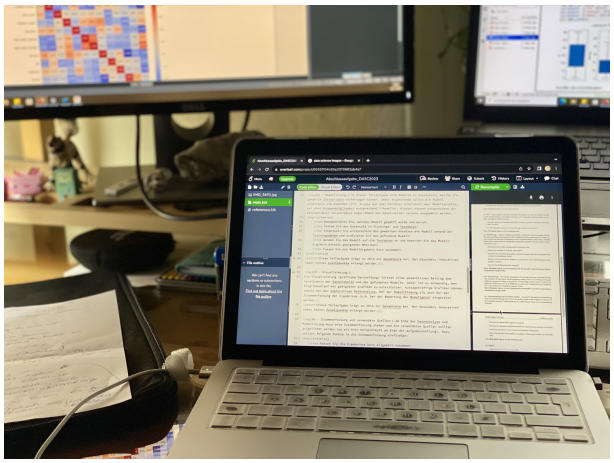
Modul Data Science - From data to knowledge

- Abschlussaufgabe -

BA Sachsen - Staatliche Studienakademie Leipzig

Abgabe bis zum: 22.12.2024

A0 - Allgemeines vorab: Die Aufgabe ist in Gruppenarbeit zu erledigen, die Zusammensetzung der Gruppen 1-4 ist bereits festgelegt (siehe Gruppen_DASC_11-2024.pdf). Die Prozessierung ist ausführlich in einem Python-Notebook zu dokumentieren, wobei hierbei auch deutlich an den Teilaufgaben zu kennzeichnen ist, wer zu dieser Aufgabe beigetragen hat. Für die einzelnen Teil-Arbeitsschritte gibt es klare Bewertungskriterien (siehe Beschreibung der Aufgaben A2 - A6). Verwendete Quellen sind am Ende des Skripts deutlich zu benennen (siehe Beispiele wie dieses verwendete Buch [2] oder eine Webseite [1]). Aus den Qualitätskriterien der Teilaufgaben ergibt sich die Notenbildung. Die Abgabe muss bis zum **24.03.2024** erfolgen (siehe Punkt A7).



A1 - Auswahl Datensatz: Es stehen 3 Datensätze zur Auswahl, von denen Sie sich **einen** auswählen! Mit diesem Datensatz führen Sie alle Bearbeitungsschritte durch und dokumentieren diese sowie die entsprechenden Ergebnisse in einem Python-Notebook. Die Datensätze sind in diesem OPAL-Ordner für diese Abschlussaufgabe abgelegt. In den entsprechenden Unterordnern finden Sie neben den eigentlichen Datensätzen auch alle zusätzlichen Informationen / Metadaten für die Datensätze.

- **Datensatz 1: Titanic - Überleben einer Katastrophe.** Hier haben Sie die Passagierliste der Titanic in den Händen. Die Fragestellung, die daran geknüpft ist: Können Sie vorhersagen, welche Passagiere im Testdatensatz die Katastrophe überleben und welche nicht? Welche Passagiermerkmale haben den größten Einfluss auf die Überlebenschancen?
- **Datensatz 2: Kosten der Gesundheitsversorgung.** Dieser synthetische Datensatz wurde erstellt, um reale Daten aus dem Gesundheitswesen nachzuahmen und es Benutzern zu ermöglichen, ihre

Fähigkeiten zur Datenmanipulation und -analyse im Kontext der Gesundheitsbranche zu üben, zu entwickeln und zu präsentieren. Mit welcher Genauigkeit lassen sich aus charakteristischen Patientendaten die Kosten der Behandlung vorhersagen? Welche Merkmale sind die stärksten Einflussfaktoren?

- **Datensatz 3: Hauspreise in New York.** In diesem Datensatz sind Informationen über die Preise von über 4000 Immobilien in New York (USA) enthalten und bietet wertvolle Einblicke in den Immobilienmarkt der Region. Er enthält Informationen wie Maklertitel, Haustypen, Preise, Anzahl der Schlafzimmer und Badezimmer, Quadratmeterzahl der Immobilie, Adressen, Bundesstaat, Verwaltungs- und Ortsgebiete, Straßennamen und geografische Koordinaten. Mit welcher Genauigkeit lassen sich aus charakteristischen Merkmalen die Hauspreise vorhersagen? Welche dieser Merkmale sind die stärksten Einflussfaktoren?

A2 - Skript- und Datenvorbereitung: Für die Dokumentation ist ein Python Notebook anzulegen, welches alle Teilaufgaben sowie die Ergebnisse umfassend dokumentiert. Im ersten Teil des Skripts sind folgende Punkte zu bearbeiten:

- Einbindung der grundlegenden Python-Bibliotheken
- Datensatz einlesen und Überblick verschaffen (Größe, Attribute, Datentypen, etc.)
- Metadaten-Informationen zusammenstellen (Quelle, Zweck der Erhebung, etc.)
- Datenqualität beurteilen (fehlende Daten, Redundanz, etc.)
- Bereinigung des Datensatzes (Filtern, Schreibfehler, Inkonsistenzen, Zusammenfassung von Attributen, Umwandlung von Datentypen, etc.)
- Festlegung der abhängigen Zielvariable inklusive der Formulierung der Fragestellung

Diese Teilaufgabe trägt zu 10% zur Gesamtnote bei.

A3 - EDA: Dieser Abschnitt umfasst die explorative Datenanalyse. Hier sollte der Datensatz erkundet werden. Die Informationen sollten grafisch aufbereitet und in Stichpunkten zusammengefasst werden. Dabei sind folgende Punkte zu beachten:

- univariate Analyse (z.B. statistische Kennzahlen, Verteilung der numerischen Daten, Verteilung in den kategorialen Variablen)
- bi- und multivariate Analyse (z.B. Zusammenhänge zwischen den Variablen)
- Ist auf der Basis der Analysen eine weitere Datenbereinigung notwendig? Dokumentieren Sie das, falls zutreffend!

Diese Teilaufgabe trägt zu 20% zur Gesamtnote bei.

A4 - Modellierung: In dieser Teilaufgabe sind Modelle zu entwickeln, welche die gewählte Zielvariable vorhersagen können. Jeder Studierende muss ein Modell entwickeln und anwenden (d.h. eine Gruppe aus zwei Personen entwickelt zwei Modellansätze. Bitte geben Sie an, wer welches Modell entwickelt hat.). Hierbei können entsprechend der Datenstruktur verschiedene Algorithmen des maschinellen Lernens eingesetzt werden.

- Dokumentieren Sie, welches Modell gewählt wurde und warum!
- Teilen Sie den Datensatz in Trainings- und Testdaten!
- Entwickeln Sie entsprechend des gewählten Ansatzes ein Modell anhand der Trainingsdaten und evaluieren Sie das gefundene Modell!
- Wenden Sie das Modell auf die Testdaten an und bewerten Sie das Modell-Ergebnis mittels geeigneter Metriken!
- Fassen Sie das Modellergebnis kurz zusammen!

Diese Teilaufgabe trägt zu 40% zur Gesamtnote bei. Bei besonders innovativen Ideen können Zusatzpunkte erlangt werden.

A5 - Visualisierung: Die Visualisierung (grafische Darstellung) leistet einen wesentlichen Beitrag zum Verständnis der Dateninhalte und des gefundenen Modells. Daher ist es notwendig, den Programmablauf mit geeigneten Grafiken zu unterstützen. Aussagekräftige Grafiken können sowohl bei der explorativen Datenanalyse, bei der Modellfindung als auch bei der Zusammenfassung der Ergebnisse (z.B. bei der Bewertung der Modellgüte) eingesetzt werden.
Diese Teilaufgabe trägt zu 20% zur Gesamtnote bei. Bei besonders innovativen Ideen können Zusatzpunkte erlangt werden.

A6 - Zusammenfassung und verwendete Quellen: Am Ende der Datenanalyse und Modellierung muss eine Zusammenfassung stehen und die verwendeten Quellen sollten aufgelistet werden (so wie hier beispielhaft am Ende der Aufgabenstellung). Dazu sollten folgende Punkte in die Zusammenfassung einfließen:

- Fassen Sie die Ergebnisse kurz allgemein zusammen!
- Was leisten die gefundenen Modelle hinsichtlich der Untersuchung der gewählten Zielvariable?
- Vergleichen und bewerten Sie die verschiedenen genutzten Modellansätze!
- Gibt es Schwachstellen, Fehlereinflüsse und mögliche Verbesserungsvorschläge für die Modellierung?

Diese Teilaufgabe trägt zu 10% zur Gesamtnote bei.

A7- Abgabe: Details dazu entnehmen Sie bitte auch der Email von Frau Prof. Schneider, die Sie bzgl. der Abgabe der Modulleistung Ende November 2024 erhalten haben.

- Aufgabenstellung und Datensätze liegen im BA-OPAL Abschlussaufgabe Modul DASC
- Schreiben Sie sich als ersten Schritt in der Lerngruppe 'Dezember 2024' ein, um entsprechend auf die Unterlagen und die Abgabe-Option zugreifen zu können!
- Der Upload der Zip-Dateien (bestehend aus Py-Notebook, Daten- und Metadaten-File) muss bis zum 22.12.2024 23:59 Uhr (MEZ) erfolgen.

- Die Abgabe erfolgt via Upload in folgendes Repository: BA-OPAL » Abgabe
- Für den Filenamen der Zip-Dateien ist folgende Struktur einzuhalten:
  csXX-Y_Name1_Vorname1_Name2_Vorname2.zip
- **Wichtig: Bitte kennzeichnen Sie an jeder Teilaufgabe im Notebook, mit welchem Anteil die Gruppenteilnehmer an den Teilaufgaben mitgewirkt haben!**

Literatur

[1] Kaggle Inc. kaggle.com. `https://www.kaggle.com/`, 2023. Accessed: 16.11.2023.

[2] P. Gedeck P. Bruce, A. Bruce. *Praktische Statistik für Data Scientists.* dpunkt.verlag GmbH, 2021.