# Homework 2

For this homework you need to choose a dataset from the list below and test various classification methods with the ultimate goal of developing a model that classifies as accurately as possible. To complete this task, you must follow these steps:

1) Exploratory Data Analysis (EDA) – 2p
    a. Describe the structure of the data: number of rows, columns, and data types.
    b. Identify correlations between data
    c. Make plots to show data distributions
    d. Identify target and feature variables
    e. Highlight Potential Issues in the Data
2) Data preprocessing – 2 p
    a. Identify outliers
    b. Deal with missing values
    c. Deal with categorical data
    d. Standardize or normalize numerical features as needed
    e. Use principal component analysis (PCA)
3) Propose and test 2 different methods that deal with data imbalance (e.g. SMOTE, ADASYN, undersampling, etc.) – 1 p
4) Test out 5 different ML methods and print out all performance metrics (Recall, Precision, Accuracy, F1). Test the performance of Voting Classifier which combines predictions from multiple models – 4p
5) Document in a report everything that you experimented, from data preprocessing to different parameters used for the ML models - Homework won't be graded without the report
6) Use k-fold cross validation to prove robustness - 1p

You need to prepare some slides to showcase your work and present them on the 12th of December.

Details about building the report:

Structure of the Report:

Introduction: Problem definition, objectives, and dataset description
EDA: Insights from data exploration, including plots and correlations
Data Preprocessing: Techniques applied and justification of your choice
Handling Imbalance: Methods tested and their impact on results
ML Models: Model descriptions, parameters tested, and performance results
Conclusion: Summary of findings, best-performing model, and future ways of improving

Visualizations and Reproducibility:

Ensure all plots are clear, labeled, and easy to interpret.
Use tables to summarize metrics, feature importance, and preprocessing effects.
Ensure that results are reproducible by others using your methodology.

You need to write your name here in order to choose your dataset:

https://docs.google.com/spreadsheets/d/1vR04LYf7J48uqKT8UM-LtE2ZUX6RtUdfgIKfTX7MCno/edit?usp=sharing

Only 23 students can choose the same dataset

Datasets to choose from:

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data

https://www.kaggle.com/datasets/jahnavipaliwal/mountains-vs-beaches-preference

https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset