Name: Manisha Apchunde

Roll no: 281076

PRN: 22420244

Batch: A3

Assignment 02

Aim:

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)

Objective :

The objective of this assignment is to explore a dataset by computing summary statistics, visualizing feature distributions, and performing essential preprocessing steps such as data cleaning, integration, and transformation. The assignment concludes with building a basic classification model using the cleaned and processed data.

Software used:

Programming Language: Python

1. Visual Studio Code

Libraries and packages used: NumPy, pandas, matplotlib, sklearn

Theory:

Methodology:

1. Summary statistics: Computing summary statistics helps in understanding the basic properties of each feature in the dataset, such as mean, standard deviation, minimum and maximum values, percentiles, etc.
2. Data visualization: Creating histograms for each feature provides insights into the distribution of data, revealing patterns, skewness, and potential outliers.
3. Data cleaning, Integration, Transformation: These steps involve handling missing values, encoding categorical variables, scaling features, etc., to prepare the data for modeling.
4. Model Building: Building a classification model using machine learning algorithms such as Decision Trees, Random Forests, or Support Vector Machines.

Advantages:

1. EDA helps in understanding the structure and characteristics of the data, aiding in better decision making.
2. Data visualization facilitates the identification of trends, patterns, and outliers in the data.
3. Machine learning modeling enables predictive analysis, which can be used for various applications such as customer segmentation, fraud detection, medical diagnosis, etc.

Disadvantages:

1. EDA and modeling require domain knowledge and expertise to interpret the results accurately.
2. Over-reliance on machine learning models without proper understanding of the data can lead to biased or misleading conclusions.

Applications with example:

- **Health Diagnosis Dataset**:
  Patient medical records Summary stats help doctors understand blood pressure, cholesterol levels, etc. Classification model predicts diseases like diabetes based on input features
- **Banking and Finance Dataset:**

Customer loan data Summary statistics identify average income, loan amounts, and risk ranges Classification used to predict loan default.

Working/ Algorithm:

Step 1: Load the dataset

Step 2: Explore the data structure

Step 3: Compute summary statistics: mean, min, max, std, percentiles

Step 4: Visualize feature distributions using histograms

Step 5: Clean the data by handling missing values and duplicates

Step 6: Integrate datasets (if multiple)

Step 7: Transform features – encode categorical and scale numerical values

Step 8: Split data into training and testing sets

Step 9: Train a classification model on the training data

Step 10: Predict and evaluate the model on the test set.

This algorithm ensures that the dataset is fully prepared and modeled efficiently, following the best practices in data preprocessing and classification.

Conclusion:

In conclusion, this project demonstrates the importance of exploratory data analysis and machine learning modeling in understanding and extracting insights from data. By following a systematic approach, we can gain valuable insights into the data, identify patterns, and build predictive models that can be applied to real-world problems across various domains.