Name: Dev Bhangale
Roll No: 281059
Batch: A3

Assignment 4

Problem Statement:

Apply appropriate machine learning algorithm on a dataset, generate a confusion matrix, and compute the following performance metrics:
a) Accuracy
b) Precision
c) Recall
d) F1-score

Objectives:

1. Apply a supervised machine learning algorithm to predict customer responses.
2. Analyze and preprocess the dataset to enhance model performance.
3. Evaluate the model's effectiveness using a confusion matrix.
4. Compute key classification metrics, including Accuracy, Precision, Recall, and F1-score.

Resources Used:

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

Theory:

Classification

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

Confusion Matrix

A confusion matrix is a tool for evaluating the performance of classification models. It consists of the following components:

- **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).

Evaluation Metrics

- **Accuracy:** Measures the overall correctness of predictions.
- **Precision:** Determines how many predicted positive cases were actually positive.
- **Recall:** Measures how many actual positive cases were correctly identified.
- **F1-Score:** A harmonic mean of Precision and Recall, balancing both metrics.

Methodology:

1. Data Preprocessing

- Load the dataset using Pandas.
- Handle missing values (either by imputation or removal).
- Encode categorical features (e.g., gender) 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, the following algorithms are suitable:

- 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.
- Predict customer responses using the test dataset.

4. Confusion Matrix & Performance Evaluation

- Generate the confusion matrix, determining True Positives, True Negatives, False Positives, and False Negatives.
- Calculate the following performance metrics:
    - **Accuracy**
    - **Precision**
    - **Recall (Sensitivity)**
    - **F1-Score**

Conclusion:

- The chosen 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.
- Further enhancements, such as feature engineering and hyperparameter tuning, could improve the model's predictive accuracy and robustness.