REPORT

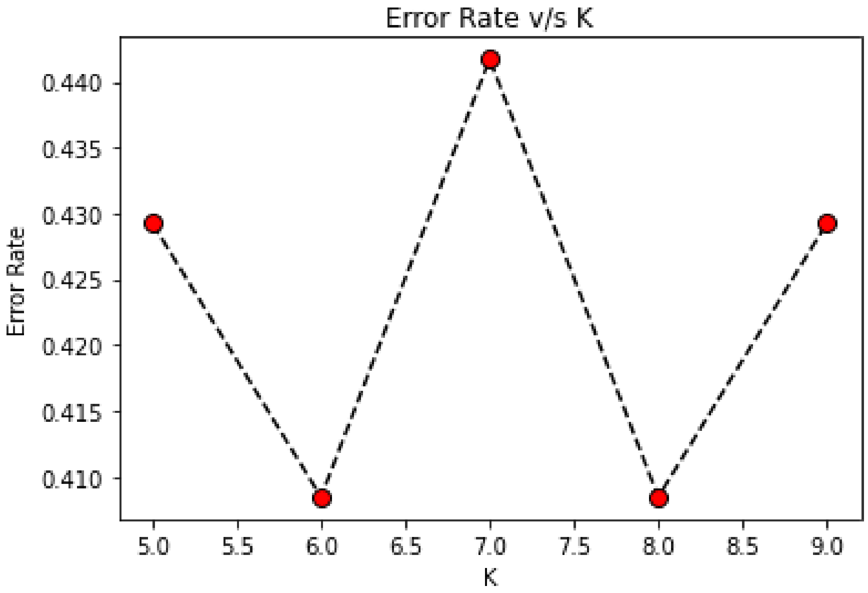
PRINCY GAUTAM (B20BB051)

Method

- We imported necessary libraries for calculation and plotting the graph.
- Assigned the name df to the data frame to read the csv file.
- Then we defined X as a data frame excluding the outcome column and Y containing the outcome column.
- Split the dataset into train test and validation in the ratio of 70:15:15, using train test split twice.
- Made a list of five different values of “K”.
- Wrote a code to perform the K-nearest neighbor classification from scratch without using any library.
- Made a class named KNN () as follows:
    - Initialized the variable, K
    - Defined a function named ‘fit’ to store the training set
    - Defined a function named ‘euclidean’ to return the Euclidean distance
    - Defined a function named ‘predict’ in which we initialized Y_predict and then found the K nearest neighbors from current test example. And finally used mode to get the most frequent class in K neighbors.
    - Defined the function named ‘k_nearest_neighbor’ to current test sample. Calculate all the euclidean distances between current test example x and training set, X_train
    - Sort Y_train according to the euclidean distances list and store it into Y_train_sorted.
    - Then we gave the driver code under which we modeled the training, put prediction on test set and printed the accuracy and confusion matrix for all values of K.
- Applied cross validation to identify the optimal value of K using cross_val_score and appended the mean of these in an empty list.
- Then we plotted the error rate v/s K graph which gave us **optimal value of k=8**
- Since it gave least error according to the graph.
- Print the accuracy and confusion matrix using sklearn libraries for sklearn model
- And print the accuracy and confusion matrix for scratch model.

Results

- Optimal value is k = 8 because it gave the least error.
- The graph of error rate v/s k obtained from scratch also gave least error at k = 8



- At k = 6 also, we are getting error similar to k = 8 but there is a minute difference between them with least error at k = 8.
- We printed the accuracy and confusion matrix for all values of k and with k = 8 having maximum accuracy.
- Thus, optimal k is the one with least error rate and maximum accuracy.
- The model from scratch and sklearn library shows us that:
- Accuracy and confusion matrix for both the models are same at k = 8, thus both the models are similar.

Colab file link:
https://colab.research.google.com/drive/1KQCWvDOv1j4I0v8W9HfrTRi8cw7RacTu#scrollTo=TF3opKkiRDir