**Department of Computer Engineering  
CO34601: Data Science & Engineering**

**(2024-2025 SEM B)**

**Lab Assignment 01: Exploration and Statistical Analysis of Dataset**

**Objective:**

To understand and apply basic data preprocessing and statistical analysis techniques using Python.

**Instructions:**

1. Select a dataset from UCI Machine Learning Repository or Kaggle. The dataset must have:

* At least 5 attributes.
* Mixed data types (numerical, categorical, etc.).

1. Perform the following tasks in Python using libraries such as pandas, numpy, and statistics.

**Tasks:**

**1. Load and Explore the Dataset**

* Import the dataset into Python and display its structure and summary.
* Identify and separate **input features** and the **class/target labels**.

**2. Analyze the Dataset**

* Determine how many unique classes (categories) are present in the target variable.
* Check if there is any **missing data** in the dataset.
  + If found, handle missing data using appropriate techniques (e.g., mean/mode imputation, or removal).

**3. Perform Statistical Analysis**

For all numerical attributes:

* Calculate and display the **mean**, **median**, **mode**, **variance**, and **standard deviation**. Use the statistics module.

**4. Calculate Distance Metrics**

* Compute the **Euclidean distance, Manhattan distance** between two randomly selected data points from the dataset.

**5. Identify Data Distribution**

* Plot the **distribution of each numerical attribute** (e.g., histograms, box plots).
* Analyze and describe the overall distribution of the dataset.

(Subject Incharge)

**Department of Computer Engineering  
CO34601: Data Science & Engineering**

**(2024-2025 SEM B)**

**Lab Assignment 02:** **Data Preprocessing**

**Objective:**  
To explore and implement various data preprocessing techniques for preparing a dataset for analysis.

**Tasks:**

1. **Data Cleaning:**

* Detect and handle duplicate records.
* Handle outliers using Z-score or IQR (Interquartile Range).

1. **Feature Scaling:**

* Apply normalization (Min-Max Scaling) to numerical attributes.
* Standardize the attributes using Z-score scaling.

1. **Feature Encoding:**

* Convert categorical data into numerical data using techniques such as:
  + - One-hot encoding.
    - Label encoding.

1. **Feature Engineering:**

* Derive new features by combining existing ones.
* Create bins for continuous variables (e.g., age groups).

(Subject Incharge)