Artificial Intelligence Assignment 3

TEAM MEMBERS:-

Ventrapragada Sai Rathan – 201601104

Anirudh Kannan V.P - 201601004

GOALS:

- K-Nearest Neighbour Classifier on the OCR data sets given 3-fold cross validation needs to be implemented .
- Naïve Bayes Classifier on the given training and testing data sets needs to be implemented**.**

Goal 1:

Idea!!

- We are given an unclassified example, and we are able to assign it a class-group by observing what its nearest neighbours belong to.
- Out of k nearest neighbours some will be assigned to class '1' ,on the other hand others will be assigned to class '2'.
- Now the class with most probability will be chosen and assigned it to our unclassified example.

Algorithm:

- In r-fold cross validation, the training set is divided into r-folds and assuming that each of the r-fold blocks is a validation set the best k will be found out by having the other blocks as the training set.
- The best k is found(out of 1 to 25) to classify an instance from the training examples.
- This k is used to classify each of the testing examples.
- The correctly classified examples are used to keep track of classification accuracy.

Goal 2:

Idea!!

- Given an unclassified example, the probability of that example being classified to each of the classes (0 to 9) needs to be found out, given its features.
- This example will be classified into the class with the most probability.
- According to Baye's Theorem, P(A/B) uses P(B/A). So with these training examples all P(features/class) values are calculated and stored.

Algorithm:

- A 3D array i.e train[b][a][c] is maintained which signifies the probability of 'a' th feature having value c, given class = b.
- This 3D array is populated with the training examples and divided by the total class members to get the probability.

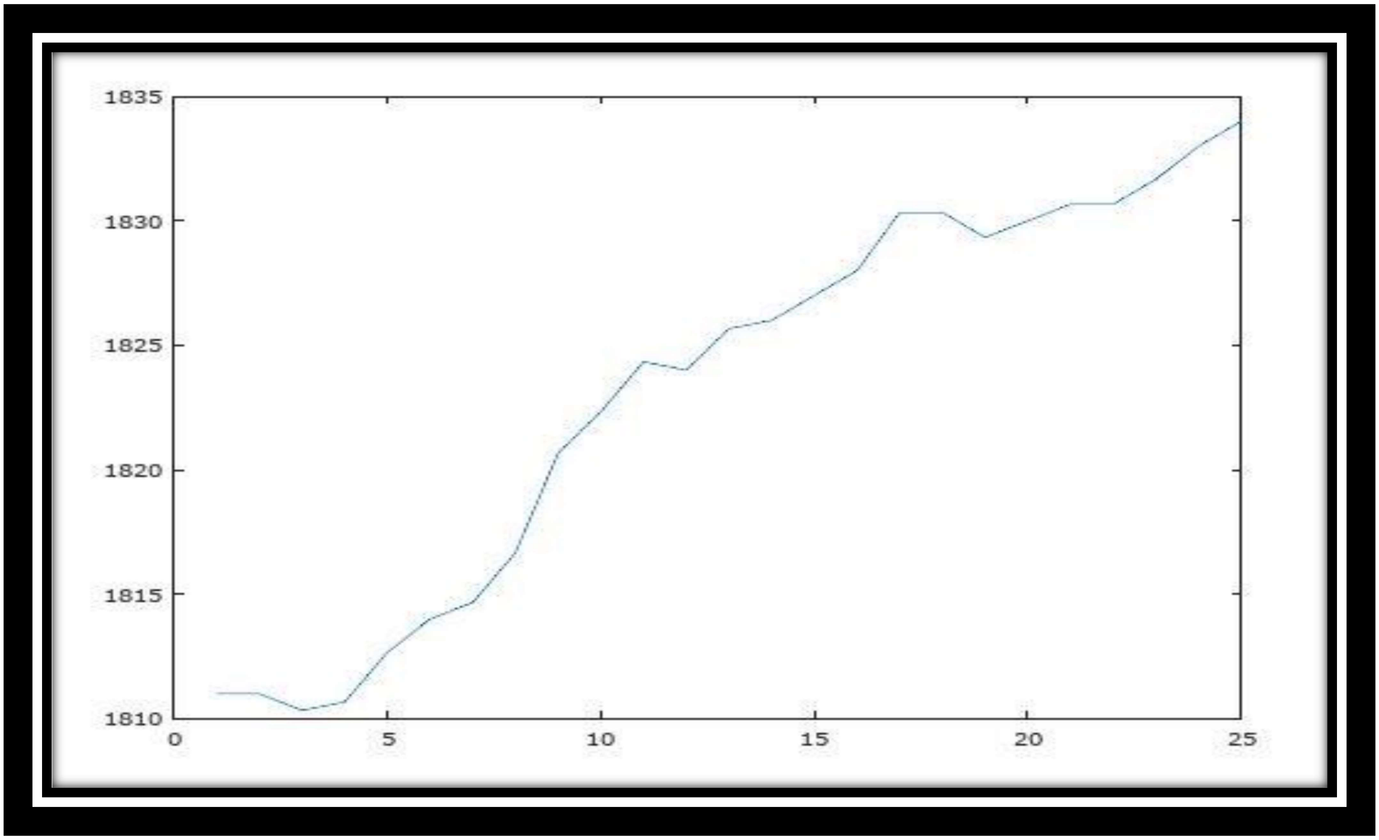
- This 3D array and Bayes classifier formulae are used to classify each of the testing examples.
- The correctly classified examples are used to keep track of classification accuracy.

Formulae:

$$
\begin{aligned}
&P\left(\frac{(f_1 = a_1, f_2 = a_2, f_3 = a_3 \ldots \ldots f_{192} = a_{192})}{class_i}\right) \\
&= P\left(\frac{(f_1 = a_1)}{class_i}\right) * P\left(\frac{(f_2 = a_2)}{class_i}\right) * P\left(\frac{(f_3 = a_3)}{class_i}\right) * \ldots \ldots \\
&* P\left(\frac{(f_{192} = a_{192})}{class_i}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\log P\left(\frac{(f_1 = a_1, f_2 = a_2, f_3 = a_3 \ldots \ldots f_{192} = a_{192})}{class_i}\right) \\
&= \log P\left(\frac{(f_1 = a_1)}{class_i}\right) + \log P\left(\frac{(f_2 = a_2)}{class_i}\right) + \log P\left(\frac{(f_3 = a_3)}{class_i}\right) \\
&+ \ \ldots \ldots + \log P\left(\frac{(f_{192} = a_{192})}{class_i}\right)
\end{aligned}
$$

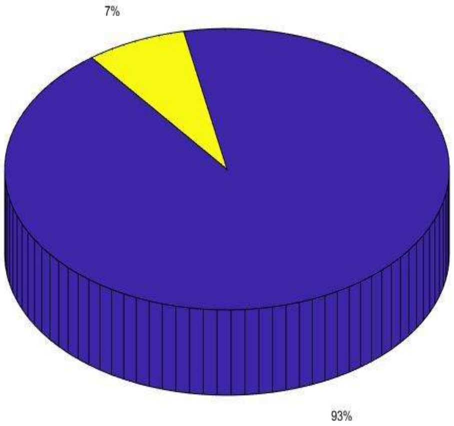
OUTPUT:



- The above plot depicts the average errors obtained for each of the k values.
- Thus it is evident that k = 3, has the minimum average value. Therefore k = 3 has to be used for classifying the test data.

3D Pie Chart for the Classification:

Goal 1:

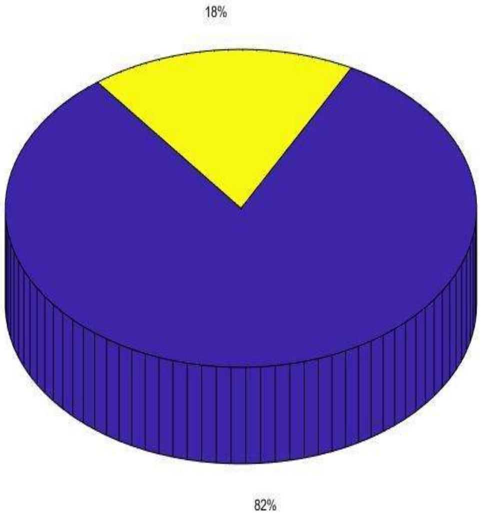


Correctly Classified

Wrongly Classified

Accuracy obtained → 92.709271%

Goal 2:



Correctly Classified

Wrongly Classified

Accuracy obtained → 81.818182%