Assignment No. 05

Aim:

Data Analytics II

1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset.
2. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, **Recall on the given dataset.**

Prerequisites:

1. Prior knowledge of Python programming.
2. Google Colab / Python IDE
3. Jupyter Notebook

Objectives: to Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision,Recall on the given dataset

Theory:

1. Importing Libraries

Social_Network_Ads.csv dataset.

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt # Importing the required libraries
import seaborn as sns
%matplotlib inline
```

Logistic Regression : Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorize, logistic regression may be able to help. It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made.

Steps:

```
1. #To check if the data is equally balanced between the target classes
2. #Defining features and target variable
3. #Splitting the data into train and test set
4. #Predicting using Logistic Regression for Binary classification
5. #Evaluation of Model - Confusion Matrix Plot
6. # Plot non-normalized confusion matrix
7. #extracting true_positives, false_positives, true_negatives, false_negatives
8. #Accuracy #Precision #Recall #F1 Score
```

Confusion Matrix Definition

A confusion matrix is used to judge the performance of a classifier on the test dataset for which we already know the actual values. Confusion matrix is also termed as Error matrix. It consists of a count of correct and incorrect values broken down by each class. It not only tells us the error made by classifier but also tells us what type of error the classifier made. So, we can say that a confusion matrix is a performance measurement technique of a classifier model where output can be two classes or more. It is a table with four different groups of true and predicted values.

Terminologies in Confusion Matrix

The confusion matrix shows us how our classifier gets confused while predicting. In a confusion matrix we have four important terms which are:

1. **True Positive (TP)**
2. **True Negative (TN)**
3. **False Positive (FP)**
4. **False Negative (FN)**

True Positive (TP)

Both actual and predicted values are Positive.

True Negative (TN)

Both actual and predicted values are Negative.

False Positive (FP)

The actual value is negative but we predicted it as positive.

False Negative (FN)

The actual value is positive but we predicted it as negative.

Performance Metrics

Confusion matrix not only used for finding the errors in prediction but is also useful to find some important performance metrics like Accuracy, Recall, Precision, F-measure. We will discuss these terms one by one.

Accuracy

As the name suggests, the value of this metric suggests the accuracy of our classifier in predicting results.
It is defined as:
Accuracy = (TP + TN) / (TP + TN + FP + FN)
A 99% accuracy can be good, average, poor or dreadful depending upon the problem.

Precision

Precision is the measure of all actual positives out of all predicted positive values. It is defined as:
Precision = TP / (TP + FP)

Recall

Recall is the measure of positive values that are predicted correctly out of all actual positive values.
It is defined as:
Recall = TP / (TP + FN)
High Value of Recall specifies that the class is correctly known (because of a small number of False Negative).

F-measure

It is hard to compare classification models which have low precision and high recall or vice versa. So, for comparing the two classifier models we use F-measure. F-score helps to find the metrics of Recall and Precision in the same interval. Harmonic Mean is used instead of Arithmetic Mean.
F-measure is defined as:
F-measure = 2 * Recall * Precision / (Recall + Precision)
The F-Measure is always closer to the Precision or Recall, whichever has a smaller value.
Calculation of 2-class confusion matrix
Let us derive a confusion matrix and interpret the result using simple mathematics. Let us consider the actual and predicted values of y as given below:

Actual y and Predicted y with threshold 0.5		
1	0.7	1
0	0.1	0

0 0.6 1		
1	0.4	0
0	0.2	0

Now, if we make a confusion matrix from this, it would look like:

N=5 Predicted 1 Predicted 0		
Actual: 1	1 (TP)	1 (FN)
Actual: 0	1 (FP)	2 (TN)

This is our derived confusion matrix. Now we can also see all the four terms used in the above confusion matrix. Now we will find all the above-defined performance metrics from this confusion matrix.

Accuracy
Accuracy = (TP + TN) / (TP + TN + FP + FN)
So, Accuracy = (1+2) / (1+2+1+1) = 3/5 which is 60%.
So, the accuracy from the above confusion matrix is 60%.

Precision
Precision = TP / (TP + FP) = 1 / (1+1) =1 / 2 which is 50%.
So, the precision is 50%.

Recall
Recall = TP / (TP + FN) = 1 / (1+1) = ½ which is 50%
So, the Recall is 50%.

F-measure
F-measure = 2 * Recall * Precision / (Recall + Precision)
= 2*0.5*0.5 / (0.5+0.5) = 0.5
So, the F-measure is 50%.

Conclusion:

Thus we have computed Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision,Recall on the given dataset(Social_Network_Ads.csv)