Name: Aziz Sayyad
Roll No: 281069
Batch: A3

Practical 4

Problem Statement:

Apply an appropriate machine learning algorithm to a dataset, generate a confusion matrix, and compute the following performance metrics:

- Accuracy
- Precision
- Recall
- F1-score

Objectives:

1. Implement a supervised machine learning algorithm to predict customer responses.
2. Preprocess and analyze the dataset to enhance model performance.
3. Assess the model's performance using a confusion matrix.
4. Calculate key classification metrics, including Accuracy, Precision, Recall, and F1-score.

Resources Used:

- **Software:** Visual Studio Code, Anaconda(Jupyter Notebook)
- **Libraries:** Pandas, Matplotlib, Seaborn, Scikit-learn

Theory:

1) Classification

Classification is a supervised learning approach where a model learns to associate input features with predefined labels. The objective is to develop a model that accurately classifies new data points into distinct categories. This assignment focuses on binary classification, where the model predicts whether a customer will respond (Yes or No).

2) Confusion Matrix

A confusion matrix is used to evaluate the performance of a classification model. It consists of four key elements:

- **True Positives (TP):** Correctly predicted positive cases.
- **True Negatives (TN):** Correctly predicted negative cases.
- **False Positives (FP):** Incorrectly predicted positive cases (Type I Error).
- **False Negatives (FN):** Incorrectly predicted negative cases (Type II Error).

3) Evaluation Metrics

- **Accuracy:** Measures the proportion of correct predictions out of the total predictions.
- **Precision:** Determines the proportion of actual positive cases among predicted positives.
- **Recall:** Measures the proportion of actual positive cases correctly identified.
- **F1-Score:** The harmonic mean of Precision and Recall, balancing both metrics.

Methodology:

1. Data Preprocessing

- Load the dataset using Pandas.
- Handle missing values (imputation or removal).
- Encode categorical variables using one-hot encoding.
- Normalize numerical features using MinMaxScaler or StandardScaler.
- Split the dataset into training and testing sets (e.g., 75% training, 25% testing).

2. Selecting the Machine Learning Algorithm

Since this is a binary classification problem, suitable algorithms include:

- **Logistic Regression**
- **Decision Tree Classifier**
- **Random Forest Classifier**
- **Support Vector Machine (SVM)**
- **K-Nearest Neighbors (KNN)**
- **Neural Networks** (optional for advanced modeling)

3. Model Training & Prediction

- Train the selected machine learning model on the training dataset.
- Use the trained model to predict customer responses on the test dataset.

4. Confusion Matrix & Performance Evaluation

- Generate the confusion matrix to analyze True Positives, True Negatives, False Positives, and False Negatives.
- Compute the following performance metrics:
    - Accuracy
    - Precision

- Recall (Sensitivity)
- F1-Score

Conclusion:

- The selected machine learning model successfully predicted customer responses with reasonable accuracy.
- The performance metrics provided insights into the model's effectiveness, highlighting areas for potential improvement.
- Additional enhancements, such as feature engineering and hyperparameter tuning, could further optimize the model's accuracy and robustness.