Name: Snehal Yelwande
Roll No: 281063
Batch: A3

Assignment 4

Problem Statement:

Apply an appropriate Machine Learning algorithm on a dataset. Create a confusion matrix based on the data and compute the following metrics:
a) Accuracy
b) Precision
c) Recall
d) F1-score

Objectives:

1. To apply a supervised machine learning algorithm for classification.
2. To preprocess the dataset for better model performance.
3. To evaluate the model using a confusion matrix.
4. To compute key classification metrics such as Accuracy, Precision, Recall, and F1-score.

Resources Used:

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

Theory:

Classification in Machine Learning:

Classification is a supervised learning technique where a model learns to map input features to predefined categories. The goal is to train a model that can classify new data points accurately. This assignment focuses on binary classification (e.g., predicting customer response as "Yes" or "No").

Confusion Matrix:

A confusion matrix helps evaluate the performance of a classification model. It consists of four key 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 the model.

- **Precision:** Measures the proportion of correctly predicted positive cases.
- **Recall (Sensitivity):** Measures how many actual positive cases were correctly predicted.
- **F1-Score:** The harmonic mean of precision and recall, balancing both metrics.

Methodology:

1. Data Preprocessing

- Load the dataset using Pandas.
- Handle missing values through imputation or removal.
- Encode categorical variables (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. Choosing the ML 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 ML model on the training dataset.
- Make predictions on the test dataset.

4. Confusion Matrix & Performance Metrics Calculation

- Compute the confusion matrix (TP, TN, FP, FN).
- Calculate the following metrics:
  a) Accuracy
  b) Precision
  c) Recall
  d) F1-Score

Conclusion:

- The selected ML model was able to classify responses with reasonable accuracy.
- The confusion matrix helped in assessing model performance.
- Based on the evaluation metrics, improvements can be made using feature engineering and hyperparameter tuning.