h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Prof. Dr. H. Zisgen
Markus Schepers
Fachbereich Mathematik und Naturwissenschaften
Sommersemester 2020

Data Mining 1

Pflichtaufgaben – Teil I
- Entscheidungsbäume -

Die folgenden Aufgaben sollen mit der Programmiersprache `Python` und der Bibliothek `sklearn` bearbeitet werden. Im Moodlekurs finden Sie im Abschnitt **Aufgaben** einen Ordner **Australische Wetterdaten**. In diesem finden Sie eine Beschreibung der Variablen und den Datensatz als CSV-Datei. Die folgenden Aufgaben beziehen sich auf diesen Datensatz.

Um den Rechenaufwand zu minimieren, sollen die Aufgaben nur auf einer Teilmenge der Daten berechnet werden. Der gefilterte neue Datensatz soll Daten aus zwei verschiedenen Jahren, entsprechend der folgenden Gruppeneinteilung, enthalten.

Gruppe	Jahr 1	Jahr 2
01	2009	2012
02	2009	2013
03	2009	2014
04	2010	2013
05	2010	2014
06	2010	2015
07	2011	2014
08	2011	2015
09	2011	2017
10	2012	2015
11	2012	2017
12	2012	2018
13	2013	2017
14	2013	2018
15	2013	2019

Aufgabe 1 (Feature Engineering)
Im nachfolgenden Link können Sie ein Tutorial zu *Feature Engineering* mit Beispielen finden.

- Überlegen Sie sich, an welchen Stellen im Datensatz *Feature Engineering* sinnvoll ist und setzen Sie dieses in den folgenden Aufgaben um.
- Seien Sie kreativ!

`https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114`

⇒ **Hinweis:**

- Bitte beachten Sie in Ihren Überlegungen, dass die Features nur numerische Werte und keine fehlenden Werte besitzen dürfen.

Aufgabe 2 (Entscheidungsbäume)

a) Teilen Sie den Datensatz in eine Trainings- und eine Testmenge auf. Überlegen Sie sich welche Aufteilungsverhältnis sinnvoll ist.

b) Erstellen Sie einen Entscheidungsbaum mit dem Modul `tree.DecisionTreeClassifier` aus der Bibliothek `sklearn`. Nutzen Sie hierzu die Default-Einstellungen des Moduls. Visualisieren Sie den Baum und interpretieren Sie diesen. Wie gut ist das Modell?

c) Erstellen Sie verschiedene Entscheidungsbäume. Variieren Sie ausgehend von den Default-Einstellungen die folgenden Parameter einzeln:

   - `max_depth`
   - `min_impurity_decrease`
   - `criterion`

   Wie unterscheiden sich die einzelnen Bäume? Für welchen Baum würden Sie sich entscheiden? Begründen Sie Ihre Entscheidung!

d) Führen Sie mit mindestens 3 Bäumen unterschiedlicher Tiefe aus Aufgabenteil c) ein *Minimal Cost-Complexity Pruning* durch. Wie verändern sich die Bäume bei Variation des Prunings? Welche Auswirkung auf die Modellgüte hat dies?

⇒ **Hinweise:**

- Zufälliges Aufteilen eines Datensatzes kann mit Hilfe des Moduls `sklearn.model_selection.train_test_split` durchgeführt werden.
- Nutzen Sie Random Seeds um reproduzierbare Ergebnisse zu erhalten! Viele Module bieten hierzu einen eigenen Parameter.
- Documentation: `sklearn.tree.DecisionTreeClassifier` (Hyperlink)
- User Guide: Decision Trees (Hyperlink)
- Example: Post pruning decision trees with cost complexity pruning (Hyperlink)