# Assignment 1: Data Preprocessing

## Objective

The objective of this assignment is to understand and apply data preprocessing techniques on a dataset. You will learn how to clean, transform, and prepare data for modeling, focusing on various steps involved in data preprocessing such as handling missing values, encoding categorical data, feature scaling, and data splitting.

## Datasets for this task

You are free to choose your datasets (you might need more than one) for this assignment. You have to use the datasets that are uploaded to blackboard.

## Tasks

1. Data Exploration (10)
    a. Explore the dataset by displaying the first few rows, summary statistics, and data types of each column.
    b. Identify missing values, outliers, and unique values in categorical columns.
2. Data Cleaning (20)
    a. Handling Missing Values
    b. Choose appropriate methods to handle missing values (e.g., mean/median imputation for numerical data, mode imputation for categorical data, or deletion of rows/columns).
    c. Justify your choices for handling missing data.
3. Handling Outliers (20)
    a. Detect outliers using methods such as the IQR method or Z-score.
    b. Decide whether to remove, cap, or transform the outliers. Justify your decisions.
4. Data Transformation (30)
    a. Encoding Categorical Data
        i. Apply label encoding or one-hot encoding to transform categorical data into numerical form.
        ii. Justify your choice of encoding method.
    b. Feature Scaling
        i. Apply feature scaling techniques such as normalization (Min-Max scaling) or standardization (Z-score normalization) to the dataset.
        ii. Explain why feature scaling is necessary and how it impacts the model.
5. Data Splitting (10)
    a. Split the preprocessed dataset into training and testing sets. Typically, an 80-20 or 70-30 split is used.
    b. Explain the importance of splitting the data and how it prevents overfitting.

6. Bonus Task (Optional, 10): Apply dimensionality reduction techniques such as Principal Component Analysis (PCA) and discuss how it affects the dataset.

## Submission Requirements

Submit a PDF file (free template) with the results and code (where necessary). Include the code only when it is absolutely necessary and the explanation can not be completed without it.

Be as precise as possible in your answers and justifications. There are no extra points for extra text.

If the justification is missing from any task where it is asked for, 10 marks will be deducted from the task.

**Word limit: 3000. Please be within this limit (the code and references are not counted, but the table and figure captions are counted).**

## Assessment Criteria

1. Correct application of preprocessing techniques.
2. Clarity and justification of decisions made during preprocessing.
3. Quality of documentation and readability.
4. Completeness of the assignment and adherence to instructions.