

## 2. Praktikumsaufgabe IVDA

- a) Lesen Sie den Datensatz „wein.csv“ ein und visualisieren Sie die Verteilungen der Attribute. Verwenden Sie dabei verschiedene Darstellungsformen wie zum Beispiel Histogramme oder Boxplots und bewerten Sie die Vor- und Nachteile der Darstellungen.
- b) Berechnen Sie die paarweisen linearen Regressionen der Attribute und stellen Sie diese visuell dar. Geben Sie zudem die Qualitätswerte der Regressionen aus.
- c) Clustern Sie die Daten mittels k-means. Wählen Sie eine geeignete Distanzmetrik für die Daten und transformieren Sie die Daten, falls notwendig. Berechnen Sie k-means Clusterings mit 2, 3, 4 und 5 Clustern und evaluieren Sie mittels des Silhouette Coefficient, welche Clusteranzahl am Besten für diese Daten geeignet sind. Stellen Sie hierfür die Silhouette Coefficients der verschiedenen Läufe visuell dar.
- d) Stellen Sie die Ergebnisse des Clusterings visuell als Scatterplot dar. Nutzen Sie hierfür die Dimensionsreduktionsmethode PCA.

Die Abgabe der Praktikumsaufgabe ist bis zum 13. Dezember 2019. Vereinbaren Sie hierfür rechtzeitig einen Termin mit Ihrem Betreuer.

Die Gruppen A, B, C, D und E werden von Yves Annanias ([annanias@informatik.uni-leipzig.de](mailto:annanias@informatik.uni-leipzig.de)) betreut.

Die Gruppen F, G, H, I und K werden von Dr. Daniel Wiegrefte ([daniel@informatik.uni-leipzig.de](mailto:daniel@informatik.uni-leipzig.de)) betreut.

Die Gruppen L, M, O, und P werden von Dr. Dirk Zeckzer ([zeckzer@informatik.uni-leipzig.de](mailto:zeckzer@informatik.uni-leipzig.de)) betreut.