**Name:** Yash Gavali  
**Roll No:** 281032  
**Batch:** A2  
**Assignment 2**

**Statement:**In this assignment, we perform various data preprocessing and analysis operations on a dataset using R/Python. The tasks include computing summary statistics, visualizing feature distributions, data cleaning, integration, transformation, and building a classification model.

**Objective:**

* Compute and display summary statistics for dataset features.
* Visualize feature distributions using histograms.
* Perform data cleaning, integration, and transformation.
* Build and evaluate a classification model.

**Resources Used:**

* **Software used:** Jupyter Notebook / RStudio
* **Libraries used:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn (Python)
* **Methodology:**

1. **Data Collection and Exploration:**
   * Load the dataset into a DataFrame.
   * Display the first few rows to understand its structure.
   * Identify missing values and data inconsistencies.
2. **Summary Statistics Computation:**
   * Compute minimum, maximum, mean, range, standard deviation, variance, and percentiles for numerical features.
   * Generate descriptive statistics for categorical features.
3. **Feature Distribution Visualization:**
   * Use histograms to illustrate the distribution of continuous features.
   * Identify skewness and outliers in the dataset.
4. **Data Preprocessing:**
   * **Data Cleaning:** Handle missing values by imputation or removal.
   * **Data Integration:** Merge relevant datasets if needed.
   * **Data Transformation:** Normalize or standardize numerical features. Encode categorical variables.
5. **Model Implementation (Classification):**
   * Split the dataset into training and testing sets.
   * Select a suitable classification algorithm (e.g., Logistic Regression, Decision Tree, Random Forest).
   * Train the model and make predictions.
6. **Model Evaluation:**
   * Evaluate model performance using accuracy, precision, recall, and F1-score.
   * Use confusion matrix for better analysis.

**Results:**

* Summary statistics were successfully computed for all features.
* Histograms provided clear insights into feature distributions.
* Data cleaning, transformation, and integration improved dataset quality.

**Advantages:**

* Helps in understanding dataset characteristics before modeling.
* Improves data quality and consistency through preprocessing.
* Provides a structured approach to classification modeling.

**Disadvantages:**

* Data preprocessing can be time-consuming.
* Incorrect transformations may lead to model performance issues.

**Conclusion:**This assignment provided practical experience in exploratory data analysis, preprocessing, and classification modeling. By computing summary statistics, visualizing distributions, and building a classification model, we gained essential skills for data-driven decision-making.