Wissenschaftliche Dokumentation

Thema:

Vorhersage von Schülerleistungen

vorgelegt von:

Selin Yildirim, 814319

im Studiengang Digital Business Engineering

an der Hochschule Reutlingen

Studienfach: Artificial Intelligence

Wintersemester 2024/25

Prüfer: Prof. Alexander Rossmann

Am Abgabedatum: 07.01.2025

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Unternehmensbeschreibung

Die Fallstudie befasst sich mit dem Unternehmen EduPro Group als führender Anbieter von Bildungsanalytik. EduPro Group ist auf die Anwendung datengetriebener Ansätze spezialisiert, um die Leistungen von Schülern zu verbessern und Bildungseinrichtungen bei der Entscheidungsfindung zu unterstützen. Das Unternehmen nutzt moderne Methoden wie Machine Learning, Datenvisualisierung und prädiktive Modellierung, um wertvolle Erkenntnisse aus Schülerleistungsdaten zu gewinnen und diese in praktische Empfehlungen umzuwandeln. EduPro Group zielt darauf ab, Bildung gerechter und effizienter zu gestalten. Die Schüler sollen mit den notwendigen Werkzeugen ausgestattet werden, um individuelle Fördermaßnahmen zu entwickeln, die den individuellen Bedürfnissen der Schüler gerecht werden. Der Fokus liegt auf der optimalen Nutzung vorhandener Ressourcen und der langfristigen Steigerung der Bildungsqualität.

Das Dienstleistungsangebot von EduPro Group umfasst mehrere Bereiche. Die Analyse von Schülerleistungsdaten beinhaltet Faktoren wie Lernzeit, Anwesenheit und familiäre Unterstützung. Prädiktive Analysen ermöglichen die frühzeitige Erkennung von Herausforderungen und die Einleitung entsprechender Maßnahmen. Zusätzlich dazu entwickelt das Unternehmen benutzerfreundliche Dashboards, die Lehrkräften und Schuladministratoren einen schnellen Überblick über relevante Daten bieten. Durch die Kombination dieser Ansätze werden Lehr- und Lernstrategien nachhaltig optimiert. EduPro basiert auf umfangreichen Forschungen im Bereich Educational Data Mining (EDM) und Learning Analytics.[1] Studien belegen, dass datenbasierte Ansätze die Qualität von Entscheidungen in Bildungseinrichtungen signifikant verbessern können.[2] Insbesondere die Anwendung prädiktiver Modelle ermöglicht es, die Leistungsentwicklung von Schülern besser zu verstehen und Maßnahmen darauf abzustimmen.[3]

Zusammenfassend zeigt die Fallstudie, wie EduPro als Pionier im Bereich der Bildungsanalytik Schulen dabei unterstützen kann, datenbasierte Ansätze in ihre Praxis zu integrieren. Die Ergebnisse der Analyse zeigen, dass die Kombination aus moderner Technologie und fundierter Datenauswertung die Bildungsqualität nachhaltig verbessern kann.

[1] eduPro Group. (2024).Bildung kennt keine Grenzen.

[2] Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, 46(5), 30–40.

[3] Baker, R. S. J. d., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 3–17.

2 Problem- und Fragestellung

Im Rahmen dieser Fallstudie untersucht die EduPro Group, wie Schülerleistungsdaten systematisch analysiert werden können, um relevante Muster und Einflussfaktoren zu identifizieren.

Eine zentrale Problematik besteht darin, dass Schulen oft auf begrenzte Ressourcen zurückgreifen müssen und somit nicht immer die besten Entscheidungen für individuelle Fördermaßnahmen treffen können. Durch den Einsatz moderner Datenanalysen und maschinellen Lernens soll jedoch eine präzisere Bewertung der Schülerleistungen ermöglicht werden. Insbesondere Machine-Learning-Modelle bieten die Möglichkeit, große Datenmengen zu verarbeiten und Muster zu erkennen, die durch klassische Analyseansätze schwer zu identifizieren sind.

In diesem Kontext stellt sich die Frage, wie EduPro seine Technologien effektiv nutzen kann, um die Haupttreiber von Schülerleistungen zu identifizieren. Dabei ergeben sich die folgenden spezifischen Fragestellungen:

1. Welche Faktoren stehen in einem signifikanten Zusammenhang mit den Klausurpunktzahlen der Schüler, und welche Merkmale haben den größten Einfluss auf ihre Leistungen?
2. Wie genau können maschinelle Lernmodelle wie Random Forest oder Lineare Regression die Klausurpunktzahlen vorhersagen, und welche Modelle liefern die zuverlässigsten Ergebnisse?
3. Welche praktischen Empfehlungen können aus den Analyseergebnissen abgeleitet werden, um die Förderung der Schüler effizienter zu gestalten und die verfügbaren Ressourcen optimal einzusetzen?

3 Einordnung in den Bereich Machine Learning

Die vorliegende Fallstudie lässt sich eindeutig im Bereich des maschinellen Lernens (ML) verorten. ML umfasst eine Vielzahl von Techniken und Algorithmen, die es ermöglichen, Muster und Zusammenhänge in großen und komplexen Datensätzen zu erkennen. Im Bildungsbereich bietet ML das Potenzial, entscheidende Einblicke in Schülerleistungsdaten zu gewinnen und diese für die Entwicklung gezielter Fördermaßnahmen zu nutzen. Maschinelles Lernen wird in dieser Fallstudie eingesetzt, um zwei zentrale Ziele zu erreichen: Erstens die Identifikation der wichtigsten Einflussfaktoren auf Schülerleistungen und zweitens die Entwicklung eines prädiktiven Modells zur Vorhersage der Klausurpunktzahlen. Diese Vorgehensweise basiert auf der Analyse von historischen Daten, die Variablen wie Lernzeit, Anwesenheit und familiäre Unterstützung enthalten.

Durch die Verwendung geeigneter Modelle wie Random Forest und Linearer Regression lassen sich komplexe nichtlineare Beziehungen zwischen diesen Variablen und den Schülerleistungen modellieren. Random Forest, ein Ensemble-Algorithmus, ist besonders geeignet, da er robuste Vorhersagen ermöglicht und gleichzeitig die Bedeutung einzelner Merkmale quantifiziert.[4] Dies macht ihn zu einem wertvollen Werkzeug für die Untersuchung, welche Faktoren den größten Einfluss auf die Schülerleistungen haben. Die Lineare Regression wird ergänzend verwendet, um einfache lineare Zusammenhänge zu analysieren und einen direkten Vergleich der Modellleistungen zu ermöglichen. Der Einsatz von Machine Learning in der vorliegenden Fallstudie steht im Kontext der stetig wachsenden Anwendung von Educational Data Mining (EDM) und Learning Analytics. Studien haben gezeigt, dass maschinelles Lernen nicht nur die Genauigkeit von Vorhersagen im Bildungsbereich verbessern kann, sondern auch dabei hilft, Entscheidungsprozesse datengetriebener zu gestalten.[5]

Die Analyse komplexer Wechselwirkungen zwischen Variablen wie Lernzeit und Prüfungsleistung ist dabei ein entscheidender Schritt, um fundierte Handlungsempfehlungen für Bildungseinrichtungen zu entwickeln. Durch den systematischen Einsatz von Machine-Learning-Methoden trägt diese Fallstudie zur Weiterentwicklung von datenbasierten Ansätzen im Bildungswesen bei. Die Erkenntnisse können nicht nur dazu genutzt werden, individuelle Schülerleistungen zu verbessern, sondern auch, um Schulen dabei zu unterstützen, Ressourcen effektiver einzusetzen und langfristig eine nachhaltige Steigerung der Bildungsqualität zu erzielen.

[4] IBM (2024).Was ist Random Forest?.

[5] Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, 46(5), 30–40.

4 Datenerfassung und Datenbeschreibung

Der Datensatz basiert auf anonymisierten Schülerdaten, die systematisch erhoben wurden, um die relevanten Faktoren für Schülerleistungen zu untersuchen. Die Erhebung erfolgt durch standardisierte Fragebögen und schulische Aufzeichnungen. Zusammenfassend enthält der verwendete Datensatz 15 Variablen, die sowohl numerische als auch kategoriale Werte umfassen. Die numerischen Variablen, wie etwa Exam_Score, Hours_Studied und Attendance, zeigen kontinuierliche Daten mit individuellen Skalen, die im späteren Trainingsprozess berücksichtigt werden müssen, insbesondere im Hinblick auf die Normalisierung und Skalierung der Werte. Dabei wurden Variablen berücksichtigt, die individuelle, soziale und schulische Merkmale abdecken, wie z. B.:

- **Individuelle Merkmale:** Diese Variablen beschreiben individuelle Eigenschaften der Schüler, einschließlich Hours_Studied, Sleep_Hours und Health Status. Sie dienen der Erfassung persönlicher Faktoren, die die Prüfungsleistung beeinflussen können.
- **Soziale Merkmale:** Hierzu zählen Daten, die das soziale Umfeld der Schüler reflektieren, wie Parental_Education_Level und Family_Income. Diese Merkmale bieten Einblicke in die sozialen und wirtschaftlichen Bedingungen, die den Lernerfolg beeinflussen können.
- **Schulische Merkmale:** Diese Kategorie umfasst Variablen, die direkt mit dem schulischen Kontext verbunden sind, wie Attendance, Teacher_Quality und Extracurricular_Activities. Diese Daten liefern Hinweise darauf, wie schulische Unterstützung und Engagement mit den Klausurpunktzahlen zusammenhängen.

Der Datensatz beinhaltet außerdem kategoriale Werte, wie etwa Parental_Education_Level und Teacher_Quality, die nominale oder ordinal strukturierte Kategorien darstellen. Nominale Variablen besitzen keine Rangordnung, wie beispielsweise die Teilnahme an Extracurricular_Activities (Ja/Nein), während andere Variablen, wie Teacher_Quality (niedrig, mittel, hoch), eine hierarchische Struktur aufweisen.

Die Datenerhebung wurde so gestaltet, dass alle relevanten Variablen berücksichtigt werden, die potenziell Einfluss auf die Klausurpunktzahlen haben könnten. Der Datensatz umfasst insgesamt 6.607 Beobachtungen und 20 Variablen, die in numerische und kategoriale Merkmale unterteilt sind.

5 Modellbildung, Datenanalyse und Algorithmen

In diesem Kapitel wird das methodische Vorgehen zur Modellbildung für die Vorhersage von Schülerleistungen beschrieben. Im Mittelpunkt stehen die Vorbereitung der Daten, die Auswahl geeigneter Algorithmen und die Evaluierung der Modelle. Ziel ist es, die Merkmale mit dem größten Einfluss auf die Klausurpunktzahlen zu identifizieren und ein präzises Vorhersagemodell zu entwickeln.

5.1 Modellbildung

Die Modellbildung erfordert eine gründliche Vorbereitung der Daten, um sie für maschinelle Lernverfahren nutzbar zu machen. Um die Daten für die maschinelle Lernmodellierung nutzbar zu machen, wurden zunächst fehlende Werte in den kategorischen Variablen wie Teacher_Quality und Parental_Education_Level durch den Modus (häufigster Wert) ersetzt. Dieser Ansatz stellt sicher, dass keine Verzerrungen entstehen und die Daten vollständig für die Analyse vorliegen.

```
data = data.copy()
data['Teacher_Quality'] = data['Teacher_Quality'].fillna(data['Teacher_Quality'].mode()[0])
data['Parental_Education_Level'] = data['Parental_Education_Level'].fillna(data['Parental_Education_Level'].mode()[0])
data['Distance_from_Home'] = data['Distance_from_Home'].fillna(data['Distance_from_Home'].mode()[0])
```

Abbildung 1: Behandlung fehlender Werte in kategorialen Variablen

Anschließend wurden kategoriale Variablen mit One-Hot-Encoding in numerische Formate umgewandelt. Dieser Schritt ist essenziell, da maschinelle Lernmodelle numerische Eingaben erwarten. Das Argument drop_first=True wurde verwendet, um sicherzustellen, dass redundante Informationen zwischen den erzeugten Dummy-Variablen vermieden werden.

```
data_encoded = pd.get_dummies(data, drop_first=True)
```

Abbildung 2: Kodierung kategorialer Variablen mittels One-Hot-Encoding

Die Daten wurden danach in Trainings- und Testdatensätze aufgeteilt (80 % Training, 20 % Test), um die Modelle unabhängig zu trainieren und zu evaluieren.

5.2 Datenanalyse

Die explorative Datenanalyse ist ein entscheidender Schritt, um ein tieferes Verständnis für die Verteilung der Variablen und deren Beziehungen zur Zielvariable Exam_Score zu gewinnen. Mithilfe von Visualisierungen wurden die wichtigsten Muster und Zusammenhänge im Datensatz identifiziert. Die Verteilung der Zielvariable Exam_Score wurde mithilfe eines Histogramms untersucht. Dieses zeigt eine nahezu symmetrische Struktur, die typisch für eine Normalverteilung ist. Der Schwerpunkt der Noten liegt zwischen 60 und 75 Punkten. Diese Erkenntnis bestätigt, dass die Zielvariable repräsentativ für die Schülerpopulation ist und sich gut für die Modellierung eignet.

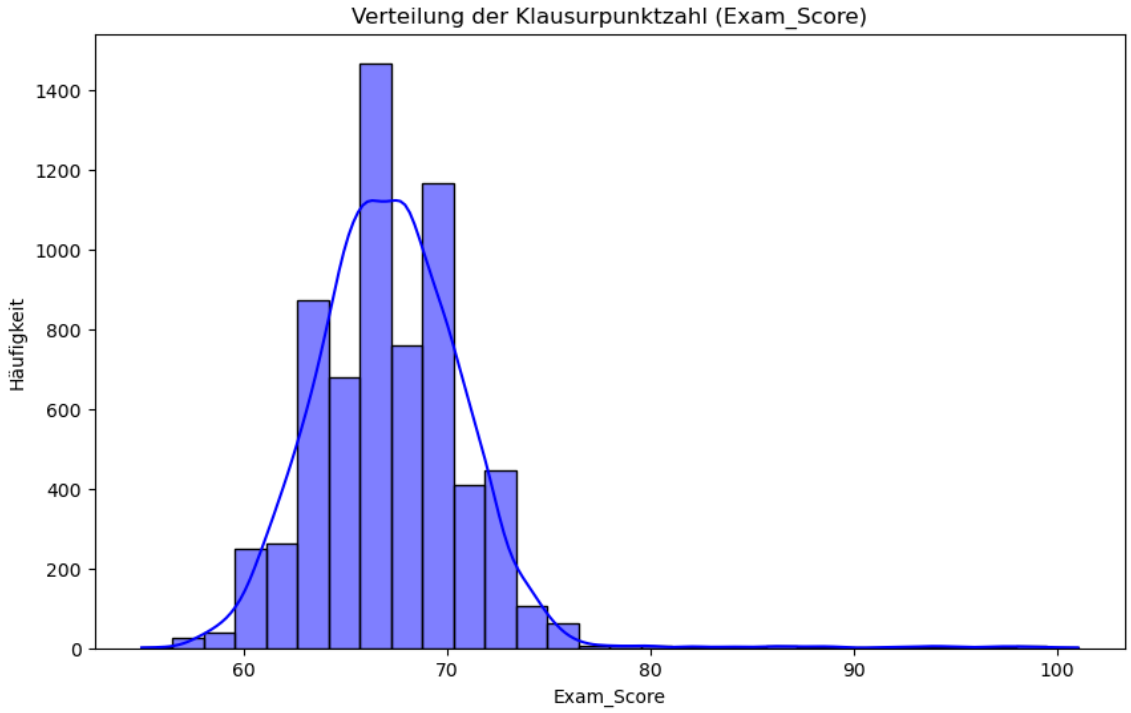

Abbildung 3: Verteilung der Klausurpunktzahlen

Ein Boxplot wurde verwendet, um die Verteilung der Lernzeit (Hours_Studied) in Abhängigkeit von den Quartilen der Klausurpunktzahlen zu analysieren. Die Schüler wurden in vier Leistungsgruppen aufgeteilt, und die Analyse zeigt, dass Schüler mit höheren Noten (oberstes Quartil) im Durchschnitt mehr Zeit in das Lernen investieren.

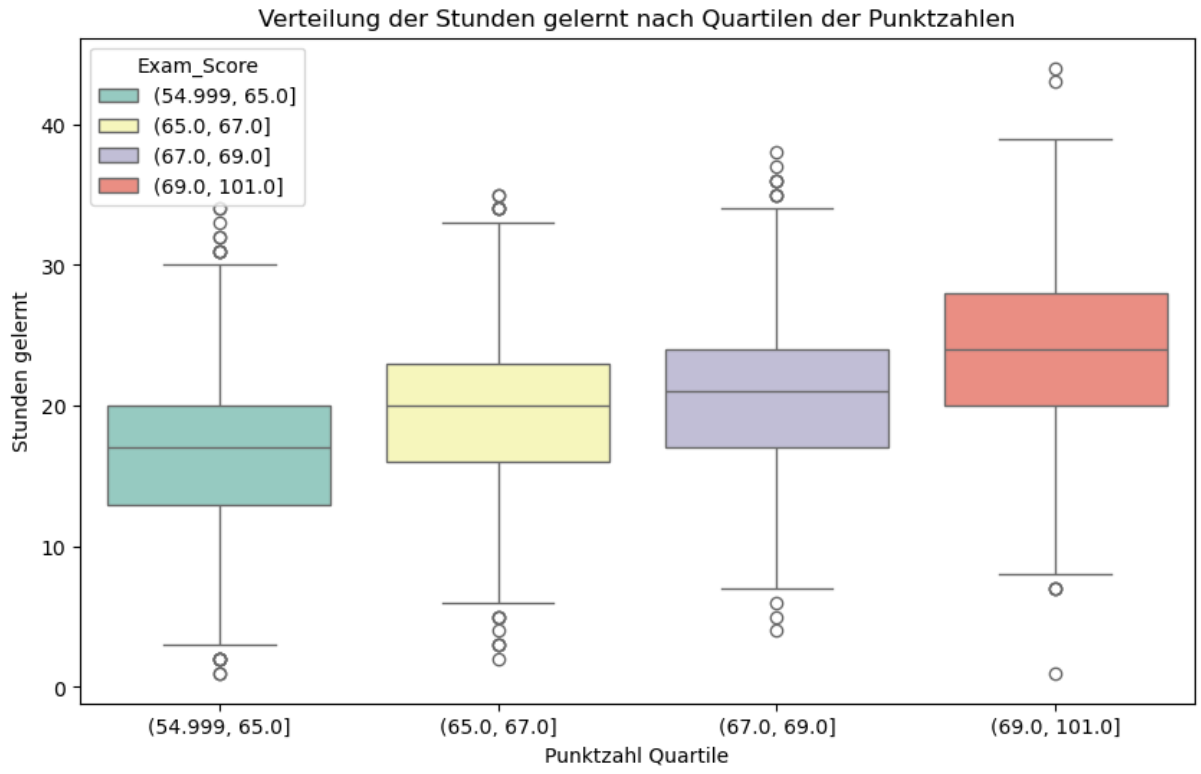

Abbildung 4: Boxplot der Lernzeit nach Quartilen der Punktzahlen

Ein Scatterplot wurde erstellt, um die Beziehung zwischen Anwesenheit (Attendance) und Klausurpunktzahlen (Exam_Score) zu untersuchen. Die Farbvariable Extracurricular_Activities wurde integriert, um den Einfluss außerschulischer Aktivitäten zu visualisieren. Die Analyse zeigt, dass Schüler mit höherer Anwesenheit bessere Noten erzielen, während außerschulische Aktivitäten keinen signifikanten Einfluss haben.

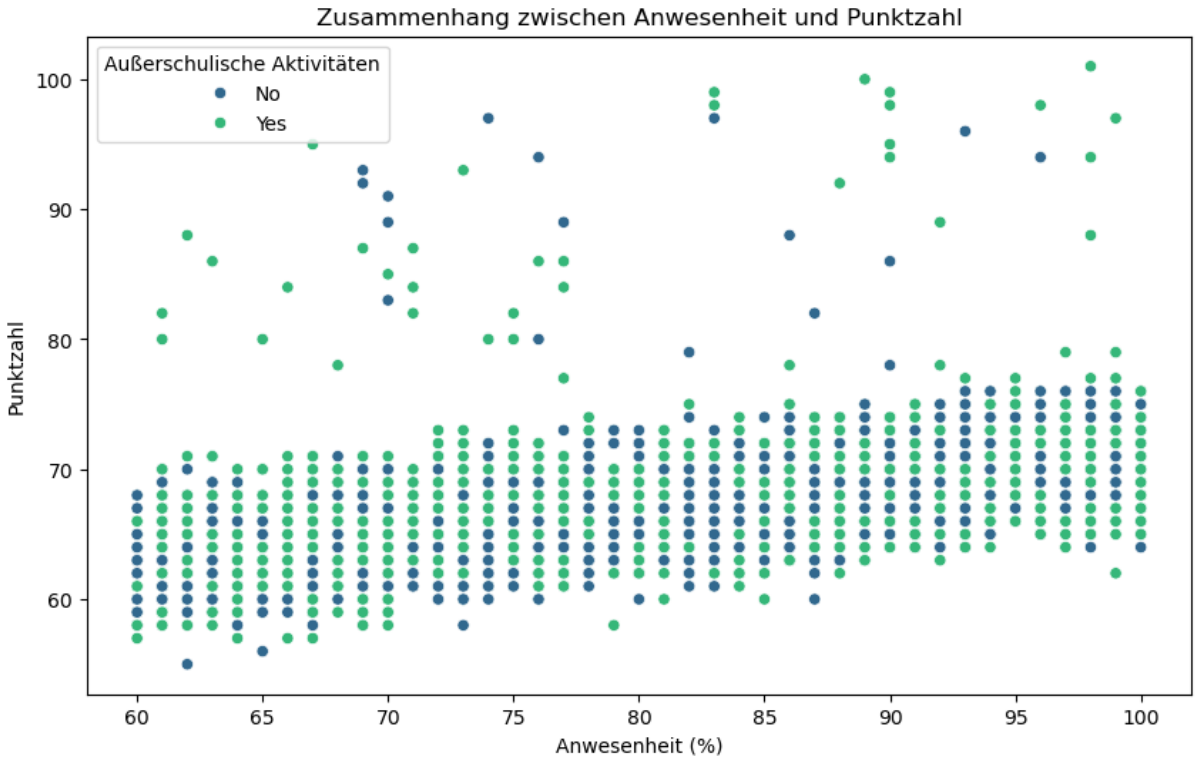

Abbildung 5: Zusammenhang zwischen Anwesenheit und Punktzahlen

Eine Heatmap der Korrelationsmatrix wurde erstellt, um die Beziehungen zwischen den numerischen Variablen zu analysieren. Die stärksten Korrelationen wurden bei Hours_Studied (r=0.68) und Attendance (r=0.55) festgestellt. Diese Variablen sind zentrale Prädiktoren für die Klausurpunktzahlen. Die diagonale Linie (dunkelrot) zeigt die perfekte Korrelation jeder Variable mit sich selbst (Korrelation = 1). Variablen wie Hours_Studied, Attendance und Previous_Scores weisen hohe positive Korrelationen mit der Zielvariable Exam_Score auf, was ihre Relevanz für die Modellbildung unterstreicht. Kategoriale Variablen wie Parental_Involvement und Teacher_Quality zeigen moderate Zusammenhänge mit Exam_Score, was darauf hindeutet, dass auch soziale und schulische Faktoren eine Rolle bei den Schülerleistungen spielen.

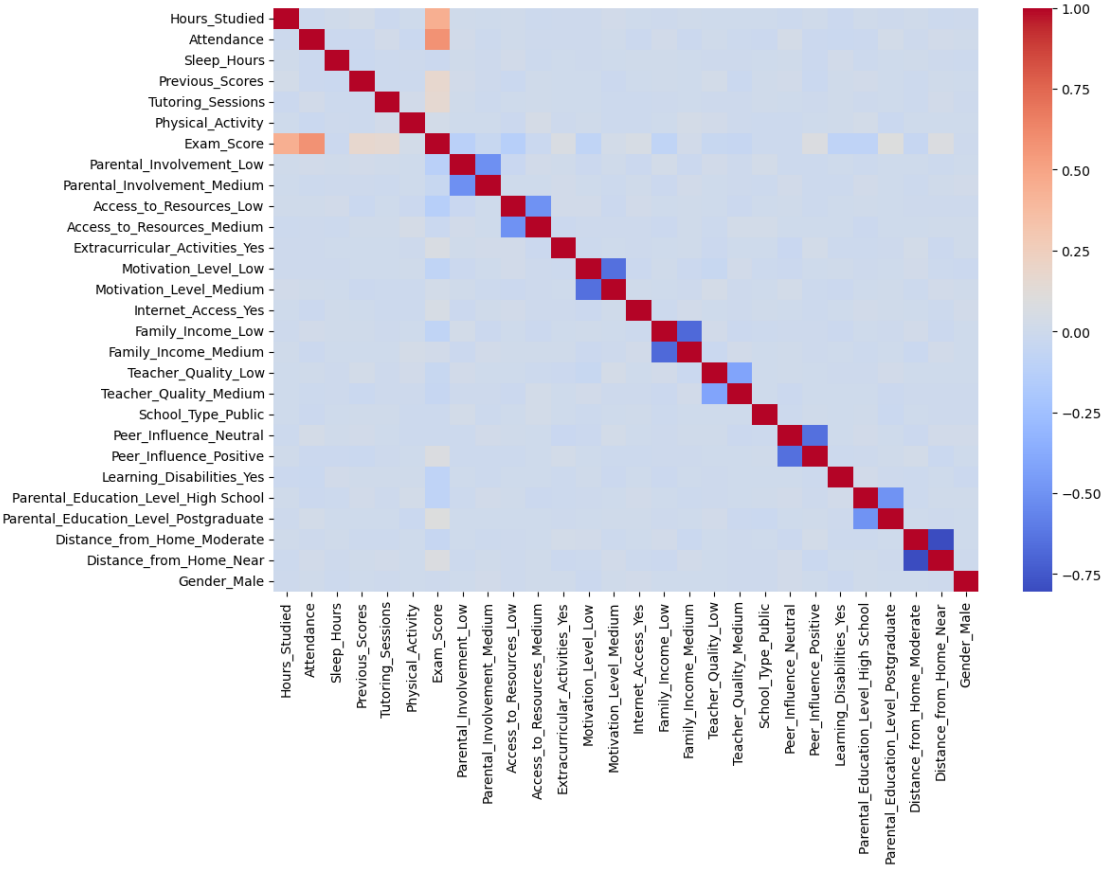

Abbildung 6: Heatmap der Korrelationsmatrix

5.3 Algorithmen

Zur Modellbildung wurden zwei Algorithmen eingesetzt: Random Forest und Lineare Regression. Beide Modelle wurden ausgewählt, da sie unterschiedliche Stärken besitzen: Während Random Forest nichtlineare Beziehungen zwischen Variablen modellieren kann, liefert die Lineare Regression einfache interpretierbare Ergebnisse.

1. Random Forest Regressor Der Random Forest Algorithmus ist ein Ensemble-Lernverfahren, das aus mehreren Entscheidungsbäumen besteht. Er ist besonders robust gegenüber Ausreißern und kann die Wichtigkeit einzelner Merkmale quantifizieren.

```
# Random Forest Modell
## Random Forest wurde als Modell gewaehlt, da es gut mit nichtlinearen Zusammenhaengen und kategorialen Daten umgehen kann
rf_model = RandomForestRegressor(random_state=42)
rf_model.fit(X_train, y_train)


# Vorhersagen und Evaluierung
rf_predictions = rf_model.predict(X_test)
```

Abbildung 7: Erstellung und Evaluierung des Random-Forest-Modells

2. Lineare Regression Die Lineare Regression wurde als Vergleichsmodell

genutzt, um zu evaluieren, wie gut einfache lineare Zusammenhänge die Klausurpunktzahlen erklären können.

```
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
```

```
# Vorhersagen und Evaluierung
lr_predictions = lr_model.predict(X_test)
print("Lineare Regression - MSE:", mean_squared_error(y_test, lr_predictions))
print("Lineare Regression - R2 Score:", r2_score(y_test, lr_predictions))
```

Abbildung 8: Erstellung und Evaluierung des Modells für Lineare Regression

Die Modelle wurden anhand des Mean Squared Error (MSE) und des R^2-Scores evaluiert, um deren Genauigkeit zu messen. Die Ergebnisse zeigen, dass der Random Forest Algorithmus eine deutlich bessere Leistung erzielt als die Lineare Regression.

```
print("Random Forest - MSE:", mean_squared_error(y_test, rf_predictions))
print("Random Forest - R2 Score:", r2_score(y_test, rf_predictions))
```

Abbildung 9: Evaluierung des Random-Forest-Modells

Um die wichtigsten Einflussfaktoren auf die Klausurpunktzahlen zu identifizieren, wurde die Feature-Importance des Random Forest Modells analysiert.

6 Datenbasierte Services

Basierend auf den Ergebnissen der vorliegenden Analyse und der entwickelten Modelle wurde ein datenbasierter Service konzipiert, der Schulen und Bildungseinrichtungen dabei unterstützt, gezielte Maßnahmen zur Verbesserung der Schülerleistungen zu ergreifen. Dieser Service verwendet die identifizierten Schlüsselmerkmale – darunter Hours_Studied, Attendance und Previous_Scores – sowie das trainierte Random-Forest-Modell, um Vorhersagen über die Klausurpunktzahlen zu treffen und datenbasierte Handlungsempfehlungen abzuleiten. Der Service zielt darauf ab, Schulen eine intuitive Plattform bereitzustellen, mit der sie:

1. Schüler mit potenziellen Leistungsproblemen frühzeitig identifizieren können,
2. gezielte Maßnahmen für individuelle Schüler vorschlagen und priorisieren können,
3. strategische Entscheidungen über die Allokation von Ressourcen basierend auf den Daten treffen können.

Der Service verarbeitet die Schülerdaten – wie Lernzeit, Anwesenheit und vorherige Leistungen – und speist sie in das trainierte Random-Forest-Modell ein. Das Modell klassifiziert daraufhin die zu erwartenden Klausurpunktzahlen und identifiziert die wichtigsten Einflussfaktoren, die zur Vorhersage beitragen. Die Ergebnisse werden in einem leicht verständlichen Format präsentiert, das sowohl Lehrkräften als auch Schuladministratoren als Entscheidungsgrundlage dient.

Eine Lehrkraft gibt die Daten eines Schülers wie folgt ein:

- Hours_Studied: 5 Stunden pro Woche,
- Attendance: 75 %,
- Previous_Scores: 60 Punkte.

Der Service verarbeitet diese Eingabedaten, transformiert sie in das erforderliche Format des Modells und liefert eine Vorhersage der zu erwartenden Prüfungsnote. Zusätzlich werden Empfehlungen zur Verbesserung der Schülerleistung gegeben.

7 Relevante Effekte für das Unternehmen

Die im Rahmen dieser Analyse gewonnenen Erkenntnisse und die daraus abgeleiteten datenbasierten Services können bedeutende Effekte für das Unternehmen EduPro Group und dessen Zielgruppen, insbesondere Schulen und Bildungseinrichtungen, haben. Die Implementierung der identifizierten Features und des trainierten Random-Forest-Modells erlaubt es, die Leistungen von Schülern datengetrieben zu analysieren, vorherzusagen und gezielt zu verbessern.

1. Verbesserung der Schülerleistungen

Ein zentraler Effekt des entwickelten Services ist die Möglichkeit, Schülerleistungen präzise vorherzusagen. Dies ermöglicht es Schulen, frühzeitig auf potenzielle Leistungsprobleme einzugehen und gezielte Fördermaßnahmen einzuleiten. Die identifizierten Schlüsselmerkmale – Hours_Studied, Attendance und Previous_Scores – bieten einen klaren Ansatzpunkt für individuelle Unterstützung. Schulen können beispielsweise Schülern mit geringer Lernzeit oder niedriger Anwesenheit spezifische Maßnahmen wie zusätzliche Nachhilfestunden oder Förderprogramme anbieten. Dies führt langfristig zu einer Verbesserung der individuellen und kollektiven Schülerleistungen.

2. Effizienzsteigerung in der Entscheidungsfindung

Ein weiterer Effekt des entwickelten Systems liegt in der Effizienzsteigerung für Lehrkräfte und Schuladministratoren. Durch die Bereitstellung datenbasierter Einblicke und prädiktiver Analysen entfällt die Notwendigkeit zeitaufwendiger manueller Auswertungen. Lehrkräfte können sich stattdessen auf die Umsetzung der vorgeschlagenen Maßnahmen konzentrieren, was die Effektivität des gesamten Bildungsprozesses erhöht.

3. Individuelle Förderung und Chancengleichheit

Die Analyse ermöglicht es, Schüler mit spezifischen Bedürfnissen oder Risikofaktoren zu identifizieren. Beispielsweise könnten Schüler aus Haushalten mit niedrigerem Bildungsniveau der Eltern *(Parental_Education_Level)* oder geringeren Ressourcen gezielt gefördert werden. Dies trägt zur Verbesserung der Chancengleichheit bei, da die Unterstützung auf Basis individueller Bedürfnisse und nicht pauschaler Annahmen erfolgt.

6. Wettbewerbsvorteil für das Unternehmen

Für EduPro Group bietet der entwickelte Service einen klaren Wettbewerbsvorteil im Markt für Bildungsanalytik. Durch die Bereitstellung eines leistungsstarken und benutzerfreundlichen Tools zur Vorhersage und Optimierung von Schülerleistungen kann das Unternehmen seine Position als führender Anbieter in diesem Bereich festigen.

8 Literatur

Baker, R. S. J. d., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 3–17.

Edupro Group. (2024). Bildung kennt keine Grenzen. https://edupro-group.com/

IBM (2024).Was ist Random Forest?. https://www.bing.com/search?q=IBM+(2024).Was+ist+Random+Forest%3F.&cvid=f491a2021920435c9147763f7948041a&gs_lcrp=EgRlZGdlKgYIABBFGDkyBggAEEUYOTIICAEQ6QcY_FXSAQc1MjBqMGo5qAIAsAIB&FORM=ANAB01&PC=U531

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, 46(5), 30–40.