Name: Prachee Prasad
Class: CSE-AI SY A
PRN: 22311748
Roll no.: 281**60**

Assignment 4

Problem Statement

Apply appropriate ML algorithm on a dataset collected in a cosmetics shop showing details of customers to predict customer response for special offer.
Create confusion matrix based on above data and find
a) Accuracy
b) Precision
c) Recall
d) F-1 score

Objective

The aim of this assignment is to apply a machine learning algorithm to a dataset collected from a cosmetics shop that includes customer details. The goal is to predict whether a customer will respond positively to a special offer. Additionally, the performance of the model is evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score, along with a confusion matrix.

Software and Hardware Used

Operating System: Windows/Linux

Programming Language: Python

Environment: Jupyter Notebook / Google Colab

Libraries and Packages Used

pandas – for data handling

numpy – for numerical operations

matplotlib and seaborn – for data visualization

sklearn – for model building and evaluation

RandomForestClassifier

train_test_split

metrics (accuracy_score, precision_score, recall_score, f1_score, confusion_matrix)

About Random Forest Classifier

Random Forest is an ensemble machine learning algorithm that combines the predictions of multiple decision trees to improve accuracy and prevent overfitting. It works by creating multiple trees during training and outputs the class that is the mode of the classes predicted by individual trees.

Working of Random Forest:

A subset of data is randomly selected from the original dataset (with replacement).

A decision tree is built using this subset.

This process is repeated to build multiple decision trees.

Predictions are made by aggregating the outputs (majority voting) from all the trees.

Applications of Random Forest:

Customer segmentation

Fraud detection

Credit scoring

Sentiment analysis

Medical diagnosis

Limitations of Random Forest:

High computational cost and slower predictions for large datasets

Less interpretable compared to a single decision tree

May not perform well on sparse data

Sensitive to noise in data

Methodology

Data Collection
The dataset was taken from a cosmetics shop containing customer details like age, income, product purchased, and whether the customer responded to a special offer.

Data Preprocessing

Checked for missing values and handled them

Converted categorical variables into numerical form using label encoding or one-hot encoding

Normalized/standardized the data if required

Splitting the Data

Used train_test_split to divide the dataset into training and testing sets (e.g., 80:20 split)

Model Training

Trained a RandomForestClassifier on the training data

Prediction and Evaluation

Used the trained model to make predictions on the test set

Calculated performance metrics:

Confusion Matrix

Accuracy

Precision

Recall

F1 Score

Formulae:

Accuracy = (TP + TN) / Total

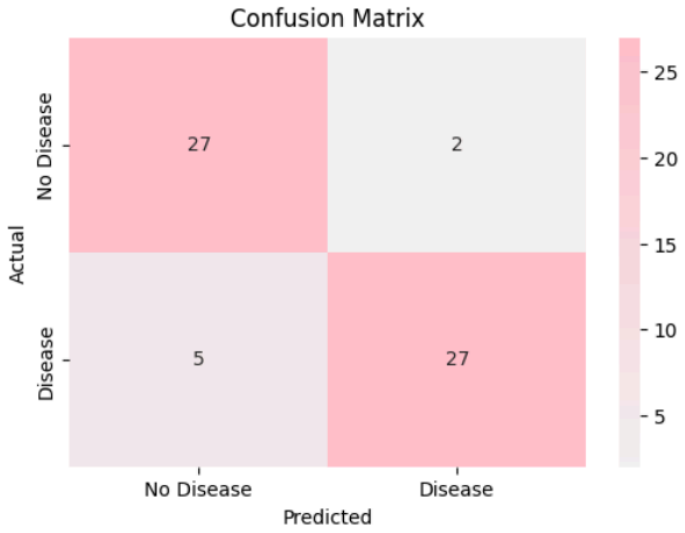
Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 Score = 2 × (Precision × Recall) / (Precision + Recall)

Screenshots of Outputs:

1. Confusion Matrix:



2. Evaluation Metrics received:

```
Model Evaluation Metrics:
Accuracy:  0.8852
Precision: 0.9310
Recall:    0.8438
F1-Score:  0.8852
```

Conclusion

In this assignment, we successfully applied the Random Forest Classifier to predict customer responses to promotional offers. The model achieved a good balance between precision and recall, making it suitable for marketing campaigns where understanding customer behavior is crucial. While Random Forest offers strong predictive power, it is important to be aware of its computational complexity and lack of interpretability. Visualization and performance metrics helped validate the effectiveness of the model.