Assignment:01

Course: Data Mining

DATASET GUIDELINES

- **Regression/Classification:** Use standard datasets (e.g., *Titanic*, *Boston Housing*).
- **Association Rules:** Use a Transactional dataset (e.g., *Groceries* or *Market Basket Optimization*).

PART 1: DATA PREPROCESSING

Objective: Clean the data to ensure model stability.

1. **Error Detection & Fixing:**
   - Identify columns with missing values (NaNs).
   - Fill missing numerical values with the **Mean** and categorical values with the **Mode**.
2. **Noise/Outlier Detection:**
   - Visualize the distribution of a numerical feature using a **Boxplot**.
   - Programmatically detect outliers using the **Interquartile Range (IQR)** method.

PART 2: SUPERVISED LEARNING

Objective: Predict outcomes and evaluate performance metrics.

1. **Regression Analysis:**
   - **Linear Regression:** Predict a target variable using a single feature.
   - **Multiple Linear Regression:** Predict the same target using all available features.
2. **Classification (Logistic Regression):**
   - Implement **Logistic Regression** for binary classification.
   - *Note:* Focus on tuning this single model rather than implementing multiple classifiers.
3. **Model Evaluation:**
   - Calculate: **Precision**, **Recall**, and **F1-Score**.
   - Plot the **Confusion Matrix** and the **ROC Curve**.

PART 3: ASSOCIATION RULE MINING

Objective: Discover frequent patterns in transactional data.

1. **Apriori Algorithm:**
    - Apply the Apriori algorithm to a transaction dataset.
    - Identify **Frequent Itemsets** with a minimum support of 2 (adjust based on dataset).
    - Generate **Association Rules** using a minimum confidence threshold.
    - *Deliverable:* List the top 5 rules with the highest **Confidence** value.

Submission Requirements:

- Submit a single Python Notebook (`.ipynb`)
- Include brief interpretations of the Association Rules found (e.g., "Customers who buy X are likely to buy Y").