Projektarbeit – Students Performance in Examns

M259

Noé Farese



Inhaltsverzeichnis

[1 Business Understanding 3](#_Toc160115068)

[2 Data Understanding 3](#_Toc160115069)

[2.1 Spaltenbeschreibung 3](#_Toc160115070)

[2.2 Einblick in das Datenset 4](#_Toc160115071)

[2.3 Wichtige Variablen 4](#_Toc160115072)

[2.4 Ausreisser 5](#_Toc160115073)

[2.4.1 Geschlecht vs Score 5](#_Toc160115074)

[2.4.2 Race/Enthnicity vs Score 5](#_Toc160115075)

[2.4.3 parental level of eduction vs score 5](#_Toc160115076)

[2.4.4 test preparation course attendance vs score 6](#_Toc160115077)

[2.5 Hypothesen 7](#_Toc160115078)

[2.5.1 Hypothese 1 7](#_Toc160115079)

[2.5.2 Hypothese 2 7](#_Toc160115080)

[2.5.3 Hypothese 3 7](#_Toc160115081)

[3 Data Preparation 8](#_Toc160115082)

[4 Modeling 8](#_Toc160115083)

[4.1 Lineare Regression 8](#_Toc160115084)

[4.2 Classification Decision Tree 9](#_Toc160115085)

[4.3 Untersuchung Zusammenhang von zwei Variabel-Paaren 9](#_Toc160115086)

[5 Evaluation 10](#_Toc160115087)

[5.1 Lineare Regression 10](#_Toc160115088)

[5.2 Classification Decision Tree 11](#_Toc160115089)

[5.3 Vergleich 11](#_Toc160115090)

[6 Fazit 12](#_Toc160115091)

[6.1 Resultate 12](#_Toc160115092)

[6.2 Ethische Probleme 12](#_Toc160115093)

# 1 Business Understanding

Ich arbeite mit dem Datenset namens *Students Performance in Exams*. Mein Ziel ist es nach dem Prinzip von CRISP-DM ein Regressionsmodell zu machen, welches die Punktzahlen in Mathematik-, Lese- und Schreibeprüfungen anhand von den anderen Variablen wie Geschlecht, Ethnizität, Bildungsniveau der Eltern und Testvorbereitungskurs hervorsagt. Ausserdem werde ich drei Hypothesen aufstellen und diese untersuchen.

# 2 Data Understanding

## 2.1 Spaltenbeschreibung

|  |  |
| --- | --- |
| gender | Geschlecht des Schülers/der Schülerin. **Values:** *male* und *female*  **Datentyp**: object |
| race/ethnicity | ethnische Zugehörigkeit des Schülers/der Schülerin. **Values:** *group A, group B, group C, group D, group E*  **Datentyp**: object |
| parental level of education | Bildungsstand der Eltern. **Values**:  *some college, associate’s degree, high school, some high school, bachelor’s degree*  **Datentyp**: object |
| lunch | Mittagessen. **Values**: *standard, free/reduced*  **Datentyp**: object |
| test preparation course | Angabe ob Schüler/Schülerin am Prüfungsvorbereitungskurs teilgenommen hat. **Values**: *none, completed*  **Datentyp**: object |
| math score | Erreichte Punktzahl in der Mathematikprüfung  Punktzahl geht maximal bis 100  **Datentyp**: int64 |
| reading score | Erreichte Punktzahl in der Leseprüfung  Punktzahl geht maximal bis 100  **Datentyp**: int64 |
| writing score | Erreichte Punktzahl in der Schreibeprüfung  Punktzahl geht maximal bis 100  **Datentyp**: int64 |

## 2.2 Einblick in das Datenset

**Aufbau:** 1000 Zeilen und 8 Spalten

**Datentypen:** Die Spalten math score, reading score und writing score haben als   
Datentyp int64. Die restlichen Spalten sind alle object. Die Datentypen stimmen damit.

**Erste zehn Spalten:** Ich habe die ersten zehn Spalten ausgegeben.

**Überblick/Describe:** Ich habe ein describe gemacht, um mir einen Überblick über die Score-Spalten zu schaffen. Bei allen drei Spalten ist die maximale Punktzahl, die man erreichen kann 100. Im Durchschnitt ist die Punktzahl beim *reading score* also bei der Leseprüfung am höchsten mit 69 Punkten. Die tiefste Punktzahl wurde bei der Mathematikprüfung erreicht mit 13 Punkten.

**Diagramme:** Ich habe zwei Säulendiagramme gemacht. Beim Ersten sieht man, dass es mehr Schüler gibt, welche männlich sind. Beim Zweiten kann man deutlich erkennen, dass die meisten Schüler nicht am *test preparation course* teilgenommen haben.

Danach habe ich noch für die drei Prüfungen boxplots gemacht. Dort sieht man noch visuell was ich beim Überblick/Describe beschrieben habe.

*Siehe Code 2.2*

## 2.3 Wichtige Variablen

Ich habe eine Correlation Matrix erstellt, um die wichtigen Variablen herauszufinden.

Für mich sind alle Variablen wichtig ausser *lunch*. *Lunch* hat zwar eine hohe Korrelation, aber ist für mein Modell nicht relevant. Es macht für mich nicht wirklich Sinn, dass ein Schüler einen höheren Score hat, wenn er kostenlos/reduziert oder Standard Mittag isst. Die anderen Variablen können alle interessant sein für die Verwendung. Zum Beispiel kann ich dann auswerten ob Schüler, welche am Prüfungsvorbereitungskurs teilgenommen haben besser waren als die Schüler welche nicht teilgenommen haben.

*Siehe Code 2.3*

## 2.4 Ausreisser

Ich habe für die Spalten *math score*, *reading score* und *writing score* jeweils ein Boxplot gemacht. Dort sieht man, dass es bei *math score* zwar nur zwei Ausreisser gibt, aber diese sind im Vergleich zu den anderen Ausreisser weit entfernt von einander. *Reading score* und *writing score* haben beide mehrere Ausreisser aber diese sind eher nah aneinander.

*siehe Code 2.4*

### 2.4.1 Geschlecht vs Score

Bei den Punkten in der Mathematikprüfung sieht man, dass der Durschnitt der Männer höher ist als bei den Frauen. Bei beide Geschlechtern wurde die volle Punktzahl erreicht.

Bei den Frauen gab es ein Ausreisser. Dort war jemand unter 20 Punkten.

Bei der Leseprüfung ist der Durschnitt bei den Frauen höher als bei den Männern. Bei den Frauen hat es dort einen Ausreisser. Der Ausreisser war wieder negativ, also nach unten, jedoch immer noch höher als das Minimum bei den Männern.

Bei der Schreibeprüfung ist der Durschnitt der Frauen wieder höher als bei den Männern. Bei den Frauen gab es zwei Ausreisser und bei den Männern ein Ausreisser.

*siehe Code 2.4.1*

### 2.4.2 Race/Enthnicity vs Score

Bei der Mathematikprüfung gibt es zwei Ausreisser. Einer bei der Gruppe C und einer bei der Gruppe D. Gruppe E hat den besten Durchschnitt von allen.

Bei der Leseprüfung gibt es zwei Ausreisser bei der Gruppe B. Gruppe E hat auch dort den besten Durchschnitt.

Bei der Schreibeprüfung gibt es bei der Gruppe B 3 Ausreisser. Bei der Schreibeprüfung hat die Gruppe D den besten Durchschnitt.

*siehe Code 2.4.2*

### 2.4.3 parental level of eduction vs score

Bei allen Prüfungen haben Schüler, welche Eltern mit dem *some high school* Abschluss haben den tiefsten Durchschnitt.

Bei der Mathematikprüfung gibt es drei Ausreisser. Zwei von diesen sind bei *associate’s degree* und einer bei *some college.*

Bei der Leseprüfung gibt es ein Ausreisser bei *associate’s degree* und einer bei *high school*.

Bei der Schreibeprüfung gibt es bei *high school* drei Ausreisser.

*siehe Code 2.4.3*

### 2.4.4 test preparation course attendance vs score

Die Schüler, die am Prüfungsvorbereitungskurs teilgenommen haben, sind im Durchschnitt bei allen drei Prüfungen besser als die Schüler, welche nicht teilgenommen haben.

Bei der Mathematikprüfung gibt es ein Ausreisser bei *none*, also bei der Gruppe welche nicht teilgenommen hat.

Bei der Leseprüfung gibt es zwei Ausreisser. Beide sind bei der Gruppe, die an dem Prüfungsvorbereitungskurs teilgenommen hat.

Bei der Schreibeprüfung gibt es einen Ausreisser bei *none,* dieser ist knapp unter dem Minimum.

*siehe Code 2.4.4*

## 2.5 Hypothesen

### 2.5.1 Hypothese 1

«Eine Schülerin erzielt im Durschnitt in jedem Fach (Math, Reading, Writing) bessere Noten als Schüler.»

Ich habe für jedes Fach beziehungsweise jeden Score ein Boxplot gemacht mit den Punkten von den Schülerinnen und Schüler beziehungsweise female und male.

Meine Hypothese stimmt nicht ganz. Beim *math score* sind haben die Schüler einen höheren Median als die Schülerinnen. Aber beim *reading score* und beim *writing score* haben die Schülerinnen einen höheren Median. Das Minimum ist bei diesen zwei Fächer bei den Schüler tiefer als bei den Schülerinnen.

*Siehe Code 2.5.1*

### 2.5.2 Hypothese 2

«Schüler und Schülerinnen mit ethnischer Zugehörigkeit von der Gruppe C sind, haben im Durchschnitt Eltern mit einem höheren Bildungsgrad (*parental level of education*)»

Ich musste zuerst *parental level of education* zu numerischen Werten umwandeln. Ich habe dafür einfach den Abschluss durch eine Zahl ersetzt (1-6). Je besser der Abschluss ist desto höher ist die Zahl. Danach habe ich von jeder Ethnizitätsgruppe den Durschnitt von dem *parental level of education* genommen und damit ein Balkendiagramm erstellt.

Meine Hypothese stimmt nicht. Gruppe C ist zwar von der Personenanzahl die grösste Gruppe (*siehe unter 2.2 im Code*), aber hat nicht den höchsten *parental level of education* Durchschnitt. Der höchste Durschnitt hat die Gruppe A. Von der Personenanzahl ist die Gruppe A die kleinste.

*Siehe Code 2.5.2*

### 2.5.3 Hypothese 3

«Schüler und Schülerinnen, die an einem Vorbereitungskurs teilnehmen, erzielen im Durchschnitt bessere Testergebnisse als diejenigen, die nicht teilnehmen.»

Ich habe für jeden Score zwei Boxplots gemacht, einen für *completed* und einen für *none*, also ob man teilgenommen hat oder nicht. Zusätzlich habe ich noch ein Säulendiagramm gemacht, dort sieht man es noch ein bisschen besser.

Meine Hypothese stimmt. In allen Scores sind die Schüler und Schülerinnen, welche an dem Vorbereitungskurs teilgenommen haben, besser als die welche nicht teilgenommen haben.

*Siehe Code 2.5.3*

# 3 Data Preparation

Bei der Datenaufbereitung habe ich überprüft ob folgende Sachen existieren:  
**Duplikate**

Es existierte ein Duplikat und dieses habe ich dann auch gleich gelöscht.

**Fehlende Werte**

Es gab keine fehlende Werte oder NaN Werte.

**Falsch geschrieben Werte**

Ich habe das CSV File durchgeschaut und habe keine falsch geschriebene Werte gesehen.

**Werte in falschen Feldern**

Ich habe das CSV File durchgeschaut und habe keine Werte in den falschen Feldern gesehen.

**Irregularitäten**

Ich habe das CSV File durchgeschaut und keine Irregularitäten.

*Sehe Code 3 Data Preparation*

# 4 Modeling

## 4.1 Lineare Regression

Zuerst habe ich das Mapping gemacht, damit ich überall Zahlen habe und keine objects mehr. Danach habe ich mein x und y gesetzt. Y ist *math score.* X waren die restlichen Spalten (zuerst ohne *lunch* und dann noch mit *lunch*). Danach habe ich die Daten in Train und Test gesplittet und dann das Modell mit linear regression trainiert.

**Ziel von Modell**

Das Ziel ist, dass das Modell anhand von den Features hervorsagen kann, wie hoch der *math score* einer Person ist.

**Features**

*Gender, race/ethnicity, parental level of education, test preparation course, writing score, reading score, lunch*

## 4.2 Classification Decision Tree

Das Mapping musste ich hier nicht mehr machen, weil ich es schon oben beim Regression Modell gemacht habe. Danach habe ich wieder mein x und y gesetzt. Y war die Spalte *race/ethnicity* und x allen anderen Spalten (auch hier wieder zuerst ohne *lunch* und dann mit *lunch*). Danach habe ich meine Daten wieder in Train und Test gesplittet und dann das Modell mit dem Decision Tree trainiert.

**Ziel von Modell**

Das Ziel ist, dass das Modell anhand von den Features erkennen kann, welche *race/ethnicity* der Schüler oder die Schülerin hat.

**Features**

*Gender, parental level of education, test preparation course, math score, writing score, reading score*

## 4.3 Untersuchung Zusammenhang von zwei Variabel-Paaren

**Erstes Variabel-Paar**

In der Korrelation Matrix sieht man, dass sich *math score* und *writing score* stark gegenseitig beeinflussen (0.81). Also man könnte daraus schliessen, dass Schüler und Schülerinnen, welche gut in Mathematik auch gut um Schreiben sind.

**Zweites Variabel-Paar**

In der Korrelation Matrix sieht man, dass sich writing score und test preparation course gegenseitig auch beeinflussen, aber nicht wirklich stark (0.32/-0.32). Daraus kann man schliessen, dass Schüler oder Schülerinnen welche am Kurs teilnahmen besser abgeschlossen haben, als die welche nicht am Kurs teilgenommen haben.

*Siehe Code 4.3 (ganz unten)*

# 5 Evaluation

## 5.1 Lineare Regression

Zuerst habe ich *lunch* nicht reingenommen. Danach habe ich aber noch *lunch* reingenommen. Der R2 Score ist gerundet 0.85. Also eigentlich ziemlich gut finde ich. Zuerst wollte ich die Gesamtpunktzahl vorhersagen, also math score, writing score und reading score zusammen, aber dort war der R2 Score tiefer, ungefähr zwischen 0.2 und 0.3

**Scatterplot**

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

Automatisch generierte Beschreibung

*Siehe Code 5.1 direkt unter 4.1*

## 5.2 Classification Decision Tree

Beim Classicfication Decision Tree habe ich zuerst *lunch* nicht ins x reingenommen. Dort war die Genauigkeit meistens 0.41. Danach habe ich *lunch* ins x reingenommen und dann war die Genauigkeit 0.4 Also hier hat sich nicht viel verändert. Das war mit einer max\_depth von 5. Wenn ich die max\_depth auf 15 setze, ist die Genauigkeit bei 0.95.

**Accuracy mit max\_depth=5**

0.4

**Accuracy mit max\_depth=15**

0.95

**Classification report mit max\_depth=15**

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

*siehe Code 5.2 direkt unter 4.2*

## 5.3 Vergleich

Der Decision Tree ist bisschen genauer als das Regressionmodell, weil wenn ich beim Decision Tree die max\_depth auf 15 habe ist er sehr genau. Er hat dann einen f1-score von etwa 0.9 was beinahe perfekt ist. Aber das Regressionmodell hat ein R2 Score von 0.85 und dies ist aber auch ziemlich gut.

# 6 Fazit

## 6.1 Resultate

Das erste Modell, das Regression Modell war zuerst nicht sehr gut. Ich wollte zuerst eben den Gesamtscore voraussagen aber ich habe es dann auf *math score* geändert und da *math score*, writing score und reading score eine hohe Korrelation haben, wurde das Modell viel besser. Das zweite Modell mit dem Classification Decision Tree war schon genauer, vorallem ab dann wo ich max\_depth auf 15 gestellt habe. Es war ein eher kleiner Datensatz, weil er nur 8 Spalten hatte. Darum gab es zum Beispiel nur drei int64 und der Rest waren objects. Positiv daran war aber, dass ich bei der Data Preparation nicht viel aufräumen musste, weil es ein ziemlich sauberer Datensatz ist.

## 6.2 Ethische Probleme

Ethische Probleme gibt es bei diesem Modell definitiv. Bei der Spalte von der Herkunft der Personen, also die *race/ethnicity* Spalte waren die Values anonymisiert. Sie wurden in Gruppen aufgeteilt. Wären sie nicht in Gruppen aufgeteilt gewesen, hätte es auch schnell rassistisch werden können.

Die Vergleiche mit männlich und weiblich sind auch heikel, weil es als sexistisch wahrgenommen werden kann. Ausserdem gab es noch die Spalte mit dem Bildungsstand der Eltern. Dort kann es auch als diskriminierend wahrgenommen werden, wenn Eltern von einer bestimmten Gruppe einen besseren Ausbildungsstand haben als andere Eltern.