EINFÜHRUNG IN DAS MASCHINELLE LERNEN – L1

EXPLORATORISCHE DATENANALYSE MIT PANDAS

Bekommen in L1

Abgabe in L2

Vergewisse dich, dass du auf deiner Maschine Python, Version mindestens 3.11, eine IDE mit Jupyter Notebooks (z.B. VSCode mit den passenden Extensions), und die Bibliotheken pandas, matplotlib und sklearn, hast.

Optional, aber nutzlich als "good practice": erstelle ein virtuelles Environment für deine Projekte. Dies hilft sicherzustellen, dass alle Abhängigkeiten und Bibliotheken die richtige Version haben. Benutze zum Management der Pakete eine Software wie z.B. poetry. Dies stellt sicher, dass für alle Pakete die Versionen kompatibel sind, sowie dass verschiedene Anwendungen und ihre Abhängigkeiten nicht in Konflikt geraten. Dies ist Standardpraxis für große Projekte in der Industrie. Siehe hier ein Tutorial darüber. Es wird empfohlen, auch ein Linter, z.B. die pylint-Extension aus VSCode zu installieren.

Das sogenannte Ames Housing Dataset wird als Testdatensatz für zahlreiche Data Science Experimente benutzt. Erstelle ein Jupyter Notebook, wo du mithilfe der Pandas-Bibliothek die vorgegebene csv-Datei einliest.

Untersuche den Datenset. Denke an Outliers, fehlerhafte Daten, usw. Zeichne Plots. Denke z.B. an die Verteilung der Werte. Identifiziere numerische, bzw. kategorische Features.

Folgende Funktionen aus Pandas oder Matplotlib könnten nutzlich sein: isnull, isna, value_counts, dropna, describe, head, avg, min, max, groupby, scatter, hist, corr...

Beispiel: wenn man in einer Notebook-zelle housing[housing["Lot Area"] > 40000], dann werden alle Anwesen mit der betreffenden Eigenschaft angezeigt. Man kann auch komplexere Filter anlegen. Pandas hat Funktionalitäten ähnlich dem join, bzw. groupby aus SQL, die Syntax kann aber verschieden sein.

Sei kreativ! Stelle sicher, dass du zusammen mit deinen Kollegen / Kolleginnen den Datensatz und die Ergebnisse der Analyse verstehst und kommunizieren kannst. Als Data Scientist muss man immer den Stakeholdern Ergebnisse und Aktionsvorschläge auf Grund von diesen kommunizieren. Präsentiere die Ergebnisse als Grafiken und / oder Text in Markdown-Zellen in deinem Notebook.

Mit was für Daten von außen könnte man das Dataset bereichern? Welche vorhandenen Features sind nicht nützlich und können entfernt werden?

Bonus: Finde zwei Features, zwischen welche eine lineare Relation existieren könnte. Hinweis: Benutze dabei Scatter Plots. Implementiere eine lineare Regression auf die so gefundenen Features. Siehe hier (sklearn dokumentation). Zeichne die Regressionslinie und die Punkte mit verschiedenen Farben. Vergleiche den tatsächlichen mit dem vorhergesagten Wert für ein Punkt aus dem Datenset.

Hinweis: Die nächsten Laboraufgaben beziehen sich auf demselben Datensatz.