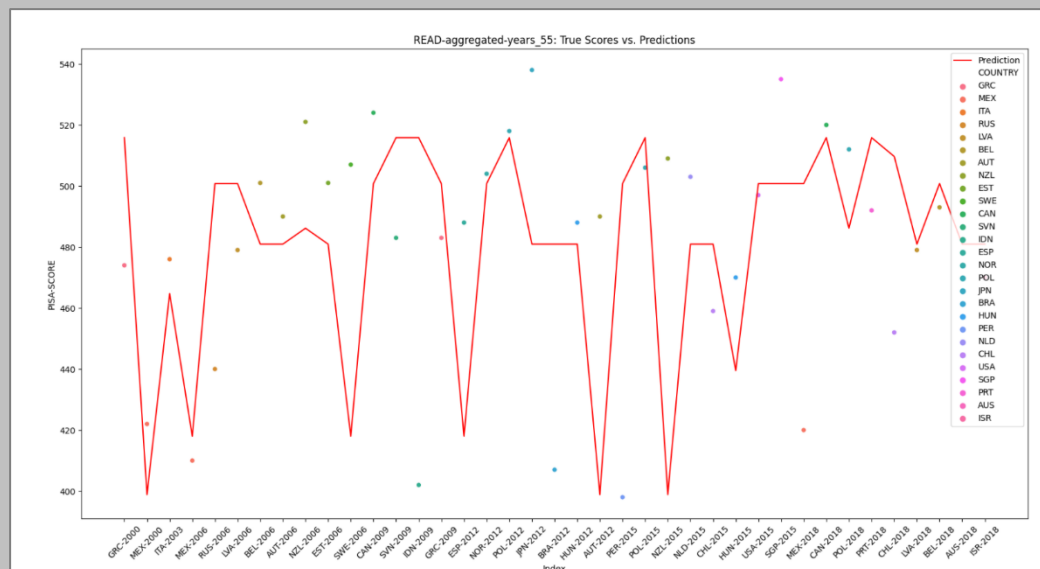


Abb. 7: Vorhersage des optimalen Modells READ_aggregated-years_55.

So weicht die Vorhersage des PISA Scoress für Read für die USA 2015 deutlich von dem realen Ergebnis ab.

Abbildung 8 weist darauf hin, dass der durchschnittliche Fehler, in den Jahren 2000 und 2012 besonders dominant auftraten, während die Daten in 2018 und 2006 kaum Fehler zuließen. Die hohe Fehlerrate im Jahr 2000 könnte damit zusammenhängen, dass der PISA Test in diesem Jahr zum ersten Mal erhoben und dementsprechend noch nicht auf einem hohen Niveau standardisiert durchgeführt werden konnte.

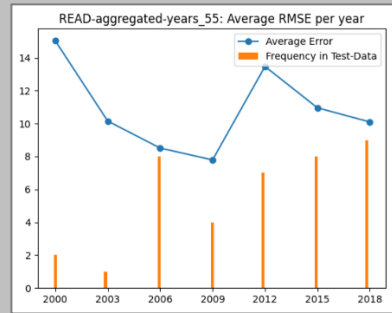


Abb. 8: Durchschnittlicher Fehler nach Jahren der Testdaten in Relation zu deren Auftreten.

5.1.3 Science Datensatz

<i>Non-NA [%]</i>	<i>Total Features</i>	<i>Features</i>	<i>RMSE</i>	<i>R²-Test</i>	<i>R²-Train</i>	<i>max_depth</i>	<i>min_leaf</i>	<i>min_split</i>
50	39072	46	16,10	87,19%	99,53%	9	3	3
55	33811	38	16,94	85,82%	99,19%	9	3	7
60	28042	20	16,91	85,88%	97,94%	5	3	3
65	22038	52	16,21	87,02%	99,53%	9	2	5
70	19334	36	19,84	80,56%	99,28%	7	3	3
75	16135	21	19,90	80,43%	97,91%	7	6	3
80	11547	38	17,98	84,03%	99,18%	9	3	7
85	6814	33	20,89	78,44%	98,88%	7	3	7
90	4546	23	20,91	78,41%	97,84%	7	4	9
95	2427	16	26,59	65,06%	96,26%	7	5	11
100	30	18	22,48	75,04%	93,59%	8	2	3

Tab. 5: Vergleich der Single Tree Modelle für den Datensatz Science mit Einbezug der aggregierten PISA Jahre.

Auch bei der Analyse des Science Datensatzes kann ein Zusammenhang zwischen Performance und Anteil der Non-NA Werte der Modelle gezogen werden. So kann ein Abwärtstrend der Performance mit sinkendem Non-NA Anteil beobachtet werden. Diese Beobachtung führt zur gleichen Schlussfolgerung wie im Rahmen der Analyse der Read Modelle beschrieben. Anders als bei den optimalen Modellen der beiden

anderen Datensätze bedient sich das optimale Modell dieses Datensatzes einer deutlich höheren Anzahl an Features. Dies weist darauf hin, dass es für die Vorhersage der Science Scores keine besonders dominanten Features gibt, welche einen besonders hohen Einfluss auf die Vorhersage hätten.

Die Abbildung 9 gibt einen Eindruck über die Performance des optimalen Modells auf dem Datensatz Science. Obwohl die meisten Werte gut prognostiziert werden konnten, sticht die Vorhersage für das PISA Ergebnis von Singapur im Jahr 2015 als Ausreißer heraus. Auffällig ist, dass diese vergleichenden Abbildungen des Datensatzes Read ebenfalls einen Ausreißer im Jahr 2015 darstellten.

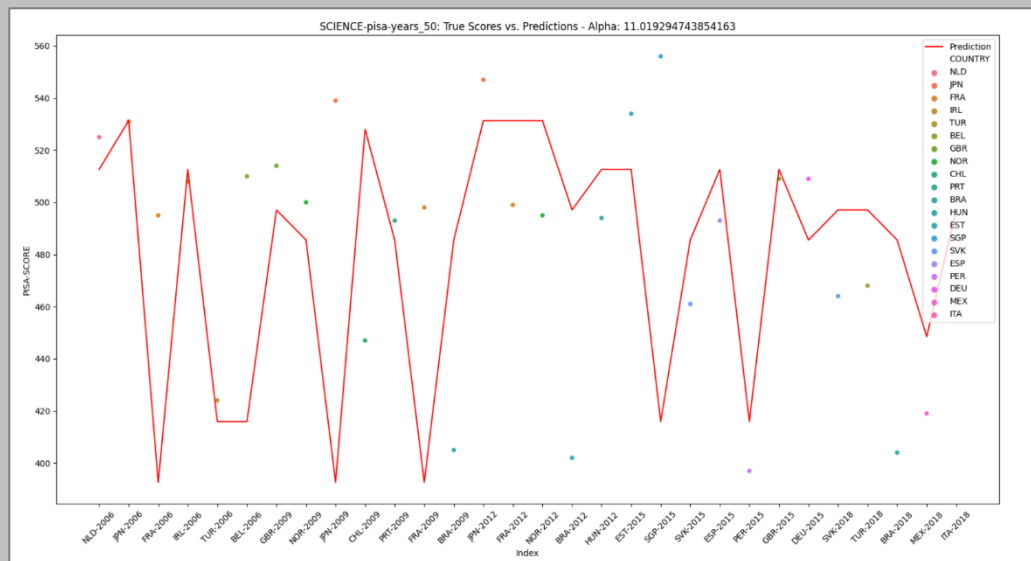


Abb. 9: Vorhersage des optimalen Modells SCIENCE _50.

Auch Abbildung 10 verstärkt den Eindruck, dass die im Jahr 2015 (wie auch 2018) erhobenen Daten eine hohe Varianz im Vergleich zu den in den vorherigen Jahren erhobenen Daten aufwiesen und aus diesem Grund zu Fehlern im Modell führten.

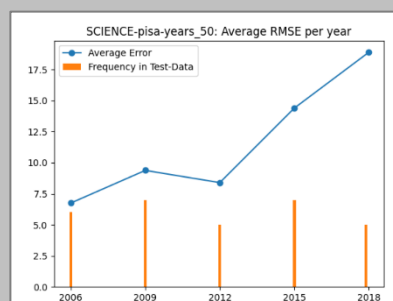


Abb. 10: Durchschnittlicher Fehler nach Jahren der Testdaten in Relation zu deren Auftreten.

5.2 Pruning

In Tabelle 6 sind die Auswirkungen des Prunings auf den optimierten Single Trees des Math Datensatzes zu sehen. Performanceverbesserungen bezüglich des R^2 bzw. RMSE sind nur bei einem Alpha von 0,153 zu beobachten und betreffen nur die 3. Nachkommastelle (Promillebereich). Generell kann beobachtet werden, dass die Performance der Modelle wie erwartet mit steigendem Alpha abnimmt (Abbildung 11). Auffällig ist zudem, dass der Baum bei maximalem Pruning mit Hilfe nur noch eines Features immer noch eine Performance von R^2 gleich 74,55% erreicht. Diese Beobachtung kann auch für die beiden anderen Datensätze gemacht werden. Für den Datensatz Read, wie aus Anhang 2 ersichtlich, verstärkt Pruning die Performance überhaupt nicht. Für Science hingegen, wie aus Anhang 3 entnommen werden kann, verbessert sich das R^2 im Promillebereich mit einem Alpha im Intervall [0,034; 0,080] und [0,934; 2,707] sowie [12,140]. Eine Verbesserung der Metrik auf die zweite Nachkommastelle kann durch ein Alpha im Intervall [3,343] und [9,672; 11,019] erreicht werden. Zusammenfassend kann festgestellt werden, dass Pruning die Performance der Modelle nur geringfügig verbessert.

<i>Alpha</i>	<i>Features Tree</i>	<i>RMSE</i>	<i>R²</i>	<i>Nodes Tree</i>
0	26	11,32	93,928%	53
0,15	25	11,29	93,968%	51
0,24	24	11,33	93,919%	49
...
2,41	18	11,94	93,254%	37
2,44	17	12,06	93,108%	35
...
6,47	10	12,27	92,875%	21
8,86	9	12,14	93,020%	19.0
...
173,69	1	23,19	74,552	3.0
1.140,6	0	46,82	-0,372%	1.0

Tab. 6: Vergleich der Performance der verschiedenen Alphas für das Pruning des optimalen Modells des Datensatzes Math.

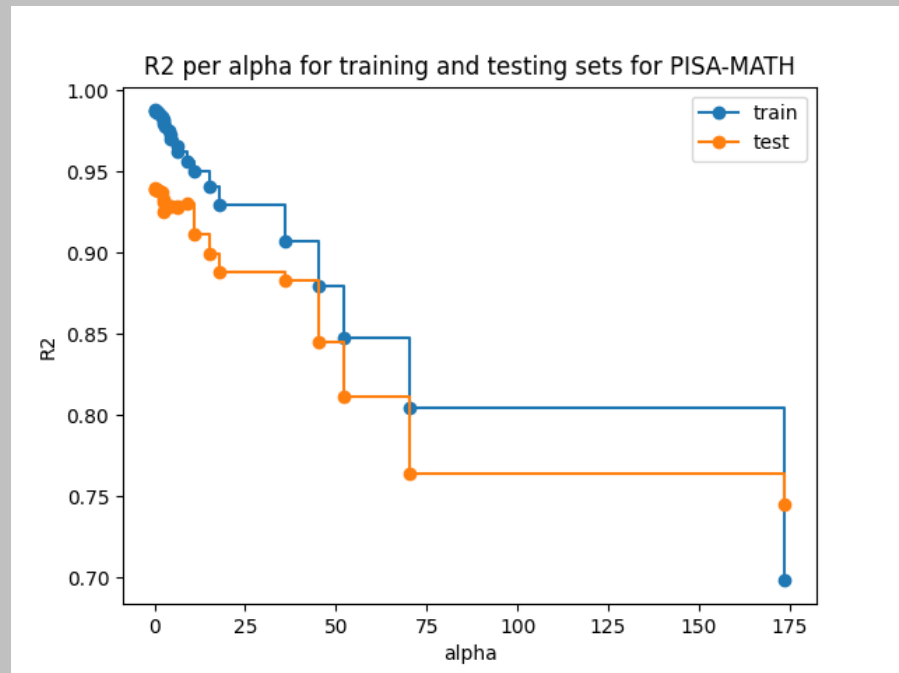


Abb. 11: Abnahme des R^2 je Alpha bei Pisa-Math

5.3 Random Forest

Der Random Forest (im Weiteren RF) wurde für jeden Datensatz mit verschiedenen Parametern gebaut. Die Ergebnisse sind in der untenstehenden Tabelle 7 gelistet und wurden mit dem optimalen Modell aus dem Single Tree Durchlauf mit Pruning erstellt. Der Random Forest wurde jeweils a) ohne Restriktionen bezüglich der Parameter, b) mittels der Parameter des optimalen Single Trees ausgenommen Pruning und c) auf Basis des geprunten optimalen Baumes gebaut. Auffällig ist, dass der RF mit zunehmender Ähnlichkeit mit den optimalen Single Tree Modellen an Performance verliert. Außer für den Read Datensatz haben die RF Modelle, welche ohne Restriktionen gebaut wurden, die beste Performance. Zusammenfassend ist jedoch zu beobachten, dass eine Erweiterung der Modelle durch den RF Ensemble Ansatz nicht lohnend scheint. Dies könnte damit zusammenhängen, dass es mit Blick auf die Performances der besten Single Trees der Datensätze kaum Verbesserungspotential für den RF Ansatz gibt.

<i>Model</i>	<i>Alpha</i>	<i>RMSE</i>	<i>R²_Test</i>	<i>R²_Train</i>	<i>max_depth</i>	<i>min_leaf</i>	<i>min_split</i>
MATH_65	0,153	11,29	93,97%	99%	8	5	3
MATH_RF_no_restrictions	0,153	13,50	91,38%	98,74	/ ⁶	1	2
MATH_RF_opt	0,153	13,71	91,10%	97,76%	8	5	3
MATH_RF_Pruned	0,153	13,72	91,10%	98%	8	5	3
READ_agg_55	0	13,36	86,83%	93,11%	4	4	9
READ_RF_no_restrictions	0	14,41	84,68%	98,06	/	1	2
READ_RF_Pruned	0	15,02	83,33%	94,75%	4	4	9
SCIENCE_50	11,019	14,62	89,45%	94,65%	9	3	3
SCIENCE_RF_no_restrictions	11,019	17,21	85,36	98,52	/	1	2
SCIENCE_RF_opt	11,019	16,90	85,89%	98,04%	9	3	3
SCIENCE_RF_Pruned	11,019	17,56	84,76%	95,83%	9	3	3

Tab. 7: Vergleich der Performance der verschiedenen Random Forest Modelle für die drei Datensätze

⁶ Für den Random Forest ohne Beschränkungen wurde kein spezifischer max-depth definiert.

6 FEATUREANALYSE

Allgemein kann beobachtet werden, dass die Vielzahl an möglichen Features für die Vorhersage eine Austauschbarkeit der vom Baum gewählten Vorhersageindikatoren zur Folge hat. Dies kann am Beispiel in Abbildung 11 und 12 gesehen werden. Beide Abbildungen zeigen einen gleichen Baum, mit identischen Parametern und Metriken. Trotzdem werden im zweiten und dritten Split nicht dieselben Features verwendet.

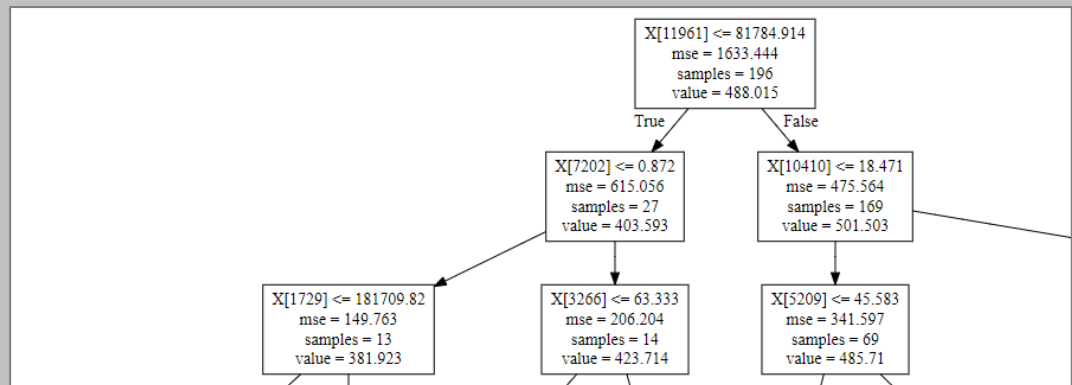


Abb. 12: Ausschnitt eines Baumes für MATH_aggregated_years_75.

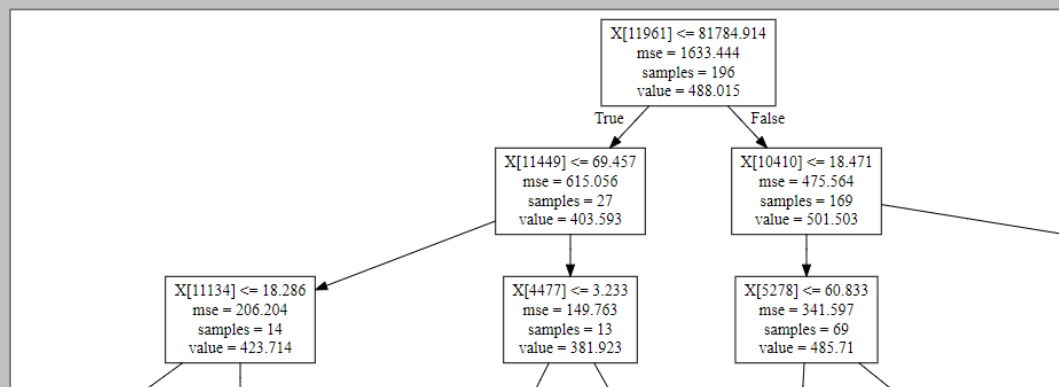


Abb. 13: Ausschnitt des Baumes für MATH_aggregated_years_75 mit Pruning bei einem Alpha=0.

Eine weitere Beobachtung ergibt sich aus der Korrelation der Features mit den Ländern. Um diesen auszudrücken wurde der Variationskoeffizient genutzt. Dieser setzt die Standardabweichung in Verhältnis zum Mittelwert. Dabei fällt auf, dass ein Großteil der genutzten Features nicht unbedingt mit den PISA Scores an sich sondern vielmehr mit den Ländern korrelieren (siehe Abb. 14: am Beispiel des Datensatzes Math). Dies wird dadurch bedingt, dass die Länder immer ähnliche PISA Scores erreichen (vgl. Kapitel 4.1).

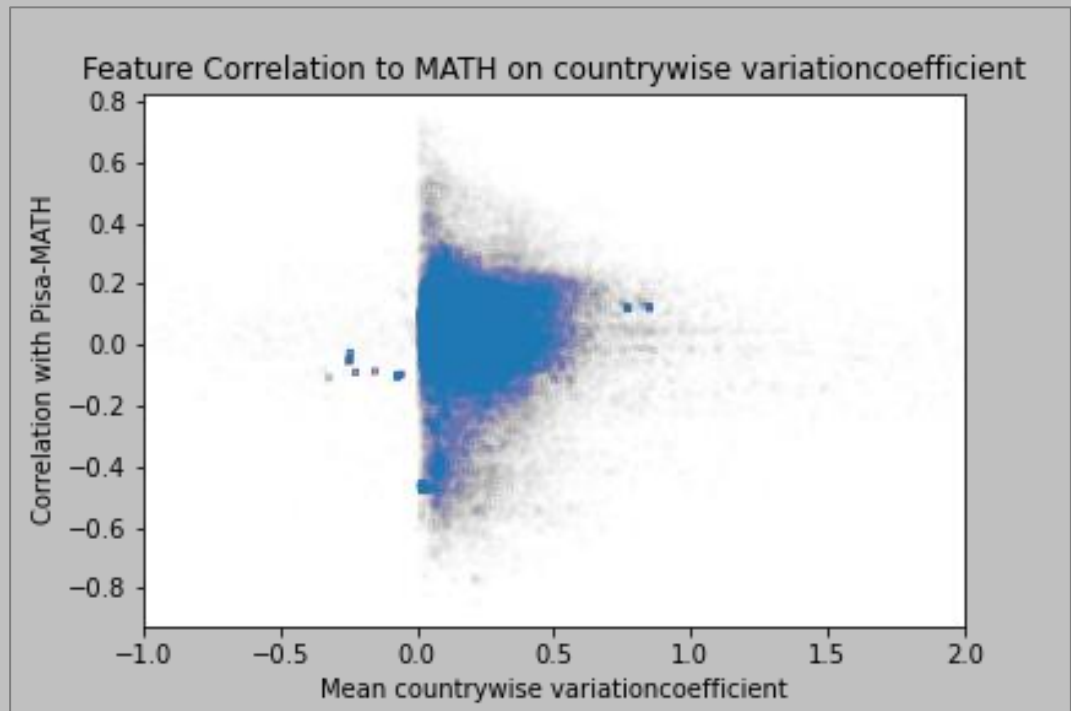


Abb. 14: Zusammenhang zwischen der Varianz innerhalb eines Landes und der Korrelation des Features mit den Pisa-Daten am Beispiel des Datensatzes Math.

Die folgenden Tabellen zeigen die Top 3 Features für jeden optimierten Entscheidungsbaum nach dem Kriterium „kumulierte Beobachtungen pro Feature“. Die zugrundeliegende Überlegung ist, dass der Entscheidungsbaum im Rootnode die meisten Beobachtungen bündelt und die darauffolgenden Childnodes immer weniger Beobachtungen enthalten. Da der Entscheidungsbaum immer sukzessive das beste Feature pro Teildatensatz auswählt, ist anzunehmen, dass die Beobachtungen pro Node (ein Feature pro Node) ein Kriterium für die Wichtigkeit des Features im Entscheidungsbaum sind. Wurde ein Feature mehr als einmal in einem Entscheidungsbaum verwendet, wurden die durch dieses Feature geteilten Beobachtungen addiert.

6.1 Math Datensatz

<i>Dataset</i>	<i>Feature</i>	<i>Samples</i>	<i>NA-Values (in Feature)</i>	<i>Annual Mean (in Feature)</i>	<i>Variation coefficient</i>
Math_65_0,153	TM.TAX.MANF.BR. ZS-Bound rate, simple mean,	196	3,46%	0,43%	2,1%

	manufactured products (%)				
Math_65_0,153	SI.DST.FRST.20- Income share held by lowest 20%	170	34,63%	12,55%	3,3%
Math_65_0,153	GREEN_GROWTH - : Population density, inhabitants per km2 [Units Inhabitants]	129	1,29%	1,29%	4,1%

Tab. 8: Vergleich der Top 3 Features des optimalen Modells für den Datensatz Math.

Das Top 1 Feature für das optimale Math Modell lautet „*TM.TAX.MANF.BR.ZS-Bound rate, simple mean, manufactured products (%)*“ und beschreibt den mittleren, maximalen Zollsatz eines Landes auf Produkte anderer Länder. Das Feature kann auf Grund seines niedrigen Anteils an NA-Werten und Annual Mean-Schätzer als robust beschrieben werden. Der Split erfolgt bei einem Prozentsatz von 17,12. Liegt der Zollsatz unter diesem Wert, so kann allgemein ein höherer Math PISA Score [449, 546] als beim Darüberliegen [370, 433] erreicht werden. Das Feature mit der zweithöchsten Anzahl an Samples ist „*SI.DST.FRST.20-Income share held by lowest 20%*“ und beschreibt den prozentualen Anteil am Gesamteinkommen eines Landes der 20%, die am geringsten verdienen. Mit einem NA Wert Anteil von 34,63% und einem Annual Mean Schätzer Anteil von 12,55 innerhalb der Feature Daten kann dieses Feature, trotz der hohen Wichtigkeit für die Vorhersage der Math PISA Scores die niedrigste Robustheit der gesamten Top 3 Features zugesprochen werden. Der Schwellwert dieses Features liegt bei 7,219%. Wie zu erwarten war, schneiden Länder, welche unter diesem Schwellwert liegen bei Math PISA schlechter ab [449, 493], als solche die darüber liegen [480, 546]. 129 Samples werden auch durch das Feature „*GREEN_GROWTH -: Population density, inhabitants per km2 [Units Inhabitants]*“ entschieden. Hierbei handelt es sich um die Bevölkerung Mitte des Jahres geteilt durch Landfläche in Quadratkilometern. Dieses Feature ist wie das erste, mit einem NA-Wert Anteil von 1,29% und einem Annual Mean Schätzer Anteil von 4,1%, ein sehr robuster Indikator für die Vorhersage. Der Schwellwert liegt bei 296,729. Für dieses Feature kann keine eindeutige Aussage bezüglich einer möglichen Richtung seines Einflusses gemacht werden. Bei diesem Feature ist das Zusammenspiel mit anderen Features tieferer Hierarchiestufen entscheidend.

Zusammenfassend sind die drei Top Features ein Indikator für Wohlstand und dessen gleichmäßige Verteilung in einem Land. Dies deutet auch darauf hin, dass die Features viel mehr als Proxys der Länder zur Vorhersage der PISA Werte dienen. Dies passt

auch zur Beobachtung in Kapitel 4.1, dass die PISA Scores eines Landes alle drei Jahre nur schwach voneinander abweichen, und zu den Ergebnissen des Variationskoeffizienten der Feature.

6.2 Read Datensatz

<i>Dataset</i>	<i>Feature</i>	<i>Samples</i>	<i>NA-Values (in Feature)</i>	<i>Annual Mean (in Feature)</i>	<i>Variation coefficient</i>
Read_agg__55	SP.ADO.TFRT- Adolescent fertility rate (births per 1000 women ages 15-19)	221	0,38%	0,38%	21%
Read_agg__55	PT6 - : Euros (millions) + 1. Domestic undertakings + Loans + Non-Life + Total + Direct insurer [Millions Euro]	198	22,69%	6,92%	80%
Read_agg__55	TIM_2019_MAIN - Domestic compensation of employees content of gross exports : Austria + Chemicals and pharmaceutical products [Millions US Dollar]	114	42,30%	3,46%	20%

Tab. 9: Vergleich der Top 3 Features des optimalen Modells für den Datensatz Read.

Mit 221 Samples splittet das Feature „*SP.ADO.TFRT-Adolescent fertility rate (births per 1000 women ages 15-19)*“ alle Samples deren Vorhersage in diesem Vorgang, durch den optimalen Baum des Read Datensatzes, getestet werden. Das Feature beschreibt die Jugendfruchtbarkeitsrate, also der Anteil an Geburten pro 1000 Frauen im Alter von 15-19 Jahren. Mit einem Anteil von 0,38% der NA-Werte und Annual Mean Schätzer, gilt es als sehr robust. Der Schwellwert dieses Features liegt bei 46,704 pro 1000 Geburten. Aus dem Baum wird klar ersichtlich, dass alle Samples, die über diesem Wert liegen ein schlechteres PISA Read Ergebnis erlangen [379, 440], während Samples, welche über diesem Wert liegen besser abschneiden [442, 536]. Annähernd genauso viele Samples werden durch das Feature „*PT6 - : Euros (millions) + 1.*“

Domestic undertakings + Loans + Non-Life + Total + Direct insurer [Millions Euro]” geteilt. Das Feature beschreibt die "Bruttobetriebskosten" der Länder. Diese sollten in der Regel die Summe der Anschaffungskosten, der Veränderung der latenten Anschaffungskosten und der Verwaltungsaufwendungen darstellen. Der NA-Wert Anteil und der Anteil an Annual Mean Schätzern lassen auf ein mittelmäßig robustes Feature schließen. Der Schwellwert liegt hier bei 189,378. Für die meisten Samples gilt, dass höhere Bruttobetriebskosten zu höheren Read PISA Scores führen. Das dritt wichtigste Feature „*TIM_2019_MAIN - Domestic compensation of employees content of gross exports : Austria + Chemicals and pharmaceutical products [Millions US Dollar]*” beschreibt den inländischen Arbeitnehmerentgelt Anteil am Bruttoexport. Dieses kryptisch wirkende Feature zeigt inwieweit die Arbeitskräfte eines Landes von deren Integration in die Weltwirtschaft abhängen. Dabei bezieht es sich ausschließlich auf Österreich als Partner und auf Chemikalien und pharmazeutische Erzeugnisse als Industrie. Der NA-Wert Anteil von 42,30% und der Annual Mean Anteil von 3,46% lässt auf eine eher schlechte Robustheit schließen, welche sich dadurch auszeichnet, dass sie vor allem auf interpolierten Werten basiert. Der Schwellwert des Features liegt bei 16. Länder, die diesen unterschreiten, können ein höheres PISA Ergebnis [442, 510] erreichen als solche die ihn überschreiten [486, 563].

Die durchschnittlich schlechte Robustheit der Top 3 Features des optimalen Baums für den Datensatz Read, gibt einen Erklärungsansatz dafür, dass das Modell bezüglich seiner Performance schlechter abschneidet als das Math Modell. Die schlechte Robustheit, sowie ein Überdenken der kausalen Zusammenhänge, insbesondere zwischen den PISA Scores und dem Top 3 Feature, lässt auf einen eher zufälligen Fit dieses letzten Features schließen. Allgemein stellen auch die Features des Read Modells Indikatoren des Wohlstandes und der sozialen Gerechtigkeit innerhalb eines Landes dar, und können trotz etwas höherer Variationskoeffizienten als Proxy für dieses gewertet werden.

6.3 Science Datensatz

<i>Dataset</i>	<i>Feature</i>	<i>Samples</i>	<i>NA-Values (in Feature)</i>	<i>Annual Mean (in Feature)</i>	<i>Variation coefficient</i>
Science_50_11, 019	TM.TAX.MANF.BR. ZS-Bound rate simple mean manufactured products (%)	167	3,55%	0,50%	21%

Science_50_11, 019	STIO_2014 - Top performers in science among 15-year olds, % total : []	146	43,65%	4,56%	8,3%
Science_50_11, 019	GOV - Local government debt as a percentage of general government debt : [Units Percentage]	81	40,60%	22,84%	24%

Tab. 10: Vergleich der Top 3 Features des optimalen Modells für den Datensatz Science.

Das Feature „*TM.TAX.MANF.BR.ZS-Bound rate simple mean manufactured products (%)*” wurde auch schon für den optimalen Baum des Math Datensatzes als entscheidendes Feature verwendet. Die Anteile der NA-Werte und der Annual Mean-Schätzer sind etwas höher als im Math Modell, da aufgrund der späteren Erhebung andere Featurewerte in die Vorhersage miteinbezogen wurden. Für das Feature wurde ein Schwellwert von 15,245% angelegt. Das Feature teilt die Samples deutlich in den Bereich mit schlechteren PISA Scores [416, 149] (Überschreitung des Schwellwertes) und besseren PISA Scores [468, 531] (Unterschreitung des Schwellwertes). Das zweitwichtigste Feature „*STIO_2014 - Top performers in science among 15-year olds, % total*” beschreibt den Anteil an Top Performern unter den 15-jährigen und kann demnach als Proxy für die PISA Scores im Bereich Science definiert werden. Umso erstaunlicher ist es, dass dieses Feature nicht als wichtigstes Feature herangezogen wurde. Dennoch kann festgestellt werden, dass die Unterschreitung des Schwellwertes eher zu einem niedrigeren PISA Ergebnis [468, 528] führt als die Überschreitung [513, 531]. Das Feature hat eine hohe Quote an NA-Werte, da es nur alle zwei Jahre erhoben wird. Das Top 3 Feature „*GOV - Local government debt as a percentage of general government debt: [Units Percentage]*” beschreibt den Anteil der kommunalen Verschuldung an der Staatsverschuldung. Die Robustheit des Features kann auf Grund der hohen Anteile an NA-Werten und Annual Mean-Schätzern nur eingeschränkt attestiert werden. Dies lässt eine höhere Zufälligkeit bei der Vorhersagebeeinflussung durch dieses Feature erwarten. Der Schwellwert dieses Features liegt bei 3,248%. Beobachtungen, die diesen Wert unterschreiten haben ein Science PISA Ergebnis von 468, während Beobachtungen, die diesen überschreiten bei einem Ergebnis im Intervall [486, 528] liegen.

Zusammenfassend ist das Science Modell aus Sicht der Featureanalyse das schlechteste Modell, da es einerseits durch das Top 2 Feature einen hohen Proxy mit den

Lösungsvariablen erfährt und andererseits durch das Top 3 Feature von einem eher zufälligen Indikator beeinflusst wird.

7 FAZIT

Das Ziel der Studie war es die verschiedenen PISA Scores (Reading, Math und Science) der einzelnen Länder und Jahre mithilfe eines maschinellen Lernen Ansatzes anhand verschiedener volkswirtschaftlicher Indikatoren vorherzusagen. Daraufhin wurden im Rahmen der Studie große Datenmengen (ca. 20GB) aus den Datenbanken der Weltbank und der OECD heruntergeladen und mittels Python eine umfangreiche Datenaufbereitung durchgeführt. Auch nach der Datenaufbereitung sind noch viele NA's enthalten gewesen, die mittels Interpolation oder Mittelwert ersetzt wurden, um die Daten für die Modellierung mit Entscheidungsbäumen nutzbar zu machen. Dafür wurden die Daten (Spalten) in Features umgewandelt, indem sie mit den PISA Scores (Reading, Math, Science) zusammengeführt wurden. Die Daten und Modelle wurden unter den Aspekten NA's pro Feature und über 3 vorherigen Jahre aggregiert oder pro exaktem Jahr der PISA Erhebung betrachtet. Für jeden PISA Score wurden die gleichen Features unter gleichen Bedingungen betrachtet. Es wurden einzelne Entscheidungsbäume optimiert anhand von den Hyperparametern `max_depth`, `min_leaf` und `min_split`. Anschließend wurden die besten Modelle mit Pruning (Verbesserungen im Promillebereich) und Random Forest (keine Verbesserungen) weiter optimiert (Kapitel 5). Die Features, die für die Vorhersage der besten Modelle am wichtigsten waren wurden identifiziert und kritisch betrachtet (Kapitel 6). Mit den verfügbaren Daten und Modellen war es möglich den PISA Score mit einer Genauigkeit von +- 11,33 (RMSE Math), 13,36 (RMSE Read), 16,10 (RMSE Science) vorherzusagen. Pruning hat nur zu Verbesserungen im Promillebereich geführt, dennoch ist dabei deutlich geworden, dass die Entscheidungsbäume mit deutlich weniger Features bereits in der Lage waren gute Vorhersagen zu treffen. Während der Studie sind zudem verschiedene allgemeine Erkenntnisse bezüglich der Modellierung mit Entscheidungsbäumen deutlich geworden. Entscheidungsbäume sind durch ihr breites Einsatzgebiet (Klassifikation und Regression) sowie ihre vielen Erweiterungen wie Pruning, Random Forest u.v.m. für die unterschiedlichsten Problemklassen anwendbar. Durch verschiedene Visualisierungsmöglichkeiten und simple Berechnung (CART) lassen sich die Vorhersagen und Entscheidungen des Modells gut nachvollziehen und erklären.

Hinsichtlich des Ansatzes der Verwendung von vielen Daten für das maschinelle Lernen auf PISA Daten sind folgende Kritikpunkte deutlich geworden. Es existieren noch relativ wenige Datenpunkte, da die PISA Studien nur alle drei Jahre stattfinden und erst seit Jahr 2000 (Reading), 2003 (Math) und 2006 (Science) erhoben werden. Außerdem sind einige Länder wie z.B. Großbritannien und Taiwan erst später dazugekommen. Diese geringe Menge an Lösungsvariablen macht eine

Generalisierung der Modelle und das Ableiten von Kausalitäten schwierig. Es kann aufgrund des Ansatzes beim maschinellen Lernen mit vielen Daten, bei dem zuerst Modelle auf möglichst vielen Daten trainiert werden, um dann anschließend die für die Vorhersage wichtigen Variablen zu identifizieren, zufällige Korrelationen geben. Dieses Vorgehen widerspricht dem üblichen Prozedere empirischer Studien, welche versuchen durch vorherige Auswahl der Variablen bereits Hypothesen aufzustellen und diese dann empirisch zu validieren. Viele Features weisen innerhalb eines Landes beinahe identische Werte auf. Da dies auch bei den Pisa-Daten der Fall ist, wird in den Entscheidungsbäumen eine Tendenz deutlich, statt den Pisa-Scores das Land vorherzusagen. In diesem Fall wird ein Baum erstellt, der für die Pisa-Werte eines Landes dieselben Werte vorhersagt. Die Features treten somit als Repräsentation des Landes auf (Länderproxies). ⁷Ein weiteres Problem stellt die Austauschbarkeit der Features innerhalb der Modelle dar (vgl. Kapitel 5.2).

Mögliche Verbesserungen oder anschließende Studien könnten sich stärker dem Aspekt Feature Robustheit widmen. Hier könnte beispielsweise ein Term analog zum Pruning für die Robustheit des jeweiligen Features in Kostenfunktion des Algorithmus zur Baumkonstruktion eingeführt wird. Bezüglich der Modelle gibt es die Möglichkeit stärker in Richtung Ensemble Methoden zu forschen mit z.B. Adaboost o.ä. sowie den Einfluss anderer Hyperparameter weiter zu untersuchen. An dieser Stelle muss allerdings erwähnt werden, dass hier aufgrund der großen Datenmenge überdurchschnittlich große Rechenkapazitäten, die den Heimcomputer übersteigen, gebraucht werden. Zudem könnte auch ein anderer Ansatz verfolgt werden, indem zuerst die Features mit evtl. menschlichem Bias ausgewählt werden, um zufällige Korrelationen möglichst auszuschließen. Auch im Bereich Feature Engineering würde dies weitere Möglichkeiten bieten, wie z.B. Nutzung von kategorischen Variablen. Allerdings sei hier gewarnt, dass dies bei einem viel-Feature Ansatz an seine Grenzen stößt und Kosten/Nutzen an dieser Stelle abgewogen werden muss.

⁷ Dieses Problem wird in der englischsprachigen Literatur oft unter dem Stichwort „redundant encoding“ betrachtet.

LITERATURVERZEICHNIS

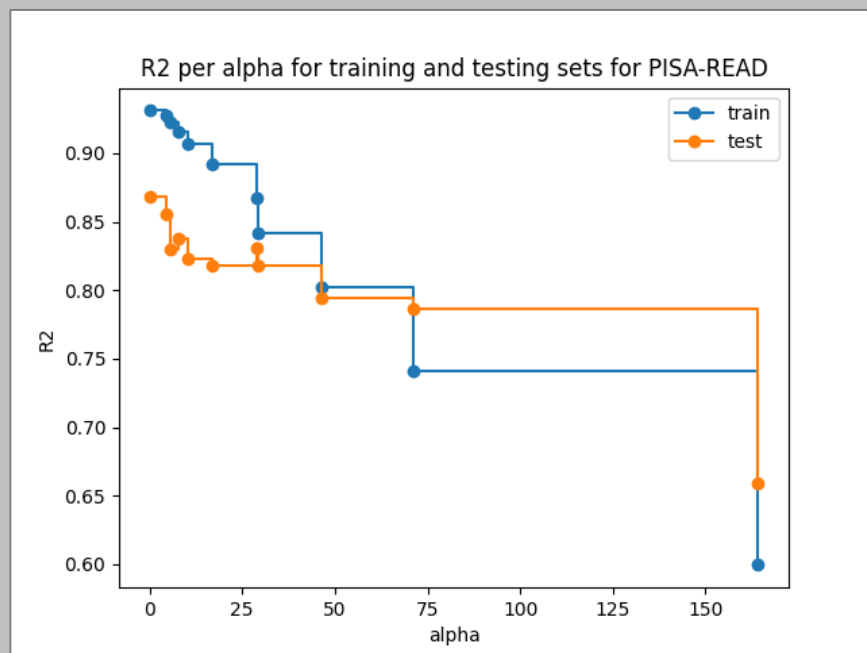
- Arantza Gorostiagaa, J. L.-Á. (11. Juli 2015). On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain. *Elsevier - Neurocomputing*.
- B. Naik, S. R. (2004). Using neural networks to predict MBA student success. *Coll. Stud. J.* 38 , 15.
- br.de. (03. 12 2019). Von br.de: <https://www.br.de/nachrichten/wissen/pisa-studie-deutsche-schueler-werden-wieder-schlechter,RjVOq9s> abgerufen
- C. Romero, S. V. (2007). Educational data mining: a survey from 1995 to 2005. *Expert Syst. Appl.* 33, 125–146.
- C. Romero, S. V. (2008). Data mining algorithms to classify students, in: Proceedings of the First International Conference on Educational Data Mining. pp. 8–17.
- C. Romero, S. V. (2013). Data mining in education, Wiley Interdiscip. . Rev.: *Data Min. Knowl. Discov.* 3 , 12–27.
- Chong Ho Yu, C. K.-P. (10 2012). A Data Mining Approach to Comparing American and Canadian Grade 10 Students' PISA Science Test Performance. *Journal of Data Science*, S. 441-464.
- E. Hanushek, L. W. (kein Datum). *Handbook of the Economics of Education vol.3*. NY, USA: North Holland.
- E. Hanushek, S. L. (2013). Does school autonomy make sense everywhere? panel estimates from PISA. *J. Dev. Econ.* 104 , 212–232.
- Gareth James, D. W. (2017). *An Introduction to Statistical Learning*. London: Springer Verlag.
- Géron, A. (2019). Hands-On Machine Learning Scikit-Learn & TensorFlow. In A. Géron, *Hands-On Machine Learning Scikit-Learn & TensorFlow*. O'Reilly Media, Inc.
- Hastie, T. T. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer Science & Business media.
- Herberholz, L. (2014). Effekte der Drittmittelfinanzierung auf den Output von Hochschulen. *Mimeo*, 1(1), S. 1-40.
- J.P. Vandamme, N. M. (2007). Predicting academic performance by data mining methods. *Educ. Econ.* 15 , 405–419.
- Luigi Guiso, F. M. (2008). Culture, Gender, and Math. *ScienceMag*, 1164-1165.

- Michler, I. (06. 12 2016). *welt.de*. Von <https://www.welt.de/wirtschaft/article160016765/Deutsche-Schueler-finden-Naturwissenschaften-laestig.html#:~:text=Pisa%20ist%20die%20umfangreichste%20international%20durchgef%C3%BChrte%20Bildungsstudie.%20Ziel,Naturwissenschaften%20im%20Fokus.%20Die%20Erg> abgerufen
- OECD. (2000). *web.archiv.org*. Von *web.archiv.org*: https://web.archive.org/web/20090715211650/http://www.pisa.oecd.org/document/7/0,2340,en_32252351_32236159_33688711_1_1_1_1,00.html abgerufen
- S. Jenkins, J. M. (2008). Social segregation in secondary schools: how does England compare with other countries? *Oxford Rev. Educ.* 34 , 21–38.
- S. Schnepf. (2007). Immigrants' educational disadvantage: an examination across ten countries and three surveys. *J. Popul. Econ.*, 527–545.
- Sprietsma, M. (2010). The effect of relative age in the first grade of primary school on long-term scholastic results: international comparative evidence using PISA 2003. *Educ. Econ.* 18, 133–156.
- T. Fuchs, W. (2007). What accounts for international differences in student performance? a reexamination using PISA data. *Econ.*, 433–464.
- Vincent Vandenberghe, S. R. (2016). Evaluating the Effectiveness of Private Education Across Countries: A Comparison of Methods. *Labour Economics*, 487–506.
- welt.de*. (29. 09 2020). Von *welt.de*: <https://www.welt.de/politik/deutschland/article216803170/Pisa-Studie-Das-grosse-Problem-der-deutschen-Lehrer.html> abgerufen

werden. Die Features bezogen sich insbesondere auf die Themen Armut und Ungleichheit, Menschen, Umwelt, Wirtschaft, Staaten und Märkte und globale Verknüpfungen. Um den Datensatz in eine für die Modelle funktionale Form zu bringen, wurden primär Funktionen der Pandas Library genutzt, die im Code „prepare_WB.py“ genauer beschrieben werden.

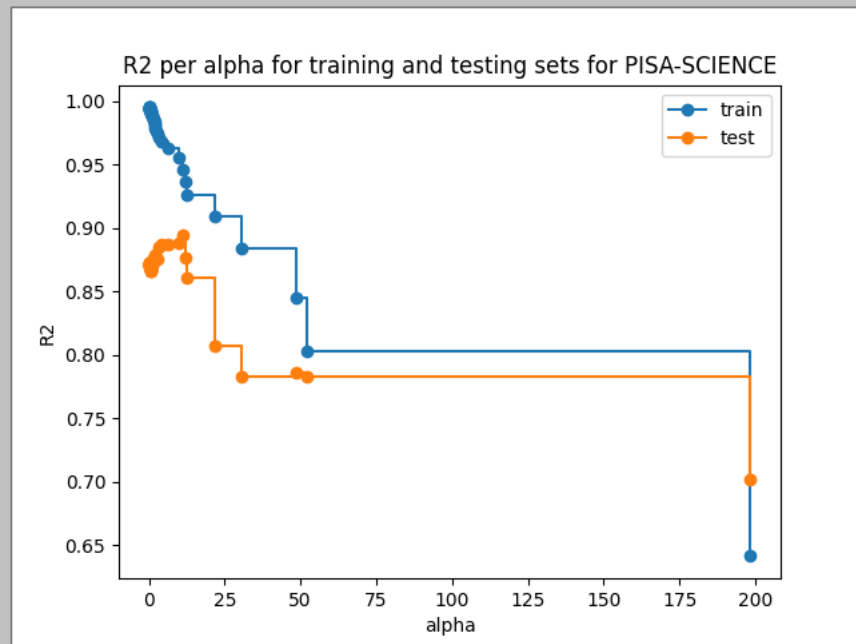
A. 2: Ergebnisse des Prunings für READ_55_aggregated_years

	Percentage	n	Total Feature	Features Tre	RMSE	R2	Percentage	_Alpha	max_Depth	min_leafs	min_splits	Nodes_Tree
0	55	19414	11	13.3550592	0.8682944	9.93006993	0	4	4	9	23	
4.42793971	55	19414	10	13.9983273	0.85530122	8.84615385	4.42793971	4	4	9	21	
5.57767722	55	19414	9	15.1711331	0.83003925	9.7008547	5.57767722	4	4	9	19	
7.87815126	55	19414	8	14.81646	0.8378931	10.4807692	7.87815126	4	4	9	17	
10.2738975	55	19414	7	15.4739738	0.82318614	10.3846154	10.2738975	4	4	9	15	
17.0178588	55	19414	6	15.6935374	0.81813284	7.75641026	17.0178588	4	4	9	13	
29.0382618	55	19414	5	15.15389	0.83042538	8.92307692	29.0382618	4	4	9	11	
29.3555747	55	19414	4	15.7147788	0.81764019	3.84615385	29.3555747	3	4	9	9	
46.4871681	55	19414	3	16.6847765	0.79443301	3.58974359	46.4871681	3	4	9	7	
71.1260341	55	19414	2	17.0150525	0.78621404	3.65384615	71.1260341	2	4	9	5	
164.195633	55	19414	1	21.4791708	0.65931923	0.38461539	164.195633	1	4	9	3	
696.723709	55	19414	0	37.2071653	-0.0222703		696.723709	0	4	9	1	



A. 3: Ergebnisse des Prunings für SCIENCE_50

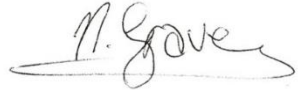
	Percentage	Total Featur	Features Tre	RMSE	R2	Percentage	Alpha	max_Depth	min_leafs	min_splits
0	50	39072	46	16.1009831	0.87190605	17.0256551	0	9	3	3
0.03450242	50	39072	45	16.0854903	0.87215244	17.0473773	0.03450242	9	3	3
0.06387226	50	39072	44	16.0652154	0.87247453	17.0848441	0.06387226	9	3	3
0.06387226	50	39072	43	16.0652154	0.87247453	17.5276202	0.06387226	9	3	3
0.0641574	50	39072	42	16.0652154	0.87247453	17.5276202	0.0641574	9	3	3
0.08083832	50	39072	41	16.0652154	0.87247453	17.366559	0.08083832	9	3	3
0.08083832	50	39072	40	16.0652154	0.87247453	17.5444162	0.08083832	9	3	3
0.08732535	50	39072	39	16.1392146	0.87129702	17.5444162	0.08732535	9	3	3
0.09238666	50	39072	38	16.1541921	0.87105803	17.0296381	0.09238666	9	3	3
0.12075848	50	39072	37	16.1541921	0.87105803	17.4619289	0.12075848	9	3	3
0.13575706	50	39072	36	16.1387556	0.87130434	17.1538596	0.13575706	9	3	3
0.14371257	50	39072	35	16.1387556	0.87130434	17.5126904	0.14371257	9	3	3
0.18567864	50	39072	34	16.1744791	0.87073396	17.5220906	0.18567864	9	3	3
0.2352438	50	39072	33	16.1744791	0.87073396	17.1222179	0.2352438	8	3	3
0.28842315	50	39072	32	16.1744791	0.87073396	17.1222179	0.28842315	8	3	3
0.29198745	50	39072	31	16.1744791	0.87073396	17.1222179	0.29198745	8	3	3
0.29585828	50	39072	30	16.2428561	0.86963872	16.893401	0.29585828	8	3	3
0.5102263	50	39072	29	16.3860266	0.86733049	16.893401	0.5102263	8	3	3
0.51097804	50	39072	28	16.3679343	0.86762329	16.5397631	0.51097804	8	3	3
0.63453094	50	39072	27	16.3997543	0.8671081	16.5747076	0.63453094	8	3	3
0.64959367	50	39072	26	16.4693863	0.86597721	17.0050761	0.64959367	8	3	3
0.79512404	50	39072	25	16.4693863	0.86597721	15.6876964	0.79512404	7	3	3
0.80479042	50	39072	24	16.2046295	0.87025159	16.2944162	0.80479042	7	3	3
0.89600798	50	39072	23	16.2219244	0.86997449	16.4306706	0.89600798	7	3	3
0.9344644	50	39072	22	15.948019	0.87432835	16.6948675	0.9344644	7	3	3
1.16407186	50	39072	21	15.948019	0.87432835	15.8256196	1.16407186	7	3	3
1.36626747	50	39072	20	15.9275982	0.87464998	16.180203	1.36626747	7	3	3
1.73053892	50	39072	19	15.9483383	0.87432332	14.3485618	1.73053892	7	3	3
1.80218135	50	39072	18	15.8717143	0.87552805	14.2494561	1.80218135	7	3	3
2.02095808	50	39072	17	15.8717143	0.87552805	14.4474815	2.02095808	7	3	3
2.05838323	50	39072	16	15.6275785	0.87932781	14.9746193	2.05838323	7	3	3
2.12303099	50	39072	15	15.9179447	0.87480188	14.9746193	2.12303099	7	3	3
2.38988986	50	39072	14	15.818259	0.87636507	14.028611	2.38988986	6	3	3
2.70786847	50	39072	13	15.8933174	0.87518898	14.028611	2.70786847	6	3	3
3.34318287	50	39072	12	15.2684152	0.8848108	11.9796954	3.34318287	6	3	3
4.25602504	50	39072	11	15.1061534	0.88724609	11.9796954	4.25602504	5	3	3
6.46706587	50	39072	10	15.1061534	0.88724609	8.74224478	6.46706587	5	3	3
9.67218763	50	39072	9	15.0226476	0.88848924	9.39086294	9.67218763	5	3	3
11.0192947	50	39072	8	14.6157964	0.89444743	9.39086294	11.0192947	5	3	3
12.1401839	50	39072	7	15.7962053	0.87670957	10.3698332	12.1401839	5	3	3
12.6759124	50	39072	6	16.8000927	0.8605408	10.1522843	12.6759124	4	3	3
21.5506965	50	39072	5	19.7620704	0.80703048	11.6751269	21.5506965	3	3	3
30.5206441	50	39072	4	20.9364154	0.78341495	9.89847716	30.5206441	3	3	3
48.5607357	50	39072	3	20.8290741	0.78563012	9.30626057	48.5607357	3	3	3
52.0012479	50	39072	2	20.938876	0.78336404	2.53807107	52.0012479	2	3	3
198.451269	50	39072	1	24.5694197	0.70172722	0.50761421	198.451269	1	3	3
789.994697	50	39072	0	45.6822338	-0.031142		789.994697	0	3	3



EIDESSTATTLICHE ERKLÄRUNG


Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Aachen, den 14.02.2021




Nina Graves

Karlsruhe, den 14.02.2021



Marie Rahlmeyer

Karlsruhe, den 14.02.2021



Jendrik von Wardenburg