Name: Rohit Bhabire

Roll No: 281031 Batch: A2

Assignment 2

Statement

Q. 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) Illustrate the feature distributions using histograms.
c) Perform data cleaning, data integration, data transformation, and data model building (e.g., classification).

Objectives

1. To perform Exploratory Data Analysis (EDA) using statistical summaries.
2. To visualize data using histograms to understand distribution.
3. To clean, integrate, and transform data for improved quality.
4. To build a classification model for predictive analysis.

Resources Used

1. **Software**: Visual Studio Code
2. **Libraries**: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

---

Introduction to the Concepts

Summary Statistics

Summary statistics help in understanding the general characteristics of the dataset. Important statistical measures include:

- **Minimum & Maximum**: Identify smallest and largest values.
- **Mean**: Average value of data in each column.
- **Range**: Difference between max and min values.
- **Standard Deviation**: Spread of values around the mean.
- **Variance**: Square of standard deviation.
- **Percentiles**: Indicates values below which a certain percentage of observations fall.

Histograms

Histograms are used to visualize the distribution of numerical data. They help detect skewness, data spread, and outliers.

Data Preprocessing Techniques

- **Data Cleaning**: Fix or remove incorrect, corrupted, or missing data.
- **Data Integration**: Combining data from multiple sources.
- **Data Transformation**: Normalization, encoding, and scaling of features.
- **Model Building**: Use of classification models like Decision Trees, Random Forests, or Logistic Regression for prediction.

Methodology

1. **Reading the Dataset**
    - Using pd.read_csv() to load data into a DataFrame.
2. **Computing Summary Statistics**
    - Using .describe(), .mean(), .std(), .min(), .max(), .var(), and .quantile().
3. **Visualizing Data**
    - Histograms created using matplotlib.pyplot.hist() and seaborn.histplot().
4. **Data Preprocessing**
    - **Cleaning**: Use of .isnull().sum() to find missing values, handled by imputation or removal.
    - **Integration**: If multiple datasets exist, merged using pd.merge() or pd.concat().
    - **Transformation**: Categorical encoding with pd.get_dummies() or LabelEncoder, and normalization with MinMaxScaler.
5. **Model Building**
    - Data split using train_test_split().
    - Classification using models like DecisionTreeClassifier, LogisticRegression.
    - Performance measured using confusion matrix, accuracy, precision, recall, and F1-score.

Conclusion

This assignment introduced essential steps in Exploratory Data Analysis using Pandas and visualization libraries. We computed statistical summaries, visualized distributions using histograms, and processed the dataset to clean and transform it. Finally, we built a classification model to make predictions from the data. These steps are foundational for data science and machine learning workflows and equip us to handle real-world datasets efficiently.