Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 2

Sommersemester 2021

Übungsblatt 2

Aufgabe 1 (Auto MPG Datensatz)

Der Auto/Verbrauchs-Datensatz ist in der Datei `auto-mpg.csv` enthalten. Sie finden die Datei im Verzeichnis `Vorlesung/data`.

1. Der Verbrauch wird in `mpg` gemessen - *miles per gallon*. Bei uns ist ja die Angabe $l/100km$ üblich. Berechnen Sie eine Spalte `l_100km`, die die entsprechende Angabe in $l/100km$ enthält.

   Nutzen Sie dafür:

   - 1 *gallon* = 4.54609 Liter
   - 1 *mile* = 1.60935 km

2. Wie hoch ist der minimale / maximale / durchschnittliche Verbrauch der Autos in dem Datensatz?

3. Jedem Automodell ist ein Baujahr/Modelljahr zugeordnet. Plotten Sie den minimalen/maximalen/durchschnittlichen Verbrauch je Modelljahr.
   Sind die Autos über die Jahre sparsamer geworden?

4. Erstellen Sie einen Plot des Verbrauchs `l_100km` gegen das Gewicht und einen Plot von `l_100km` gegen den Hubraum `displacement`.

Aufgabe 2 (Regression mit Bäumen)

Die SciKit Learn Bibliothek enthält auch einen Baum-basierten Regressor:

```
from sklearn.tree import DecisionTreeRegressor
```

1. Normalisieren Sie die Auto MPG-Daten mit dem `MinMaxScaler`. Achten Sie dabei darauf, dass der `MinMaxScaler` erwartet, dass alle Spalten numerische Werte enthalten.

2. Teilen Sie die Daten in Trainings- und Testdaten auf (80% Trainingsdaten).

3. Trainieren Sie `DecisionTreeRegressor` Modelle mit verschiedenen maximalen Tiefen (2 bis 10) und bestimmen Sie jeweils den *mean squared error* (`mse`) auf den Testdaten.
   Erzeugen Sie einen DateFrame, der die Tiefe des Baumes und den `mse` enthält.

4. Trainieren Sie auf den gleichen Trainingsdaten verschiedene lineare Regressionsmodelle mit den Graden 1 bis 6 und erzeugen Sie ebenfalls einen DataFrame mit dem Grad des Modells und dem zugehörigen `mse`.

5. Plotten Sie die Generalisierungsfehler für ihre Baum-Lerner und die Regressionsmodelle.