Name: Prachi Shedge

Roll No: 281050

Batch: A2

Assignment 2

Statement

In this assignment, we aim to:
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 histograms.
c) Perform data cleaning, data integration, data transformation, and build a classification model.

Objective

1. Understand how to analyze structured data using Python.
2. Apply statistical techniques to summarize and explore datasets.
3. Implement data preprocessing steps like cleaning, encoding, transformation, and feature selection.
4. Build and evaluate a classification model to predict target labels.

Resources Used

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

---

Introduction to Data Analysis and Classification

Data analysis involves examining datasets to extract useful insights. This assignment focuses on key aspects of data preprocessing, visualization, and model building using Python.

Key Concepts:

- **Data Cleaning:** Handling missing values and ensuring data consistency.
- **Data Transformation:** Encoding categorical values and normalizing numerical features.
- **Data Integration:** Combining datasets meaningfully (if applicable).
- **Classification Model:** Training a logistic regression model to classify customers based on spending behavior.

Basic Functions Used

1. pd.read_csv() - Reads data from a CSV file into a Pandas DataFrame.
2. describe() - Provides summary statistics for numerical features.
3. hist() - Plots histograms to visualize feature distributions.
4. fillna() - Handles missing values by replacing them with mode values.
5. LabelEncoder() - Encodes categorical variables into numeric form.
6. train_test_split() - Splits the dataset into training and testing sets.
7. StandardScaler() - Standardizes numerical features for better model performance.
8. LogisticRegression() - Builds a classification model to predict customer spending behavior.
9. accuracy_score() - Evaluates model accuracy.
10. classification_report() - Provides a detailed performance summary of the model.

Methodology

1. Data Collection and Exploration

- The dataset used contains customer details, including **spending scores** and demographic attributes.
- The first step involves reading and displaying the dataset structure.

2. Data Preprocessing

- **Handling Missing Values:** Missing data is replaced with the most frequently occurring value (mode).
- **Encoding Categorical Features:** Convert text labels into numerical values using LabelEncoder().
- **Feature Scaling:** Standardizing numerical values to a common scale using StandardScaler().

3. Data Transformation & Feature Engineering

- **Target Variable Creation:** A new column HighSpender is created based on whether the **Spending Score** is above the median.
- **Feature Selection:** Excluding unnecessary columns (CustomerID, Spending Score) before training the model.

4. Model Building & Evaluation

- A **logistic regression model** is trained using train_test_split() with an 80-20 train-test split.
- The model is evaluated using accuracy and a classification report.

Results & Observations

- **Summary statistics** helped understand the range, mean, and spread of numerical attributes.

- **Histograms** provided insights into data distributions and possible outliers.
- **Preprocessing steps** (handling missing values, encoding categorical variables) ensured clean data for modeling.
- **Classification model** achieved an accuracy of **X.XX%**, showing its ability to distinguish high spenders.

Advantages of Data Preprocessing & Model Building

1. Enhances the dataset quality for better insights and model performance.
2. Standardization ensures numerical stability in machine learning models.
3. Classification helps businesses identify potential high-value customers.

Disadvantages

1. Handling large datasets requires significant computational resources.
2. Model accuracy depends on feature selection and preprocessing techniques.

Conclusion

This assignment demonstrated the complete data analysis pipeline, from **data cleaning and transformation to model training and evaluation**. The logistic regression model effectively classified customers, highlighting the importance of preprocessing in data science workflows.