Assignment 4

Name: Yashraj Nimbalkar

Roll No: 281039
Batch: A2

Problem Statement

The purpose of this assignment is to:

• Utilize an appropriate machine learning algorithm on a dataset from a cosmetics store containing customer information.
• Predict whether a customer will respond positively to a special promotional offer.
• Generate a confusion matrix and calculate the following metrics:
a) Accuracy
b) Precision
c) Recall
d) F1-Score

Objectives

1. Generate descriptive statistics from a dataset using Python.
2. Create histogram-based visualizations to understand data distribution.
3. Perform comprehensive data cleaning, transformation, and integration.
4. Train a classification model using the processed dataset.
5. Assess model performance using confusion matrix and relevant evaluation metrics.

Tools and Libraries Used

• IDE: Visual Studio Code
• Python Libraries: Pandas, Matplotlib, Seaborn, Scikit-learn

Understanding Pandas and Data Analysis

Pandas is a powerful open-source Python library tailored for data handling and analysis. It provides flexible data structures like Series (1D) and DataFrame (2D) that simplify operations on structured data.

Main Features of Pandas

• Easily read data from CSV, Excel, or databases
• Clean, transform, and handle missing data effortlessly
• Perform statistical analysis and create insightful visualizations
• Prepare data for modeling tasks such as classification or regression

Common Functions Used

Function	Description
pd.read_csv()	Import data from CSV files
describe()	Summarize key statistical measures
hist()	Plot histograms for feature distribution
fillna()	Replace missing values
LabelEncoder()	Transform categorical values into numeric form
train_test_split()	Partition the data into training and test sets
LogisticRegression()	Train a logistic regression model
confusion_matrix()	Generate confusion matrix
accuracy_score(), precision_score(), recall_score(), f1_score()	Evaluate prediction quality

Methodology

1. Dataset Import & Exploration
   • Loaded the dataset using read_csv()
   • Reviewed feature types, missing entries, and overall structure
2. Data Preprocessing
   • Filled in missing data using statistical imputation (mean/median)
   • Removed redundant entries and resolved format inconsistencies
3. Statistical Summary
   • Used describe() to compute measures like:

- Mean, Minimum, Maximum
- Standard Deviation and Variance
- Percentile values

4. Data Visualization
   • Utilized hist() and sns.histplot() to visualize numeric attributes
5. Feature Engineering
   • Applied LabelEncoder() on categorical fields
   • Performed correlation checks to select impactful features

6. Data Integration
   • Combined data sources (if any), ensuring integrity and alignment
7. Model Development
   • Split data using train_test_split()
   • Trained a logistic regression classifier
   • Computed performance metrics from the confusion matrix:

- Accuracy
- Precision
- Recall
- F1-Score

Evaluation Metrics Explained

Given 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)

Benefits of Using Pandas and ML

1. Simplifies complex data wrangling tasks
2. Facilitates visual understanding through graphs
3. Makes implementation of prediction models efficient

Limitations

1. Handling large datasets can strain memory
2. Raw or unstructured data may need advanced transformation techniques

Conclusion

This assignment provided hands-on experience in data cleaning, feature engineering, and building classification models. We analyzed actual customer data, visualized features, and trained a logistic regression model to predict promotional offer responses. The model's performance was evaluated using a confusion matrix, offering a complete perspective on classification-based machine learning tasks.