Data Science 1

Übungsblatt 5

Aufgabe 1 (Klassifikation von Schokokugeln)
Unter der URL

`https://datascience.hs-bochum.de/datenfabrik/camera/4/data.csv`

stehen die Daten des Farbsensors von unserem Schokokugel-Sortierer als CSV-Datei zur Verfügung. Darin enthalten sind u.a. die RGB-Werte für die einzelnen Kugelfarben.

1. Lesen Sie die Daten in einen Pandas DataFrame ein und berechnen Sie die durchschnittlichen RGB Werte für die roten und die blauen Kugeln.

2. Erstellen Sie einen DataFrame, der nur die Daten der roten und blauen Kugeln enthält. Dazu ist das Filtern mit einer OR-Verknüpfung erforderlich, was z.B. auf die folgenden Art funktioniert:

   ```
   rb = df[(df['label'] == 'rot') | (df['label'] == 'blau')]
   ```

   Danach ist `rb` ein neuer DataFrame, der nur noch die Daten der roten und blauen Kugeln enthält.

3. Schreiben Sie eine Funktion `anzahl(dframe)`, die für einen DataFrame zählt, wie oft der Wert `rot` bzw. `blau` in der Spalte `label` vorkommt. Die Funktion soll als Ergebnis ein Tupel der Art

   `(anzahlRot, anzahlBlau)`

   zurückliefern.

4. Teilen Sie den DataFrame `rb` in zwei Hälften, in dem Sie nach der Spalte `R`, `G` oder `B` filtern und dabei selbst einen Schwellwert aussuchen. Berechnen Sie dann für jede der Hälften die Anzahlen von roten und blauen Kugeln. Also z.B.:

   ```
   kleiner = rb[ rb['G'] < 50 ]
   groesser = rb[ rb['G'] >= 50 ]

   anzahl(kleiner)
   anzahl(groesser)
   ```

   Suchen Sie das beste Attribute aus `R`, `G`, `B` und den zugehörigen Schwellwert, der für Sie die beste Teilung ergibt.

5. Erstellen Sie einen DataFrame `X` aus dem DataFrame `rb`, der nur die Spalten `R`, `G` und `B` enthält. Speichern Sie zudem in der Variablen `y` die Spalte `label` aus dem DataFrame `rb`.

   Damit haben Sie den Schritt (1) aus dem Foliensatz (Folie 47) erledigt. Gehen Sie mit diesen Variablen `X` und `y` entsprechend der Schritte (2), (3) und (4) vor. Wie gut kann ihr Modell die Kugeln unterscheiden?

   Hinweis: Im Modul `sklearn.tree` gibt es die Funktion `plot_tree`, mit der Sie sich leicht den entstandenen Baum anzeigen lassen können. Dort sind allerdings nicht die Attribut-Namen sondern die Spalten-Indizes der Attribute dargestellt.