**Data Exploration Project – Report**

Dieses Projekt wird im Rahmen der Vorlesung „Data Exploration Projekt“ bei Nils Zerrer im vierten Semester durchgeführt. Ziel ist es einen Machine Learning Algorithmus anhand beliebiger Daten zu trainieren und eine Klassifikation oder Vorhersage zu erstellen.

In diesem Projekt wird mit Hilfe einer linearen Regression eine Vorhersage – basierend auf den gegebenen Daten – getätigt.

# 1. Der Datensatz: World Happiness Report

Der für dieses Projekt ausgewählte Datensatz kann auf Kaggle unter dieser URL gefunden werden: <https://www.kaggle.com/unsdsn/world-happiness>.

Der „World Happiness Report“ beinhaltet Daten, die aus einer weltweiten Umfrage ermittelt worden sind. Obwohl bereits Daten seit dem Jahr 2012 erhoben werden, werden in diesem Projekt nur die Daten aus dem Jahr 2019 betrachtet.

## 1.1 Größe des Datensatzes

Der Datensatz besteht 156 Zeilen, bei dem jede Zeile einem anderen Land entspricht. Jede Zeile beinhaltet die Daten für die neun Feature: “*Country or region*”, “*Overall rank*”, “*Score*”,” *GDP per capita*”, “*Social support*”, “*Healthy life expectancy*”, “*Freedom to make life choices*”, “*Generosity*” und “*Perceptions of corruption*”.

### 1.1.1 Country or region

Dieses Feature beinhaltet das Land – bzw. die Region – aus dem die Daten erhoben worden sind.

### 1.1.2 **Overall rank**

Dieses Feature gibt die Platzierung des entsprechenden Landes in der Rangfolge für das Jahr 2019 an.

### 1.1.3 Score

In diesem Feature ist die Punktzahl angegeben, die von dem entsprechenden Land in der Bewertung erzielt worden ist. Zehn Punkte hätten hier erreicht werden können.

### 1.1.4 GDP per capita

Dieses Feature gib das Bruttoinlandsprodukt (BIP) pro Person des entsprechenden Landes an. Dieser Wert gibt, wie sehr sich das BIP auf die Bewertung des entsprechenden Landes auswirkt.

### 1.1.5 Soical Support

In diesem Feature befindet sich der Wert zur allgemeinen sozialen Unterstützung im entsprechenden Land. Weitergehend gibt der Wert an, wie sehr sich die soziale Unterstützung auf die Bewertung des entsprechenden Landes auswirkt.

### 1.1.6 Healthy life expectancy

Dieses Feature beinhaltet den Wert, der angibt wie stark sich die Lebenserwartung in dem entsprechenden Land auf die Bewertung des Landes auswirkt.

### 1.1.7 Freedom to make life choices

Der Wert in diesem Feature gibt an wie stark sich die Möglichkeit eigene Lebensentscheidungen zu treffen auf die Beurteilung des entsprechenden Landes auswirkt.

### 1.1.8 Generosity

„Generosity“ gibt an, wie sehr sich die Großzügigkeit des entsprechenden Landes sich auf dessen Bewertung auswirkt.

### 1.1.9 Perception of corruption

In diesem Feature gibt der Wert an, wie stark in dem entsprechenden Land Korruption wahrgenommen wird und wie stark sich dieser Wert auf die Bewertung auswirkt.

# 2. Datenvorbearbeitung

Bevor die Daten für das Trainieren des Modells genutzt werden können, werden diese noch ein wenig angepasst.

Das Feature „*Country or region*” wird als eineindeutigen Index verwendet und ist damit leicht zu identifizieren und zu zuordnen.

# 3. Feature Engineering

# 4. Split des Datensatzes

Der vorbereitete Datensatz wird in drei Teile aufgeteilt. In Trainingsdaten, Validierungsdaten und Testdaten. Dabei wird die Funktion *train\_test\_splits()* von *scikit-learn* zweimal verwendet.

Im ersten Durchlauf wird der Datensatz in Trainingsdaten und Testdaten mit einem Verhältnis von 90% / 10% aufgeteilt. In diesem Durchlauf wurde mit dem Parameter *random\_state* dafür gesorgt, dass die Daten zufällig aufgeteilt worden sind, damit in den Trainingsdaten nicht nur die Länder sind, die eine niedrige Bewertung erhalten haben.

Im zweiten Durchlauf werden Trainingsdaten erneut in Trainingsdaten und Validierungsdaten mit einem Verhältnis von 75% / 25% aufgeteilt. Hier wurde nicht zufällig aufgeteilt, da die Daten bereits aus der ersten Aufteilung in einer zufälligen Anordnung waren.

## 4.1 Trainingsdaten

Die Trainingsdaten werden hierbei fürs Trainieren des Machine Learning Modells verwendet und bilden deswegen auch den größten Teil mit etwa 67% (105 Records).

## 4.2 Validierungsdaten

Die Validierungsdaten werden dazu verwendet das trainierte Modell zu überprüfen und zu optimieren und haben dabei den zweitgrößten Teil mit etwa 23% (35 Records).

## 4.3 Testdaten

Die Testdaten werden für Demonstration des Modells verwendet und haben deswegen den kleinsten Anteil mit etwa 10% (16 Records).

# 5. Das ausgewählte Machine Learning Modell

Für die Vorhersage der Bewertung der Zufriedenheit in einem Land wird eine lineare Regression verwendet.

## 5.1 Wie funktioniert dieser Datensatz?

## 5.2 Warum dieser Datensatz?

# 6. Die ausgewählten Metriken

- Mean squared Error

- Mean absolute Error

- Mean absolute percentage Error (?)

# 7. Training des Modells

# 8. Tuning der Hyperparameter

Für das Tunen der Hyperparameter wurde *MLFlow* verwendet.

# 9. Evaluation mit den Testdaten

# 10. Mögliche Schwachstellen und Verbesserungsmöglichkeiten