BGD_DatasetOverview_Altendorfer_Eckmayr

December 17, 2023

1 Datensatzbeschreibung

1.1 Gruppe Altendorfer, Eckmayr

Datensatz: StudentsPerformance

Der von uns gewählte Datensatz enthält Daten zu Ergebnissen in den Fächern Mathematik, Lesen und Schreiben von 1000 US-amerikanischen Highschool Schülern. Wir möchten in der weiteren Analyse untersuchen, ob Zusammenhänge zwischen einzelnen/mehreren Faktoren und den erzielten Testergebnissen hergestellt und durch die angegeben Features der Test Score vorhergesagt werden kann sowie welche Variablen die größte Signifikanz aufweisen. Der Datensatz ist auf Kaggle verfügbar.

1.1.1 Konkret sind folgende Variablen enthalten:

```
['gender', 'race/ethnicity', 'parental level of education', 'lunch', 'test
preparation course', 'math score', 'reading score', 'writing score']
```

	gender	race/ethnicity	parental level of education	lunch	\
0	female	group B	bachelor's degree	standard	
1	female	group C	some college	standard	
2	female	group B	master's degree	standard	
3	male	group A	associate's degree	free/reduced	
4	male	group C	some college	standard	

	test preparation course	math score	reading score	writing score
0	none	72	72	74
1	completed	69	90	88
2	none	90	95	93
3	none	47	57	44
4	none	76	78	75

1.1.2 Ein näherer Blick auf die jeweiligen Werte der Variablen:

```
gender: ['female' 'male']
race/ethnicity: ['group B' 'group C' 'group A' 'group D' 'group E']
parental level of education: ["bachelor's degree" 'some college' "master's
degree" "associate's degree"
 'high school' 'some high school']
```

```
lunch: ['standard' 'free/reduced']
test preparation course: ['none' 'completed']
```

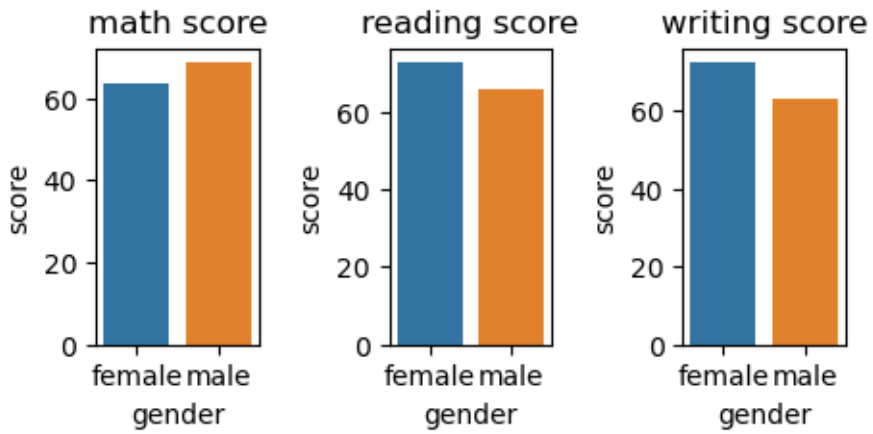
1.1.3 Zusammenfassung des Datensatzes:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64 
 6   reading score                1000 non-null   int64 
 7   writing score                1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

1.2 Forschungsfragen

Aus dem vorliegenden Datensatz können einige spannende Forschungsfragen abgeleitet werden. Wir wollen in unserer Analyse einige Variablen herausgreifen und diese hinsichtlich der erzielten Testergebnisse beurteilen und die Resultate visualisieren. Beispielhaft für das Geschlecht:

```
C:\Users\fabia\AppData\Local\Temp\ipykernel_8100\3083268956.py:7:
MatplotlibDeprecationWarning: Auto-removal of overlapping axes is deprecated
since 3.6 and will be removed two minor releases later; explicitly call
ax.remove() as needed.
  plt.subplot(1,3, idx+1)
```



1.2.1 Forschungsfrage 1

Kann ein Unterschied zwischen den ethnischen Gruppen in den erzielten Prüfungsergebnissen beobachtet werden?

1.2.2 Forschungsfrage 2

Beeinflusst das Bildungsniveau der Eltern die Wahrscheinlichkeit einen Vorbereitungskurs zu absolvieren und wirkt sich die Absolvierung eines solchen auf die Prüfungsergebnisse aus?

1.2.3 Forschungsfrage 3

Besteht ein statistisch signifikanter Zusammenhang zwischen der Bereitstellung eines Mittagessens und den Leistungen in Prüfungen, wobei höhere Prüfungsergebnisse bei Schülern beobachtet werden, die Zugang zu einem Mittagessen haben, im Vergleich zu Schülern ohne diesen Zugang?

1.2.4 Forschungsfrage 4 - Machine Learning

Kann unter Verwendung der Variablen 'Parental Level of Education', 'Lunch' und 'Test Preperation Course' mittels einer ausgewählten Machine Learning Methode die Note des Prüfungsergebnisses vorhergesagt werden, wobei die Prüfungsergebnisse in Noten nach dem österreichischen Schulsystem (1-5) eingeteilt werden?