Assignment 4

Course Name: Advanced DBMS
Course Instructor: Husnain Haider

Real-World Problem Solving Through Data Preprocessing

Objective:
Your task is to identify a real-world problem that can be addressed using data. Then, propose a feasible solution for it, find a relevant dataset, and perform thorough data preprocessing.

Instructions:

1. **Problem & Dataset Selection**
    - Think of a **real-world problem** (e.g., predicting housing prices, detecting spam messages, forecasting weather, etc.) that can be solved using data analysis or machine learning.
    - Find a **relevant public dataset** from sources such as Kaggle, UCI ML Repository, Google Dataset Search, etc.
    - Each group must work on a **unique problem and dataset**. Duplicate topics will not be allowed.

2. **Dataset Preprocessing**
   Using Python (in a `.py` script or Jupyter Notebook), preprocess the dataset by covering the following aspects:
    - **Identify column attributes**:
        - Data types (categorical, numerical, etc.)
        - Discrete or continuous variables
        - Distribution: skewed or symmetric
    - **Detect and handle outliers**
    - **Clean the data**:
        - Handle missing values (filling/removal)
        - Remove noise and inconsistencies
        - Drop or impute NaN values

3. **Submission Guidelines:**

   Your submission must be in the form of a **single compressed folder** named exactly as:
   `Group<YourGroupNumber>_<ShortTopicName>.zip`

Example: `Group05_SpamDetection.zip`

The folder must contain the following files:

- **ProblemStatement.pdf / ProblemStatement.docx**
  A document explaining the problem, its importance, proposed solution, original dataset overview, and comparison between original and preprocessed datasets.
- **Preprocessing.py** or **Preprocessing.ipynb**
  Your Python code for data preprocessing.
- **CleanedDataset.csv**
  The final preprocessed dataset.

Make sure your filenames are exactly as written above to ensure consistency and avoid confusion. Only one member of your group needs to submit the work.

4. **Grading Criteria**

   The assignment will be graded through evaluation. All group members should have equal knowledge about their submitted work for maximum marks.

Deadline for Final Submission: May 11, 2025 (Sunday) - 11:59 PM