Quiz – Introduction to Data Science

Dieses Übungsblatt beinhaltet Aufgaben zu dem Kurs *Introduction to Data Science*, die euer Wissen zu den Kursinhalten prüfen. Bitte verwendet für die Beantwortung der Aufgaben den bereit gestellten Datensatz *winequality-red.csv*. Bitte bearbeitet diese Aufgaben eigenständig und schickt eure Lösung, d.h. den dokumentierten Code mit Verweis auf die entsprechenden Fragen, bis zum **31. Januar 2023 23:59** Uhr an **cestonaro@wiwi.uni-frankfurt.de** mit dem Betreff „Lösungen zum Data-Science-Quiz“. Das Abschicken eurer Lösung ist Voraussetzung für den Erhalt eines Zertifikats für die Teilnahme am oben genannten Kurs.

Zur Bearbeitung der Fragen verwendet bitte 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 der *pandas* Bibliothek, so dass am Ende ein Data Frame Objekt als Variable in deiner Entwicklungsumgebung vorliegt.

1.2. Führe das Pre-Processing durch, indem du den DataFrame auf NaN (Not-a-Number) Values prüfst. Sind NaN-Werte enthalten?

1.3. Lasse dir einige deskriptive Kennzahlen über die Daten ausgeben. Welcher DataFrame-Befehl ist an dieser Stelle 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 DataFrame. 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 es einen Trend hinsichtlich des Medians?

2.3 Erstelle eine Abbildung für die Verteilung der Variable *fixed acidity* mit *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 gibst, zusätzliche Abstände einfügst und die Breite auf 9 Inches und die Höhe auf 6 Inches anpasst. Welche drei Variablenpaare weisen auf eine hohe Korrelation in dem Datensatz hin?

3. Wein im Machine Learning - Pre-Processing

3.1. Teile den Datensatz in guten und schlechten Wein ein. Wir nehmen an, dass guter Wein bei einer Wertung von 7 beginnt. Nutze zur Unterteilung 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 mit Hilfe der test_size getrennt? Bitte erkläre, warum wir ein Train- und ein Testset verwenden

4. Alter Wein in neuen Schläuchen: Machine Learning - Modelling and Predicting

Ziel ist es nun den Alkoholgehalt der Weine zu prognostizieren. Dazu solltest du den Datensatz zunächst in Trainings- und Testset unterteilen. Anschließend verwenden wir verschiedene Machine-Learning-Modelle zur Vorhersage des Alkoholgehalts eines Weines basierend auf verschiedenen Features.

4.1. Bitte konfiguriere einen *Regression Tree* wie in der Vorlesung beschrieben. Du solltest auch beschreiben, warum du bestimmte Parameter wie *max_depth* gesetzt hast und welche Auswirkungen diese Parameter haben.

4.2. Was war das Ergebnis deines *Regression Trees*? Konnte er den Alkoholgehalt der Weine effizient vorhersagen? Wenn nicht, was könnte das Problem sein?

4.3. Nutze den bereits definierten *Random Forest Classifier* zur Vorhersage des Alkoholgehalts. Was war das Ergebnis deines *Random Forest Classifiers*? Konnte er die Qualität der Weine effizient vorhersagen? Wenn nicht, was könnte das Problem sein?

4.4. Die Struktur eines Neuronalen Netzes ist gegeben. Bitte fülle die fehlenden Parameter bzgl. der Modellarchitektur aus. Bitte erläutere den Grund für die Parameterwahl und deren Auswirkungen auf das Modell.

4.5. Bitte fülle die fehlenden Trainingsparameter aus. Welche Anzahl hast du für die *batch size* festgelegt? Warum? Was war der Effekt, als du diesen Parameter geändert hast?

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

4.7. Was war das Ergebnis deines Neural Network Modells?

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