Name: Adhija Bachhav

Roll No: 281053

Prn: 22311656

Batch: A-3

Assignment 2

Problem Statement:
Perform the following operations using R/Python on the data sets:
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 histogram.
c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification)

Introduction:

Data analysis is a crucial step in extracting meaningful insights from raw data. It involves various processes such as computing summary statistics, visualizing distributions, cleaning, integrating, transforming data, and finally, building predictive models. This document outlines these operations using R/Python and provides a comprehensive overview of statistical and machine learning techniques used in data processing.

Objective:

The objective of this document is to:

1. Compute and display summary statistics for each feature in the dataset.
2. Illustrate the feature distributions using histograms.
3. Perform data cleaning, integration, and transformation.
4. Develop a classification model for data analysis.

Theory:

1. **Summary Statistics:**
    - **Minimum & Maximum Value:** The smallest and largest values in the dataset.
    - **Mean:** The average value of a feature.
    - **Range:** Difference between the maximum and minimum values.
    - **Standard Deviation:** Measures the dispersion of data points from the mean.
    - **Variance:** The squared standard deviation, indicating data spread.
    - **Percentiles:** Values that divide the dataset into 100 equal parts.

2. **Feature Distribution (Histogram):**
    - A histogram represents the frequency distribution of numerical data.
    - It helps in identifying skewness, normality, and presence of outliers.
3. **Data Cleaning:**
    - Handling missing values (e.g., using mean/mode imputation or removal).
    - Removing duplicate records.
    - Correcting inconsistent data entries.
4. **Data Integration:**
    - Combining data from multiple sources into a single dataset.
    - Resolving schema mismatches and redundancy issues.
5. **Data Transformation:**
    - Normalization (scaling data to a fixed range, e.g., [0,1]).
    - Encoding categorical variables.
    - Feature engineering to enhance predictive power.
6. **Data Model Building (Classification):**
    - Classification is a supervised learning technique used for predicting categorical outcomes.
    - Popular classification algorithms include:
        - **Logistic Regression:** Used for binary classification.
        - **Decision Trees:** Tree-like structure for decision making.
        - **Random Forest:** Ensemble of decision trees for improved accuracy.
        - **Support Vector Machine (SVM):** Finds optimal decision boundary.
        - **Neural Networks:** Complex multi-layered models for deep learning.
    - Model evaluation is performed using metrics such as accuracy, precision, recall, and F1-score.

Conclusion:

Data analysis is an essential step in data-driven decision-making. By computing summary statistics, visualizing feature distributions, and performing data cleaning, integration, and transformation, we enhance data quality and usability. Finally, classification models help in making predictions based on the dataset, enabling businesses and researchers to gain valuable insights from their data. Python and R provide extensive libraries for efficient data analysis and modeling, making them ideal choices for this process.