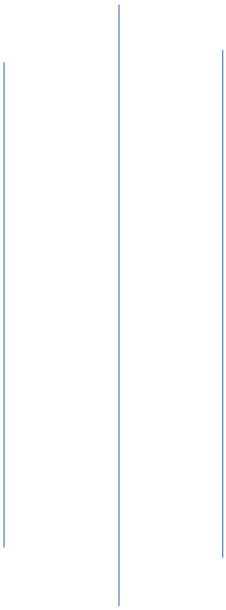
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

(Deemed to be University)
Patiala, Punjab

CSED-Experiential Learning Activities
E110-2024

REPORT SUBMISSION
ON

Handwritten Digit Recognition (KNN)



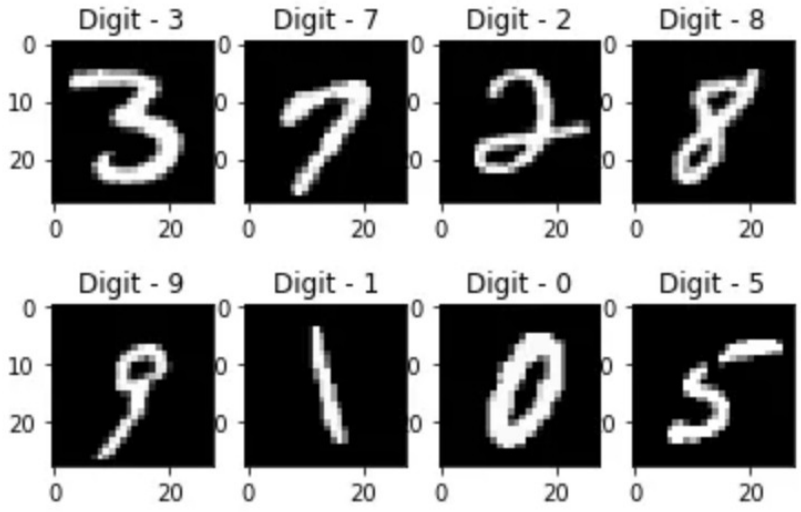
Submitted by:
Aadarsha Mahato(2CS9)(102217264)

Introduction:

Digit recognition system is the machine to train in recognizing the digits from different sources like emails, bank cheque, papers, images, etc. and in different real-world scenarios for online handwriting recognition on computer systems, recognize number plates of vehicles, processing bank cheque amounts, numeric entries in forms filled up by hand and so on

MNIST Dataset

MNIST(Modified National Institute of Standards and Technology)dataset contains total of 70,000 handwritten digits labeled images from 0 to 9.
Handwritten digits are images in the form of 28*28 gray scale intensities of images representing an image along with the first column to be a label (0 to 9) for every image.



Sample from MNIST dataset

K-Nearest Neighbors

K-nearest Neighbor is a Non parametric ,lazy and supervised machine learning algorithm used for both Classification and Regression.

-Uses the phenomenon “similar things are near to each to each other” .

-It predicts the class of the new data point by majority of votes of k nearest neighbors based on the similarity measure(ie., distance functions).

-Varying the value of K ,differs the prediction class.

-It is commonly used for its easy of interpretation and low calculation time.

Implementation of KNN includes:

1)Calculation of Euclidean distance

Euclidean distance is the square root of the sum of squared distance between two points.

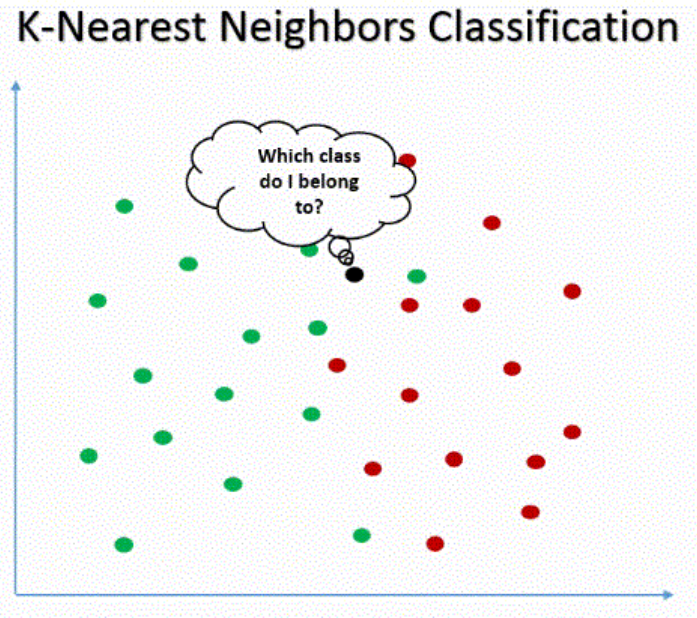
Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

2)Get the k nearest neighbours after sorting distance

-In order to find the neighbors we need to first sort the distance in ascending order,np.argsort () is used to find the index of minimum distance .

-After that we will arrange the data according to the sorted index.

-Silcing the data according to the number of neighbors.



3)Predicting the class of the new data point

-The test data will present in the class with majority of the votes .So,to find that we will use max () function

-They key in the max function groups the neighbors wrt to their classes and .count will count the number of neighbors in each class.

-Finally max returns the class with majority votes which will be the predicted class of the test data.

4)Accuracy calculation

Accuracy shows how close the measured value to the true value.

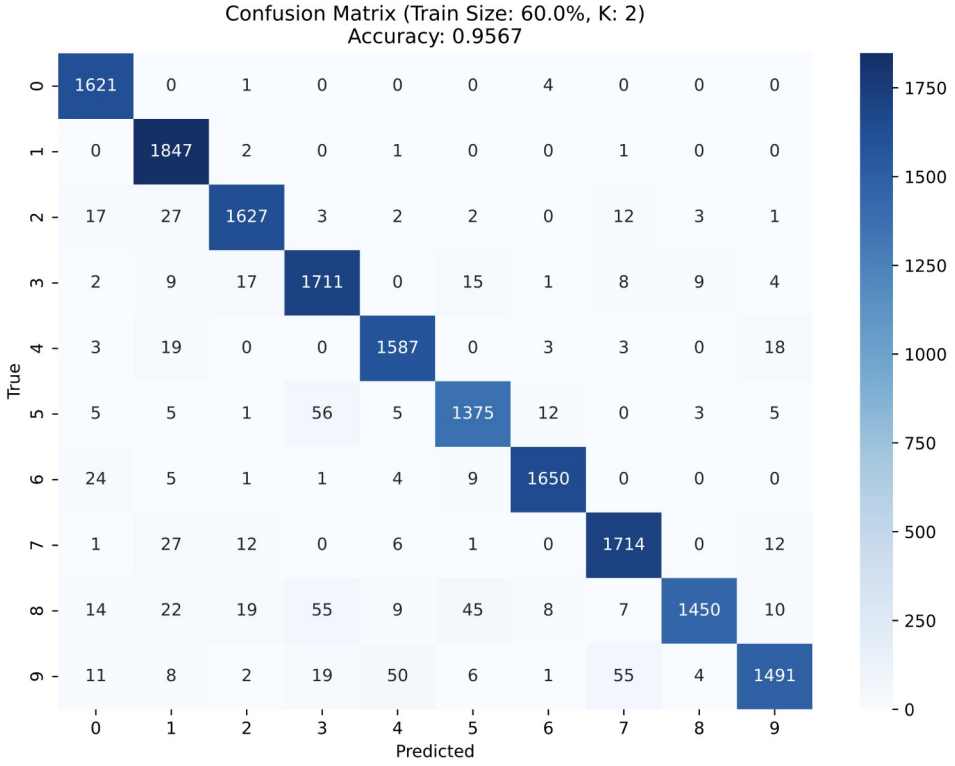
$$Accuracy = \frac{Number\ of\ Features\ classified\ correctly}{Number\ of\ Features}$$

-Accuracy is calculated by dividing the correctly classified samples count by total samples.

-Higher the accuracy ,more efficient the model .

Implementing the KNN

After training your model using K-Nearest Neighbor Algorithm with having values of K as {2,4,5,6,7,10}, over data.csv file provided with train and Test split of the data in the ratio of 60:40, 70:30, 75:25, 80:20, 90:10, 95:5 and evaluating the performance of the model over test data for all these scenarios (36 cases)
we get result as follow:



Above provided image is of the confusion matrix when K=2 and train test split ratio is 60:40 we got accuracy of 95.67% and similarly the rest 35 cases are provided on the separate pdf.