Challenge - Data Science

Dieses Übungsblatt beinhaltet Aufgaben zu dem Kurs *Data Science Python*, die euer Wissen zu den Kursinhalten basierend auf dem bereit gestellten Datensatz mit dem Wein. Bitte bearbeitet diese Aufgaben eigenständig und schickt eure Lösung, d.h. der dokumentierte Code mit den Antworten auf die entsprechenden Fragen, bis zum 18.10. 23:59 Uhr an pfeuffer@wiwi.uni-frankfurt.de mit dem Betreff [**DataSciencePython**]. Das Abschicken eurer Lösung ist Voraussetzung für den Erhalt eines Zertifikats für die Teilnahme an o.g. Kurs. Bitte verwendet folgende Python Bibliotheken:

- os
- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn
- keras

Viel Erfolg und Spaß bei den Übungen!

1. Weine in der Vorbereitung

1.1 Importiere die Daten mit Hilfe von der *pandas* Biblothek, so dass am Ende ein Data Frame Objekt mit dem Variablenname *wine* steht.

1.2 Führe das Pre-Processing durch, indem du das DataFrame auf NaN (Not-a-Nubmber) Values prüft. Sind NaN enthalten?

1.3 Lasse dir einige deskriptive Kennzahlen über die Daten ausgeben. Welcher DataFrame Befehl ist hier für sinnvoll. Wie sehen die Kennzahlen aus?

2. Klarer Blick trotz Wein?

2.1 Verschaffe dir einen Überblick über den Datensatz, indem du die Daten visualisiert. Nimm dir etwas Zeit und erstelle ein *pairplot* für dein Data Frame. Was für Erkenntnisse lassen sich aus der Darstellung gewinnen?

2.2 Analysiere den Zusammenhang zwischen *quality* und *alcohol* indem du ein *Boxplot* erstellst. Dabei soll *quality* auf der x-Achse stehen. Was können wir aus dem *Boxplot* ablesen? Gibt einen Trend hinsichtlich des Medians?

2.3 Erstelle eine Abbildung für die Verteilung der Variable *fixed acidity* mit dem *sns.distplot()*. Was lässt sich aus der Abbildung schließen?

2.4 Wie sehen die Korrelation der Variablen aus? Erstelle hierzu eine *Heatmap* die auf die Spiegeldiagonale verzichtet. Passe die Standard-Heatmap an, indem du ihr eine neue Farbe gibt, zusätzliche Abstände einfügst und die Größe auf 9,6 anpasst. Welche drei Variablenpaare weise eine relativ hohe Korrelation in dem Datensatz auf?

3. Wein im Machine Learning - Pre-Processing

3.1 Teile den Datensatz in guten und schlechten Wein ein. Nutze hierzu die cut Methode des DataFrames.

3.2 Erstelle einen Seaborn countplot für die Werte von der spalte *quality*.

3.3 Erkläre die Aufteilung des train- und testsets. Warum haben wir durch die test_size getrennt? Bitte erkläre , warum wir ein Train und ein Testset benötigen.

4. Wein im Machine Learning - Modelling and Predicting

4.1 Bitte konfiguriere den Regression Tree wie in der Vorlesung beschrieben. Du solltest auch beschreiben, warum du bestimmte Parameter wie max_depth gesetzt hast und wie diese Parameter wirkten.

4.2 Was war das Ergebnis Deines Regressionsbaums? Könnte es die Qualität der Weine effizient vorhersagen? Wenn nicht, was könnte das Problem sein?

4.3 Was war das Ergebnis Deines Zufallswaldklassifizierers? Könnte es die Qualität der Weine effizient vorhersagen? Wenn nicht, was könnte das Problem sein? Wenn es besser war als der Regressionsbaum, erkläre, warum dies der Fall ist.

4.4 Diese neuronale Netzstruktur ist gegeben. Bitte fülle die fehlenden Parameter im Modell aus. Bitte erläutere den Grund und die Wirkung der Parameter.

4.5 Bitte füllen Sie die fehlenden Trainingsparameter aus. Welche Anzahl hast du für die Losgröße festgelegt? Warum? Was war der Effekt, als du ihn geändert hast?

4.6 Was macht die Anzahl der Epochen? Warum sollte jemand höhere Werte dieser Parameters verwenden?

4.7 Was war das Ergebnis Deines neuronalen Netzwerkmodells?

4.8 Bitte beschreibe die Diskrepanz deiner Test-, Validierungs und Test Accuracy.