**Dokumentation zur Shiny App: Weinqualität (von D. Gutmann, V. Güttinger, J. Hänle)**

1. Einleitung

Im Rahmen des Moduls „Business Analytics: Anwendungsentwicklung“ soll eine interaktive Webapplikation mittels R Shiny entwickelt werden. Hierfür wurde der Kaggle Datensatz „Wine Quality“ herangezogen (Parmar 2018). Zum öffnen der App muss die Datei [shiny\_app\_wine.R](https://d.docs.live.net/9908e2104864295e/Dokumente/Studium/Data%20Science%20%5e0%20Business%20Analytics/Term%206/Anwendungsentwicklung/Prüfungsleistung/Repository/anwendungsentwicklung/shiny_app_wine.R) in R ausgeführt werden, [Wine\_Quality\_Datenbereinigng](https://d.docs.live.net/9908e2104864295e/Dokumente/Studium/Data%20Science%20%5e0%20Business%20Analytics/Term%206/Anwendungsentwicklung/Prüfungsleistung/Repository/anwendungsentwicklung/Wine_Quality_Datenbereinigung.R) zeigt die Bereinigung und Untersuchung der Daten. Alle Dateien sollten sich im selben Arbeitsverzeichnis befinden. Relevante Pakete (*shiny, ggplot2, shinywidgets, rstudioapi*) sind im Skript aufgeführt und müssen gegebenenfalls zuvor installiert werden.

1. Datensatzbeschreibung und –aufbereitung

Der verwendete Datensatz zeigt die Ergebnisse verschiedener physikochemischen Testverfahren zu portugiesischen Rot- und Weißweinen (siehe Tabelle 1) sowie das Qualitätsrating (anhand sensorischer Tests). Der Datensatz besteht aus 6497 Zeilen (1599 Rot- und 4898 Weißweine) sowie 13 Spalten, und ist u.a. prädestiniert für Klassifikationsaufgaben im Bereich des maschinellen Lernens (Cortez et al. 2009).

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute1** | **Datentyp** | **Beschreibung** | **Einheit** |
| type | nominal (factor) | Weintyp: Rot- oder Weißwein | red, white |
| fixed acidity | intervall (float) | Nicht flüchtige Säuren (z.B. Weinsäure) | g/l |
| volatile acidity | intervall (float) | Verflüchtigende Säure | g/l |
| citric acid | intervall (float) | Zitronensäure | g/l |
| residual sugar | intervall (float) | Verbleibender Restzucker (nach der Fermentation) | g/l |
| chlorides | intervall (float) | Chloridgehalt (z.B. Salze) | g/l |
| free sulfur dioxide | verhältnis (float) | Ungebundenes Schwefeldioxid | mg/l |
| total sulfur dioxide | verhältnis (float) | Gesamtschwefeldioxid (Stabilität & Haltbarkeit) | mg/l |
| density | verhältnis (float) | Dichte (Masse pro Volumeneinheit) | g/cm³ |
| pH | intervall (float) | Säurestärke (im Bereich von 3 bis 4) | pH |
| sulphates | verhältnis (float) | Sulfate (Salze oder Ester der Schwefelsäure) | g/l |
| alcohol | verhältnis (float) | Alkoholgehalt in Volumenprozent | % |
| quality2 | odinal (integer) | Bewertung (durch Tester) | Q∈ {1,..7} |

Tabelle 1: Attribute und Datentypen des Datensatzes „Wine Quality“. 1Aus Gründen der Benutzerfreundlichkeit wurden die Spaltennamen angepasst. 2Ursprünglich Werte von 3 bis 9

Ursprünglich enthielt der Datensatz 38 fehlende Werte. Während der Datenbereinigung wurden diese durch den Mittelwert der jeweiligen Spalte ersetzt (siehe [winequality\_cleaned.csv](https://d.docs.live.net/9908e2104864295e/Dokumente/Studium/Data%20Science%20%5e0%20Business%20Analytics/Term%206/Anwendungsentwicklung/Prüfungsleistung/Repository/anwendungsentwicklung/winequality_cleaned.csv)). „Mean Imputation“ hat sich in der Praxis aufgrund der Einfachheit und geringen Verfälschung als gängiges Verfahren zur Handhabung fehlender Werte etabliert. Betrachtet man die Verteilung (siehe hierzu auch QQ-Plots.png und Hist.png) der numerischen Daten fällt auf, dass bei den Variablen *Chlorid, fixierte Säure, flüchtige Säure, Restzucker, Sulfate* sowie *freies Schwefeldioxid* eine starke Schiefe vorliegt, Symmetrie liegt einzig bei *Gesamtschwefeldioxid* und *Qualität* vor. Insgesamt zeigen viele der Variablen eine deutliche Abweichung von der Normalverteilung. Dichte und pH-Wert scheinen annähernd normalverteilt. Diese Tendenzen werden auch bei den Visualisierungen in der Anwendung ersichtlich.

1. Aufbau der Anwendung
2. Datenanalyse

Aufgrund der bereits thematisierten Datenverteilung wurde bei der Untersuchung der zentralen Tendenzen auf den Mann-Whitney-U-Test bzw. Wilcoxon-Test zurückgegriffen. Dieser setzt – im Gegensatz zum T-Test - keine normalverteilte Daten voraus (McKnight und Najab 2010). Die gängigen Lageparameter und Metriken der einzelnen Variablen wie Mittelwert, Standardabweichung, Minimum bzw. Maximum und Median können der Shiny App entnommen werden.

* Eigensetzte statistische Verfahren/Metriken
* Berichten relevanter Ergebnisse

1. Fazit

* Mögliche Erweiterungen / Limitationen

Literaturverzeichnis

Cortez, Paulo; Teixeira, Juliana; Cerdeira, António; Almeida, Fernando; Matos, Telmo; Reis, José (2009): Using Data Mining for Wine Quality Assessment. In: João Gama, Vítor Santos Costa, Alípio Mário Jorge und Pavel B. Brazdil (Hg.): Discovery Science, Bd. 5808. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Computer Science), S. 66–79.

McKnight, Patrick E.; Najab, Julius (2010): Mann‐Whitney U Test. In: Irving B. Weiner und W. Edward Craighead (Hg.): The Corsini encyclopedia of psychology. 4th ed. Hoboken, NJ: Wiley, S. 1.

Parmar, Raj (2018): Wine Quality. Online verfügbar unter https://www.kaggle.com/datasets/rajyellow46/wine-quality/data, zuletzt aktualisiert am 09.07.2018, zuletzt geprüft am 28.07.2024.