**Name:** Chandrakant Dattatrey Thakare

**Roll no:** 282014  **PRN:** 22310303

**Class :** SY CSE AI**Batch:** B1

**Assignment 2: Data Analysis and Model Building**

**Problem Statement:**

Perform the following operations using **R/Python** on suitable datasets:

a) Compute and display summary statistics for each feature available in the dataset (e.g., minimum value, maximum value, mean, range, standard deviation, variance, and percentiles).  
b) Illustrate the feature distributions using **histograms**.  
c) Perform **data cleaning**, **data integration**, **data transformation**, and **data model building** (e.g., **Classification**).

**Objective**

* Strengthen the ability to analyze datasets through descriptive statistical measures.
* Visualize data distribution to understand feature behavior.
* Learn practical steps in data preprocessing and apply machine learning for classification.

**Tools and Resources**

* **Software Used**: Google Colab / Jupyter Notebook
* **Libraries Used**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

**Key Functions and Methods Used**

* df.describe(include='all'): Display summary statistics for numerical and categorical features.
* df.min(), df.max(), df.mean(), df.std(), df.var(): Compute individual statistical metrics.
* df.quantile([0.25, 0.5, 0.75]): Calculate percentiles (25th, 50th, 75th).
* df.hist(), sns.histplot(): Plot histograms to visualize feature distributions.
* df.dropna(), df.fillna(): Handle missing values for data cleaning.
* pd.concat(), merge(): Integrate multiple datasets.
* LabelEncoder, StandardScaler, MinMaxScaler: Transform data for model readiness.
* train\_test\_split, LogisticRegression, DecisionTreeClassifier, accuracy\_score: Build and evaluate classification models.

**Methodology**

**1. Summary Statistics**

* Load the dataset using Pandas.
* Use describe() and other functions to compute min, max, mean, standard deviation, variance, range, and percentiles for each feature.
* Summarize insights on feature distributions.

**2. Feature Distribution Visualization**

* Generate histograms for each numerical feature using Matplotlib or Seaborn.
* Identify skewness, modality, and spread of each feature.

**3. Data Preprocessing**

* **Cleaning**: Handle missing values using deletion or imputation.
* **Integration**: Combine multiple data sources if applicable.
* **Transformation**: Encode categorical features, scale numerical features.

**4. Model Building**

* Choose a classification algorithm (e.g., Logistic Regression, Decision Tree).
* Split the data into training and testing sets.
* Train the model and evaluate performance using accuracy or other metrics.

**Advantages of This Approach**

* **In-depth Data Understanding**: Summary statistics and histograms help uncover patterns and irregularities.
* **Better Model Accuracy**: Cleaned and transformed data leads to more effective model building.
* **Comprehensive Workflow**: Covers essential stages from raw data to actionable models.

**Challenges**

* Dealing with skewed or imbalanced datasets.
* Choosing the right data transformation techniques.
* Tuning classification models for best accuracy.

**Conclusion**

This assignment offered hands-on practice with statistical analysis, data preprocessing, and classification model building using Python. I learned how to:

* Extract meaningful statistical summaries.
* Visualize data distributions using histograms.
* Clean, integrate, and transform data for machine learning tasks.
* Apply classification techniques to build predictive models.