Name: Dev Bhangale

Roll no: 281059

Batch: A3

Assignment 2

Problem Statement:

Perform the following operations using R/Python on the given dataset:

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

Software used:

1. Python 3.x
2. Google colab

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

Methodology:

Summary Statistics: Computing summary statistics helps in understanding key characteristics of each feature in the dataset, including measures such as mean, standard deviation, minimum and maximum values, and percentiles. These statistics provide an overview of data distribution and variability.

Data Visualization: Creating histograms for each feature helps visualize data distribution, revealing trends, skewness, and the presence of outliers. This step aids in understanding the dataset's structure and potential anomalies.

Data Cleaning, Integration, and Transformation: This stage involves preprocessing tasks such as handling missing values, encoding categorical variables, and scaling numerical features to ensure the dataset is properly formatted for modeling.

Model Building: A classification model is developed using machine learning algorithms such as Decision Trees, Random Forests, or Support Vector Machines to predict outcomes based on the dataset's features.

Advantages

- Exploratory Data Analysis (EDA) enhances our understanding of the dataset's structure, leading to more informed decision-making.
- Data visualization helps identify trends, patterns, and anomalies, making data interpretation easier.
- Machine learning models enable predictive analysis, with applications in areas like fraud detection, customer segmentation, and medical diagnosis.

Disadvantages

- Effective EDA and model development require domain knowledge to accurately interpret results.
- Over-dependence on machine learning models without a deep understanding of the data can lead to biased or inaccurate conclusions.

Applications with Example

EDA and machine learning modeling can be applied in multiple domains, including:

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

Example:
Predicting customer churn for 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 with the `describe()` function.
3. Use Matplotlib and Seaborn to create histograms for data visualization.
4. Perform data cleaning, integration, and transformation where necessary.
5. Develop a classification model using Scikit-learn.
6. Evaluate the model's performance using metrics like accuracy, precision, and recall.

Conclusion

This project highlights the significance of exploratory data analysis and machine learning in extracting insights from data. By systematically analyzing and preparing the data, we can uncover patterns and trends that contribute to building predictive models. These models can then be applied across various domains to solve real-world problems effectively.