Prüfungsleistung für „Clustering & Classification“

- Für einen selbst gewählten Datensatz ist eine **Clustering- oder Klassifikationsanalyse** in Form einer **Studienarbeit** (12 Seiten) zu bearbeiten. Um eine eigenständige Bearbeitung prüfen zu können, wird die Arbeit in einem kurzen Vortrag (10 Minuten) mündlich erläutert.
- Die Arbeit besteht aus einem **Theorieteil**, in dem die verwendete Methodik vorgestellt wird und aus einem **Praxisteil**, in dem die Methodik auf ein Praxisbeispiel angewendet wird.
- Anforderungen Theorieteil:
    - Die Theorie der Methodik muss verständlich vermittelt werden (inkl. Quellen)
- Anforderungen Praxisteil:
    - Verwendung von Python oder R (in der Vorlesung verwenden wir Python)
    - Die Zielsetzung muss klar definiert sein.
    - Die Analyse soll entlang des CRISP-DM Modells oder eines ähnlichen Vorgehens durchgeführt werden.
    - Der Datensatz muss beschrieben werden
        - Merkmale mit Bedeutungen
        - Label (falls vorhanden)
    - Die Quelle des Datensatzes muss angegeben werden.
    - Im Rahmen des Datenverständnisses soll eine explorative Datenanalyse gemacht werden. Idealerweise durch Plots von Variablen und Datentransformationen (z.B. Gruppierte Tabellen, Top-10, etc.). Diese Analyse kann auch auf aufgestellten Thesen beruhen, die dadurch belegt oder widerlegt werden. (Mindestanforderung: zwei Aspekte der Daten darstellen)
    - Es müssen mindestens zwei Methoden verwendet werden (d.h. zwei Clusteransätze oder zwei Klassifikationsansätze). Diese müssen miteinander verglichen werden.
    - Die Evaluation der Methoden muss geeignete Qualitätsmaße verwenden.
- Aufwertung durch:
    - Eigens recherchierte, nicht in der Vorlesung verwendete Methoden
    - Optimierungsversuche z.B. Ausprobieren von Hyperparametern durch Gittersuche oder Ähnliches
    - Aufwändige Datenvorverarbeitung z.B. durch Text-Mining oder Zusammenführung von mehreren Datenquellen
    - …

Abgabe als PDF-Datei; zusätzliche Abgabe eines jupyter Notebooks. Wir besprechen in der Vorlesung Möglichkeiten, diese Dateien zusammenzuführen.

Zur mündlichen Vorstellung wird das PDF/Notebook verwendet – keine separaten Folien erforderlich.

Wo finde ich Daten?

Allgemein

- Google https://datasetsearch.research.google.com/
- https://www.google.com/publicdata
- https://trends.google.com/
- Kaggle https://www.kaggle.com/datasets
- Our World in Data https://ourworldindata.org/
- Amazon https://aws.amazon.com/de/data-exchange/
- https://registry.opendata.aws/

Klima- und Geodaten

- Deutscher Wetterdienst https://www.dwd.de/DE/klimaumwelt/cdc/cdc_node.html
- Klimadaten weltweit https://www.ncei.noaa.gov/access/search/dataset-search
- https://meteostat.net/de/
- Earthdata https://www.earthdata.nasa.gov/learn/find-data

Metaseiten (Listen von Datenarchiven)

- Data Portals https://dataportals.org/
- OpenDataMonitor https://opendatamonitor.eu/
- Qandl https://demo.quandl.com/

Andere Seiten, die beliebte offene Datenarchive auflisten

- Wikipedia https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Gesundheitsdaten

- WHO https://www.who.int/data/collections
- EU Covid-19 https://www.ecdc.europa.eu/en/covid-19/data

Meinungsforschung

- World Values Survey https://www.worldvaluessurvey.org/
- Pew Research Center https://www.pewresearch.org/internet/datasets/

Regierungsstatistiken

- Statistisches Bundesamt https://www-genesis.destatis.de/genesis/online
- U.S. Government data.gov, Singapur https://data.gov.sg/