

**Hinweise zur Prüfungsleistung im Wintersemester 2021/2022**

- Abgabetermin ist am 24.01.2022
  - o Dokumentation nach CRISP-DM in Form einer HTML-File
  - o Präsentation (bevorzugt PowerPoint) als PDF-File
  - o Präsentationstermin steht noch aus
- Das Projekt ist mit Jupyter-Notebook zu bearbeiten, somit ist auch eine Abgabe als HTML-File sichergestellt

Es werden nur Abgaben als HTML-File (Dokumentation) und PDF-File (Präsentation) bewertet.

**MUSS-Inhalte der Prüfungsleistung**

- **Business Understanding**
  - o Projektbeschreibung
  - o Data Dictionary
  - o Sonstige Anmerkungen bzgl. Des Projektes
- **Data Understanding**
  - o Deskriptive Analyse
  - o Die in der Vorlesung behandelten Diagrammartentypen + unterschiedliche Zusammenhänge der Variablen
  - o **Pearson vs. Spearman (weshalb verwenden Sie welche Art)**
  - o Erläuterung der Auffälligkeiten, auch Auffälligkeiten die keine Auffälligkeiten sind
  - o Beachten Sie hierbei auch Folie 14 bis 38
- **Data Preparation**
  - o Beheben von Anomalien unter Berücksichtigung der **3-W Regel**
  - o Behandlung unterschiedlicher Methoden zum Umgang von Missing Values
    - Anwendung + **Vergleich behandelter Methoden**
    - **1 zusätzliche Methode ihrer Wahl (z.B. RandomForest, KNN, MissForest ...)**
    - Falls Sie Interesse an einer Imputation mittels eines neuronalen Netzes haben, kann ich GNN empfehlen.
  - o Undersampling + Oversampling
    - Vergleich beider Verfahren anhand des Datensatzes
    - Auf **welches Verfahren fällt ihre Entscheidung?**
  - o Feature Engineering (Selection + Importance)
    - Korrelation
    - Logistische Regression *oder* PCA
  - o Beachten Sie hierbei auch Folie 33 bis 59
- **Modeling**
  - o Decision Tree (RandomForest)
  - o **Neuronales Netz**
  - o **1 zusätzliche Methode** ihrer Wahl (z.B. Gradient Boosting)
  - o Beachten Sie hierzu die Folien aus der Veranstaltung
- **Evaluation**
  - o Vergleichen Sie die Modelle anhand der Gütemaße
  - o Entscheiden Sie sich für das beste Modell und fahren Sie anschließend mit der Evaluation für das gewählte Modell fort
  - o Korrektklassifikations- und Fehlerrate
  - o Konfusionsmatrix

- True Positive Rate, False Positive Rate, True Negative Rate, False Negative Rate
- Precision und Recall
- AUC und ROC
- Beachten Sie hierbei die Folien aus der Veranstaltung
- Hierbei liegt der Fokus auf der Interpretation – Bitte erörtern Sie ihr Vorgehen + Ergebnisse
- **Deployment**
  - Spiegelt die Präsentation wider

***Bitte beachten Sie den Hinweis aus der Veranstaltung zum Thema Data Leakage und der damit verbundene Train bzw. Testsplitt an der richtigen Stelle!***