Assignment 1: Identifying a Real-World Problem, Data Collection, and Preprocessing

Course: Data Science
Class: BSCS-F22
Instructor: Ghulam Ali
Course Learning Outcome: CLO-03 GA-02

Due Date: Oct 20, 2025 11:30 PM

Objective:

The objective of this assignment is to familiarize students with the initial stages of a data science project—identifying a real-world problem, data collection, and preprocessing. Students will identify a problem they want to solve using data science, collect a relevant dataset, clean it, handle missing values, and document the entire process.

Assignment Tasks:

1. Problem Identification:

- Choose a real-world problem that can be solved using data science.
- Clearly define the problem statement and explain why it is important.
- Describe how data is involved in solving this problem.
- List the key questions (at least 5 questions) that need to be answered using data.

2. Data Collection:

- Identify and collect a dataset related to the chosen problem.
- The dataset can be obtained from online sources (e.g., Kaggle, UCI Machine Learning Repository, government databases) or collected manually.
- Clearly document the source of the dataset and the reason for choosing it.

3. Data Preprocessing:

- Load the dataset into Python (Pandas) or another relevant tool.
- Check for missing values and handle them appropriately (removal, imputation, etc.).
- Detect and remove duplicate records if any.
- Convert categorical data into numerical format if applicable.
- Normalize or scale numerical data if necessary.
- Identify and handle outliers (if applicable).

4. Documentation:

Students must submit a well-structured report covering:

I. **Introduction:** Brief description of the identified problem, its significance, and how data science can help solve it.
II. **Problem Statement:** Clear articulation of the problem and expected outcomes.
III. **Key Questions:** A list of the main questions that will be answered through data analysis.
IV. **Data Collection Process:** Source, methodology, and any challenges faced.
V. **Preprocessing Steps:** A detailed explanation of each preprocessing step along with justifications.
VI. **Code Snippets:** Include relevant code snippets in the report (not the entire script, only important parts).
VII. **Observations and Insights:** Key findings and observations from the preprocessing phase.
VIII. **Conclusion:** Summary of the work done and future possible improvements.

Submission Guidelines:

1. The report should be in PDF format.
2. The code must be submitted separately in a Jupyter Notebook (.ipynb).
3. Proper citations should be included if external sources are used.
4. Submit your assignment on Google Classroom by the due date.