Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science 1

Sommersemester 2021

Probeklausur

Die Prüfungsleistung zum Modul *Data Science 1* findet als Hausarbeit statt. Die Aufgabenstellung zur Hausarbeit finden Sie in diesem Dokument.

Für die Bearbeitung der Aufgabenstellung und die Erstellung Ihrer Hausarbeit steht wieder der Jupyter-Notebook Server zu Verfügung. Die Abgabe der Hausarbeit erfolgt dann als PDF-Export Ihres Jupyter-Notebooks. Das PDF Ihres Notebooks laden Sie als Lösung in der zugehörigen Aufgabe im Moodle Kurs hoch.
Andere Formen der Abgabe sind nicht vorgehen.

Für die Bearbeitung der Aufgabenstellung haben Sie ab Themenausgabe eine Woche Zeit. Der exakte Zeitraum wird in der zugehörigen Aufgabe im Moodle Kurs vermerkt.

Als Materialien können Sie sämtliche Unterlagen aus der Vorlesung und den Übungen mit benutzen, im Internet recherchieren oder weitere Bücher/Kurse mit verwenden. Geben Sie bitte bei Verwendung von umfangreicherem Programm-Code aus dem Netz (mehr als 3-4 Zeilen) die Quelle kurz mit an.

Aufgabe 1 (Python Basics)

Es ist eine Liste von Tupeln gegeben, von denen jedes Tupel die folgende Form hat:

```
(hoerspiel,folgenNr,laenge)
```

Die `hoerspiel` Komponente ist ein String, die `folgenNr` eine ganze Zahl und `laenge` ein String in der Art `minuten:sekunden`, also z.B.

```
("Inspektor Hase", 23, "22:37")
```

Die Liste der Tupel bekommen Sie über die Funktion `hoerspiele()` die im Modul `datascience` enthalten ist. Es gibt in der Liste Hörspiele, zu denen mehrere Tupel gehören, weil diese aus mehreren Folgen bestehen.

1. Bestimmen Sie, wieviele *verschiedene* Hörspiele es gibt! Schreiben Sie dazu eine Funktion `anzahlHoerspiele(xs)`, die für die Liste von Hörspiel-Tupeln die Anzahl der verschiedenen Hörspiele berechnet.

2. Schreiben Sie eine Funktion `anzahlFolgen(xs)`, die für die Liste der Tupel eine Liste zurückliefert, die für jedes Hörspiel ein Paar mit

   ```
   (hoerspiel,anzahlFolgen)
   ```

   enthält, also z.B.

   ```
   [("Inspektor Hase", 8), ("Bibi Blocksberg", 35), ...  ]
   ```

3. Schreiben Sie eine Funktion `sekunden(s)`, die einen String `s` im Format `minuten:sekunden` in die Anzahl der Sekunden umrechnet. Berechnen Sie aus der Liste der Hörspiel-Tupel eine neue Liste, bei dem jedes Tupel nun noch die Länge in Sekunden als neue vierte Komponente enthält.

   Hinweis: Um einen String `s` in eine ganze Zahl zu konvertieren, können Sie `int(s)` aufrufen.

Aufgabe 2 (Pandas und Statistiken)

Ein Hersteller für Kinder-Musikboxen zeichnet das Abspielverhalten der Boxen auf. Dabei wird für jedes Kind/jede Box und jede Hörspielfolge gespeichert, an welchem Tag das Hörspiel wie oft gespielt wurde.

Kind	Hörspiel	Folge	Mo	Di	Mi	Do	Fr	Sa	So
K7	Bibi&Tina	Folge 1	3	0	0	3	0	1	3
K5	Feuerwehrmann Sam	Folge 3	1	2	1	3	3	3	2
K2	Paw Patrol	Folge 9	0	3	1	2	2	1	3
K6	Benjamin Blümchen	Folge 2	0	0	0	3	1	0	3
K5	Bibi&Tina	Folge 6	3	2	3	1	1	1	2

1. Berechnen Sie eine Spalte `Gesamt`, die für jede Zeile die Summe der Häufigkeiten der Hörspielfolge enthält.

   Welche Hörspielfolge wurde am häufigsten gespielt?

   Welche Hörspielfolge wurde am seltensten gepielt?

2. Schreiben Sie eine Funktion `top_folge(df)`, die für einen DataFrame mit der zuvor definierten `Gesamt`-Spalte die Zeilen mit den am häufigsten gespielten Hörspielfolgen zurückgibt. Wenn es mehrere Hörspielfolgen gibt, die die höchste Häufigkeit haben, sollen alle entsprechenden Zeilen zurückgegeben werden.

3. Schreiben Sie eine Funktion `top_hoerspiel(df)`, die das am häufigsten gespielte Hörspiel (über alle Folgen) für einen entsprechenden DataFrame `df` zurückgibt.

   Wenn es mehrere Hörspiele mit gleicher Gesamthäufigkeit gibt, sollen alle häufigsten Hörspiele zurückgegeben werden.

4. Welcher Wochentag ist derjenige, an dem die meisten Hörspiele-Folgen gehört werden?

Aufgabe 3 (Modell-Training)

Der Hersteller für die Kinder-Musikboxen hat für einige Kunden zusätzliche Daten gesammelt. Für die Boxen dieser Kunden ist bekannt, ob die Box einem *Mädchen* oder einem *Jungen* gehört.
Die Daten finden Sie im Verzeichnis `Kurse/DataScience1/data` in der Datei `musik-kids.csv`

1. Laden Sie die Daten in einen DataFrame und geben Sie die Anzahl der Datensätze, sowie die Anzahl der Jungen/Mädchen an.

2. Teilen Sie die Daten in Trainings- und Test-Daten auf und verwenden Sie dabei 80% der Daten zum Training und den restlichen Teil für das Testen.

3. Trainieren Sie ein lineares SVM Modell auf ihren Daten. Welchen Trainingsfehler erreicht ihr Modell?

4. Bestimmen Sie den Parameter `C`, der auf den Daten für ein lineares SVM-Modell den besten Generalisierungsfehler liefert. Testen Sie für die Bestimmung des Parameters `C` 10 Werte im Bereich von 1 bis 100.

   Erzeugen Sie dazu einen DataFrame, der den Parameter `C`, den Trainings- und den Test-Fehler enthält.