Präsentation und Abgabe im Rahmen von BTX8332

Ziel ist es, die vorgegebenen Daten mit Hilfe von R so zu analysieren, dass diese durch Graphiken, deskriptive Statistiken, statistische Tests und Machine-Learning-Verfahren zu verständlichen Informationen verdichtet werden. Sie haben 22 Stunden je Studierende. Dabei sollten ca. 2 Stunden Recherche und 20 Stunden Programmierung und Präsentationserstellung sein. Folgende Kriterien spielen bei der Bewertung eine Rolle:

- Es gibt eine Einführung in die Daten, samt Kontext und was man mit ihnen anfangen kann. Dazu ist vor allem die Recherche gedacht.
- Richtigkeit: Ihre Ergebnisse sind valide und nachvollziehbar (sie legen jeweils kurz dar, warum Sie sich für eine gewisse Vorgehensweise entschieden haben)
- Data Cleaning: Sie wenden alle von uns in der Vorlesung diskutierten Verfahren sinnvoll an: (i) Behandlung fehlender Werte und Ausreißer, (ii) Erkennung fehlerhafter Daten, (iii) Typkonvertierung, (iv) Standardisierung von Strings und Werten im Allgemeinen.
- Feature Engineering: Sie wenden alle von uns in der Vorlesung diskutierten Verfahren sinnvoll an: (i) Kodierung von kategoriellen Variablen (Coding-Schema und Kontraste), (ii) Skalierung von Variablenwerten, (iii) PCA für eine erste Einschätzung, (iv) Feature Selection und (v) Ausbalancierung kategorielle Zielvariablen mit Under/Over-Sampling.
- Clustering: Sie wenden alle von uns in der Vorlesung diskutierten Verfahren sinnvoll an: (i) hierarchisches Clustering und (ii) k-Means. Sie beschreiben, welchen Mehrwert das Clustering für Ihre Daten liefert. Bewerten Sie das Clustering auch quantitativ.
- Machine Learning: Sie wenden alle von uns in der Vorlesung diskutierten Verfahren (für eine sehr gute Note, auch darüberhinausgehende Methoden wie Random-Forests oder LASSO) sinnvoll an.
- Qualität der Ergebnisse: Die Deskriptionen und Visualisierungen sind aussagekräftig. Sie dienen dazu, schnell Übersicht zu gewinnen relevante Zusammenhänge zu entdecken.
- Reflexion: Wurden die Ergebnisse reflektiert und Verbesserungsmöglichkeiten identifiziert? Welche der erworbenen Fähigkeiten können eventuell nicht oder kaum in anderen Projekten eingesetzt werden?
- Für die Visualisierungen nutzen Sie Shiny. Für den Vortrag können Sie MS Powerpoint oder R Markdown nutzen.

Es ist wichtig zu beachten, dass diese Kriterien nicht mit der Nutzung von ChatGPT erfüllt werden können. Vielmehr erfordert diese Aufgabe ein Verständnis von statistischen Methoden, R-Programmierung und Datenanalysekompetenz, die nur durch praktische Arbeit und Erfahrung erworben werden können.

Abgabe: Shiny-App samt R-Code und aussagekräftige Vortragsfolien.