Name: Jiya Sharma
Roll no: 281056
PRN: 22311741

Assignment 2

Statement:

Q. In this assignment we have to do
Perform the following operations using R/Python on the given dataset:

a) Compute and display summary statistics for each feature (e.g., minimum value, maximum value, mean, range, standard deviation, variance, and percentiles).
b) Data Visualization - Create a histogram for each feature to illustrate the feature distributions.
c) Perform data cleaning, data integration, data transformation, and data model building (e.g., classification).

Objective:

- To perform exploratory data analysis (EDA) by computing statistical summaries.
- To visualize the dataset to understand feature distributions.
- To clean, integrate, and transform data for better analysis.
- To build a classification model based on the dataset.

Resources used:

- Software used: Visual Studio Code
- Libraries used: Pandas, Matplotlib, sklearn

Theory:

Summary Statistics:

Summary statistics provide essential insights into the dataset. The key statistical measures include:

- Minimum & Maximum Values: Identify the smallest and largest data points in each feature.
- Mean: Represents the average value of a feature.
- Range: Difference between the maximum and minimum values.
- Standard Deviation: Measures the amount of variation in a feature.
- Variance: The square of standard deviation, showing dispersion in the data.
- Percentiles: Provide insights into the data distribution at specific percentage points.

Data Visualization:

Histograms are used to represent the frequency distribution of numerical data. They help in identifying skewness, outliers, and patterns in data distribution.

Data Processing Techniques:

- Data Cleaning: Handling missing values, removing duplicates, and correcting errors.
- Data Integration: Combining multiple sources of data into a unified dataset.
- Data Transformation: Scaling, normalization, and encoding categorical variables.
- Data Model Building (Classification): Applying supervised learning models such as Decision Trees, Random Forest, or Logistic Regression to classify data.

Methodology:

1. **Computing Summary Statistics:**

- Load the dataset using Pandas (Python) or dplyr (R).
- Use functions like describe(), min(), max(), mean(), std(), and percentile() to compute statistics.

2. **Data Visualization:**

- Generate histograms for each numerical feature using Matplotlib/Seaborn (Python) or ggplot2 (R).
- Interpret the distribution of each feature.

3. **Data Processing:**

- Cleaning: Handle missing values with imputation techniques or remove null values.
- Integration: Merge multiple datasets if applicable.
- Transformation: Normalize numerical values and encode categorical data.

4. **Data Model Building (Classification):**

- Choose a classification algorithm such as Decision Tree, Random Forest, or Logistic Regression.
- Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
- Train the model and evaluate its accuracy using a confusion matrix and performance metrics (Accuracy, Precision, Recall, F1-score).

Conclusion:

- Summary statistics provide an overview of the dataset’s distribution and variation.
- Histograms help in understanding feature distribution and identifying potential anomalies.
- Data preprocessing ensures that the dataset is clean and ready for analysis.
- Classification models can be built using the processed data to derive insights and make predictions.