Name: Snehal Yelwande
Roll No: 281063
Batch: A3

Assignment 2

Problem Statement:

Perform the following operations using Python on the given dataset:
a) Compute and display summary statistics for each feature (e.g., minimum, maximum, mean, range, standard deviation, variance, and percentiles).
b) Create histograms for each feature to visualize data distributions.
c) Perform data cleaning, data integration, data transformation, and build a classification model.

Objectives:

1. Perform exploratory data analysis (EDA) by computing statistical summaries.
2. Visualize the dataset to understand feature distributions.
3. Clean, integrate, and transform data for better analysis.
4. Build a classification model based on the dataset.

Resources Used:

- **Software:** Visual Studio Code
- **Libraries:** Pandas, Matplotlib, Scikit-learn

Theory:

Summary Statistics:

Summary statistics provide insights into the dataset, including:

- **Minimum & Maximum Values:** Identify the range of data.
- **Mean:** Average value of each feature.
- **Range:** Difference between max and min values.
- **Standard Deviation & Variance:** Measure the spread of data.
- **Percentiles:** Indicate the distribution of values.

Data Visualization:

Histograms represent the frequency distribution of numerical data, helping to identify patterns, skewness, and outliers.

Data Processing Techniques:

1. **Data Cleaning:** Handling missing values, removing duplicates, and correcting errors.
2. **Data Integration:** Combining multiple sources if needed.

3. **Data Transformation:** Normalization, scaling, and encoding categorical variables.
4. **Model Building (Classification):** Applying supervised learning models such as Decision Tree, Random Forest, or Logistic Regression.

Methodology:

1. Computing Summary Statistics

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

2. Data Visualization

- Generate histograms for numerical features using Matplotlib/Seaborn.
- Analyse the feature distributions.

3. Data Processing

- **Cleaning:** Handle missing values using imputation or removal.
- **Integration:** Merge datasets if required.
- **Transformation:** Normalize numerical data and encode categorical variables.

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 it using a confusion matrix and performance metrics like Accuracy, Precision, Recall, and F1-score.

Conclusion:

- Summary statistics help in understanding data distribution.
- Histograms provide insights into feature variations.
- Data preprocessing ensures the dataset is clean and usable.
- Classification models help in making predictions based on the data.
