In diesem Projekt werden Sie selbst ein Machine Learning-Modell trainieren und evaluieren.

Kaggle ist eine Plattform, die *Machine Learning Competitions* hostet. Besonders bekannt ist die Plattform für Competitions großer Unternehmen, bei denen man größere Geldbeträge gewinnen kann. Darüber hinaus gibt es allerdings auch eine Vielzahl Competitions ohne Geldgewinne, die dazu dienen, die eigenen Machine Learning-Kompetenzen zu erweitern.

In diesem Projekt werden wir zunächst gemeinsam die Titanic Competition versuchen, anschließend sollen Sie noch an einer weiteren, frei wählbaren Competition teilnehmen.

## Titanic

1. Erstellen Sie einen Account auf https://www.kaggle.com.
2. Bearbeiten Sie folgendes Tutorial: https://www.kaggle.com/code/alexisbcook/titanic-tutorial/notebook.
3. Versuchen Sie, den im Tutorial entwickelten Baseline-Classifier zu schlagen. Folgende Ideen können aussichtsreich sein:
   a) Data Imputation: Gibt es fehlende Werte (NAs) in einigen der Spalten? Um diese Features trotzdem für die Klassifikation nutzen zu können, kann es notwendig sein, die fehlenden Werte zunächst mit plausiblen Werten zu ersetzen. Dieses Tutorial kann Ihnen weiterhelfen: https://www.kaggle.com/code/residentmario/simple-techniques-for-missing-data-imputation/notebook.
   b) Feature Engineering: Oft ist es sinnvoll, nicht nur die Rohdaten als Input des Classifiers zu verwenden, sondern zuvor Features zu berechnen. Dieses Tutorial hilft: https://www.kaggle.com/learn/feature-engineering.
   c) Auswahl des Classifiers und Hyperparameteroptimierung: Je nach Datensatz erreichen unterschiedliche Classifier eine gute Performance. Random Forests sind oft eine gute, aber nicht unbedingt die beste Wahl. Dieses Tutorial liefert mehr Informationen zur Wahl des Classifiers und Hyperparameteroptimierung: https://www.kaggle.com/code/shreayan98c/hyperparameter-tuning-tutorial.

## Eine Weitere Kaggle Competition

Abschließend sollen Sie selbstständig eine selbst gewählte Kaggle Competition bearbeiten. Ihre Ideen und Ergebnisse sollen Sie dokumentieren, und den Report einreichen.

1. Wählen Sie eine beliebige Kaggle Competition aus, die Sie in bearbeiten möchten. Es ist empfehlenswert, eine Klassifikations-Aufgabe zu wählen (anstatt Regression oder unsupervised Learning), aber die Entscheidung liegt bei Ihnen.
2. Entwickeln Sie eine Baseline, d.h. ein Kaggle-Notebook, das eine valide Einreichung erzeugt und einfache Standardmethoden verwendet (z.B. einen Random Forest-Classifier). Reichen Sie das Baseline-Modell bei der Competition ein.

3. Versuchen Sie, die Baseline-Performance zu übertreffen. Feature Engineering, Hyperparameteroptimierung, Classifier-Auswahl sind sinnvolle erste Ansatzpunkte. Reichen Sie das verbesserte Modell bei der Competition ein.

4. Dokumentieren Sie Ihre Arbeit in einem PDF-Dokument. Das Dokument sollte folgende Aspekte beinhalten:
   a) Eine kurze Beschreibung des Datensatzes. Woraus bestehen die Daten? Was ist die Zielstellung? Das Dokument soll einige Plots des Datensatzes enthalten, die relevante / interessante Aspekte des Datensatzes geeignet darstellen.
   b) Eine kurze Beschreibung des Baseline-Modells und ein Link zum Kaggle-Notebook.
   c) Eine Beschreibung Ihrer Ideen und Experimente, um die Performance des Baseline-Modells zu verbessern. Geben Sie für jede Änderung des Baseline-Modells die Performance-Änderung an.
   d) Ein Link zum Kaggle-Notebook, in dem die Verbesserungen implementiert sind.

Das Dokument muss bis zum 14.03., 16:00 Uhr an stefan.luedtke@uni-leipzig.de geschickt werden, mit dem Betreff "Abgabe KI-Projekt <Ihr Name>". **Die fristgerechte Einreichung des Dokuments ist Voraussetzung, um an der Klausur teilnehmen zu können!**