Assignment 4

Name: Sakshi Vyavahare

Roll No: 281023
Batch: A1

Statement

In this assignment, we aim to:

- Apply an appropriate machine learning algorithm to a dataset collected from a cosmetics shop, containing customer details.
- Predict customer response to a special offer.
- Create a confusion matrix and compute:
  a) Accuracy
  b) Precision
  c) Recall
  d) F1-Score

Objectives

1. Compute summary statistics for a dataset using Python.
2. Visualize data distributions using histograms.
3. Perform data preprocessing, transformation, and integration.
4. Build a classification model on the cleaned dataset.
5. Evaluate model performance using confusion matrix and related metrics.

Resources Used

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

Introduction to Pandas and Data Analysis

Pandas is an open-source Python library designed for efficient data manipulation and analysis. It offers powerful data structures like Series (1D) and DataFrame (2D) for working with structured data.

Key Capabilities of Pandas

- Import data from CSV, Excel, or SQL databases

- Perform data cleaning, transformation, and handling of missing values
- Execute descriptive statistical analysis and visualizations
- Enable modeling tasks like classification, regression, and clustering

---

Basic Functions Used

Function	Purpose
pd.read_csv()	Load data from CSV file
describe()	Get summary statistics
hist()	Plot histograms
fillna()	Handle missing values
LabelEncoder()	Convert categorical data to numeric
train_test_split()	Split dataset for training and testing
LogisticRegression()	Train classification model
confusion_matrix()	Compute confusion matrix
accuracy_score(), precision_score(), recall_score(), f1_score()	Evaluate performance

---

Methodology

1. Data Collection and Exploration
    - Loaded the cosmetics customer dataset using pd.read_csv()
    - Examined data types, missing values, and feature categories
2. Data Preprocessing
    - Filled missing values using mean/median imputation
    - Removed duplicate records and handled inconsistent formatting
3. Summary Statistics
    - Used describe() to compute:
        - Mean, Min, Max
        - Standard Deviation, Variance
        - Percentiles
4. Visualization
    - Plotted histograms using hist() and sns.histplot() for numeric feature distribution analysis

5. Feature Engineering

- Applied LabelEncoder() for categorical data
- Selected features based on correlation analysis

6. Data Integration

- Merged data sources (if applicable) ensuring consistency

7. Model Building

- Used train_test_split() to split the data
- Trained a Logistic Regression model
- Evaluated using confusion matrix and computed:
    - Accuracy
    - Precision
    - Recall
    - F1-Score

Evaluation Metrics

Based on the confusion matrix:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Advantages of Pandas and Machine Learning

1. Simplifies data manipulation and preprocessing
2. Enables meaningful visual analysis of features
3. Allows implementation of predictive ML models

Disadvantages

1. Memory usage increases with larger datasets
2. Unstructured data requires complex preprocessing steps

Conclusion

This assignment enhanced our understanding of data preprocessing, feature engineering, and model evaluation. We worked with real-world customer data, visualized features, and built a logistic regression classifier to predict responses to marketing offers. Performance was evaluated using a confusion matrix and key metrics, offering a comprehensive approach to classification-based machine learning.