LAB REPORT

Name: Devyani Gorkar
Roll.no:.B20ME027

Question1:

Reading the csv file using pandas.read_csv.

Question 2

Using test_train_split split the dataset into X_train, X_test, X_validate, y_train, y_test, y_validate .The distance similarity used here is Euclidean distance

Question 3:

KNN is the function which takes 4 inputs.X_train,X_test,y_train,K.No. of nearest neighbours is K is chosen randomly.The first for loop selects a data point from the X_test and then finds its distance with every data point from X_train and store the distances in a list called 'distances'.'new' is another list which stores the sorted values of the distance list.Now depending upon the value of K store first K values from the sorted 'new' list in 'close'.The list 'index' will store the values of indices.With the help of indices we find their corresponding output i.e 1 or 0 and store it in a list called 'out'.Depending on the number of ones and zeros present we assign a value to the new datapoint.Now we do this same process for every datapoint in X_test and store the predicted output everytime in a list called 'final'.
'er' is a list which stores the error rates of the dataset for different values of k.The graph plotted shows the values of different error rates with different k values.From the graph it can be seen that minimum error is obtained when k=8.From the classification report we can see that the accuracy and F1 score is maximum when k=8 and error is minimum hence the optimal value of k will 8.On comparing the confusion matrix we can see that matrix of k=8 is better than other matrices.

Question 4:
Initially when we used the library we used k=5 and the accuracy obtained is 0.72 but while implementing from scratch we find that the optimal value of k=8 since it gives minimum error rate and maximum accuracy that is 0.75 which is pretty close to 0.72.On using k=8 in library function of knn we find that the accuracy is 0.77 which is even better than before.Hence we can conclude that k=8 is the optimal value for both the cases.The confusion matrix also seems to be similar except for the values of False positive.