Name: Rahul Kshirsagar

Class: CSE(AI) B div

Roll no.: 282018

**Assignment 4 – Customer Response Prediction Using Machine Learning**

**Objective**

The aim of this assignment is to apply an appropriate machine learning algorithm to a dataset collected from a cosmetics shop. The dataset includes customer details and is used to predict whether a customer will respond positively to a special offer. The performance of the model is evaluated using a **confusion matrix**, along with metrics such as **accuracy**, **precision**, **recall**, and **F1-score**.

**Steps Performed**

**1. Data Loading and Preprocessing**

* The dataset was imported using the pandas library.
* Categorical features were encoded into numerical format using **Label Encoding** to make them suitable for model input.
* The **target variable** is defined as AHD (presence or absence of heart disease, used here as a placeholder for offer response).
* Feature scaling was done using **StandardScaler** to normalize numerical values and ensure better model performance.

**2. Model Training**

* The dataset was split into **training (80%)** and **testing (20%)** subsets.
* A **Random Forest Classifier**, which is an ensemble learning method based on decision trees, was used for classification. This algorithm provides high accuracy and handles overfitting well.

**3. Confusion Matrix**

A **confusion matrix** is used to evaluate the performance of the classification model by comparing the predicted labels with the actual ones.

* **True Positive (TP)**: Predicted = 1, Actual = 1 (correctly predicted positive response)
* **True Negative (TN)**: Predicted = 0, Actual = 0 (correctly predicted negative response)
* **False Positive (FP)**: Predicted = 1, Actual = 0 (incorrectly predicted positive response)
* **False Negative (FN)**: Predicted = 0, Actual = 1 (missed a positive response)

A heatmap of the confusion matrix was also plotted to visualize model performance.

**4. Evaluation Metrics**

Based on the confusion matrix, the following metrics were calculated:

* **Accuracy**: Proportion of total predictions that were correct.  
  Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN​
* **Precision**: Proportion of correctly predicted positive observations to the total predicted positive observations.  
  Precision=TPTP+FP\text{Precision} = \frac{TP}{TP + FP}Precision=TP+FPTP​
* **Recall (Sensitivity)**: Proportion of correctly predicted positive observations to all actual positive observations.  
  Recall=TPTP+FN\text{Recall} = \frac{TP}{TP + FN}Recall=TP+FNTP​
* **F1-Score**: Weighted average of Precision and Recall.  
  F1-Score=2⋅Precision⋅RecallPrecision+Recall\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}F1-Score=2⋅Precision+RecallPrecision⋅Recall​

These metrics provide a holistic view of the model’s performance, especially in imbalanced datasets.

**5. Model Saving (Optional)**

The trained Random Forest model can be saved using **Pickle** for future predictions without the need for retraining.

**Conclusion**

This assignment showcases a complete machine learning pipeline:

* From data preprocessing,
* To model building,
* To performance evaluation using the confusion matrix and key classification metrics.

The Random Forest model showed strong performance, making it suitable for predicting customer response in a retail setting such as a cosmetics shop.