Name: Rushikesh Kolekar
Roll No: 282020

Batch: B3

Practical 2

Problem Statement:

Q: - 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)

Software Used:

1. Python 3.x
2. Google Colab

Libraries and Packages:

- NumPy
- pandas
- matplotlib
- sklearn

Methodology:

1) Summary Statistics:

Computing summary statistics provides insights into key attributes of each dataset feature, such as mean, standard deviation, minimum, and maximum values. These metrics help in understanding data distribution and variability.

2) Data Visualization:

Histograms are created for each feature to visualize the distribution, detect trends, and identify any skewness or outliers. This visualization aids in understanding the overall structure of the dataset.

3) Data Preprocessing (Cleaning, Integration, and Transformation):

This step includes handling missing values, encoding categorical variables, and scaling numerical features to ensure a well-prepared dataset for model training.

4) Model Development:

A classification model is built using machine learning algorithms such as Decision Trees, Random Forests, or Support Vector Machines to make predictions based on dataset attributes.

Advantages:

- Exploratory Data Analysis (EDA) enhances the understanding of dataset structure, leading to well-informed decision-making.
- Data visualization simplifies trend detection, pattern recognition, and anomaly identification.
- Machine learning models facilitate predictive analysis, enabling applications in various industries such as fraud detection, customer segmentation, and healthcare.

Disadvantages:

- A solid understanding of domain knowledge is essential for accurate interpretation of EDA and model results.
- Relying too much on machine learning models without thoroughly understanding the dataset may lead to biased or misleading conclusions.

Applications with Example:

EDA and machine learning models have widespread applications in different fields, including:

- **Finance:** Credit risk assessment
- **Healthcare:** Disease prediction
- **Marketing:** Customer segmentation

Example: Predicting customer churn in a telecom company by analyzing customer demographics, service usage patterns, and subscription details.

Working / Algorithm:

1. Load the dataset using Pandas.
2. Compute summary statistics using the describe() function.

3. Generate histograms using Matplotlib and Seaborn for data visualization.
4. Perform data cleaning, integration, and transformation as needed.
5. Train a classification model using Scikit-learn.
6. Evaluate model performance using metrics such as accuracy, precision, and recall.

Conclusion:

This project underscores the importance of exploratory data analysis and machine learning in extracting valuable insights from data. Through systematic data preparation and analysis, meaningful patterns and trends can be identified, leading to the development of predictive models. These models have practical applications across various domains, aiding in real-world problem-solving and decision-making.