HANDWRITTEN CHARACTER RECOGNITION USING PYTHON

Kashish jain , Vishal, Tintus, Abhishek
Research scholar,
Ms Richa Sharma ,Assistant professor,
Department of Computer science and technology,
Bhagwan Parshuram Institute of Technology ,Delhi ,India

Abstract

The communication of ideas in today's world is necessary and there are various symbols and characters denoting a particular meaning being used, the digits being one of them. This paper discusses the digit recognition using machine learning classifiers and comparing their performances. The input given in this case is the set of digits (0-9). And further, the model is trained using Random Forest Classifier, Naives Bayes using the Jupyter notebook in the Python language. The classifiers have been compared on various parameters, which are discussed in the paper. Though, the classifier can work only on the digits (0-9) and cannot classify more than a single digit, when given as input. In addition to that, the prerequisite is that the input digits should only be black or white in colour.

***Index Terms*- Machine Learning, Naives Bayes algorithm, Random Forest Classifier, Digit Recognition**

1.Introduction:

Machine Learning is being widely used across all the domains. And now, attempts are being made to incorporate it to solve real life problems, including object detection, pattern recognition, face recognition, classifiers, natural language The accuracy generated by the system depends on how well it is trained and tested by the user.The goal of the article is to observe and compare various classifiers, namely the Naives Bayes classifier, Random Forest Classifier on the basis of parameters, including accuracy, precision, recall,f1-score and support .The article also aims to classify the training models on the basis of the listed parameters.

In the project, we aim to compare the classifiers, namely the Random Forest Classifier and the Naives Bayes model. We will implement the model by importing the images from the Model National Institute of Standards and Technology (MNIST) dataset. It contains a total of 70000 images, out of which some is split into training and some in the testing purposes.Handwritten Character recognition is the system takes a set of input characters, from 0-9 and predicts which digit is displayed, according to the classifier used. The classifiers include the Random Forest Classifier, Naïve Bayes model, Support Vector Machines.

Several classifiers have been previously been designed using the various models such as Naïve Bayes, Random Forest Classifier, k nearest neighbour, Support Vector Machines (Support Vector Machines) and they have generated varying amount of accuracy and the other parameters. While with regular experimentations and using the Convolution Neural Network (CNN), the accuracy has been maximized. Thus, there has been major advancements in the field of machine learning and thus, their applications is being used in various fields, such

as natural language processing (NLP), pattern recognition, face recognition, spam detection, fake news detection as well as sentiment analysis and many more. Similarly, there has been several advancements in the digit recognition as well.

While several papers have also implemented various models for digit recognition to compare the classifier performance based on several factors such as accuracy, precision, f1 score, support and recall. And it helps to get an idea of which model works best depending on certain factors and under certain conditions, which further helps to implement the best classifier for getting high accuracy as well as reduce the number of false positives and false negatives in the system.In the Pattern Recognition field, growing interest has been shown in recent years for Multiple Classifier Systems and particularly for Bagging, Boosting and Random Subspaces. Those methods aim at inducing an ensemble of classifiers by producing diversity at different levels. Following this principle, Breiman has introduced in 2001 another family of methods called Random Forest. Our work aims at studying those methods in a strictly pragmatic approach, in order to provide rules on parameter settings for practitioners. For that purpose we have experimented the Forest-RI algorithm, considered as the Random Forest reference method, on the MNIST handwritten digits database. In this paper, we describe Random Forest principles and review some methods proposed in the literature. We present next our experimental protocol and results. We finally draw some conclusions on Random Forest global behavior according to their parameter tuning.

2.LITERATURE REVIEW :

Handwritten character recognition is process of converting the hand written work over page to a attractive digital format.HCR is a intelligent work done throw scanning the images will complete the analysis of character with output. CR require proper handling of complexity of written content, writing environment, materials, etc. HCR techniques are based on extracting various features of handwritten.

Isha Vats,Shamandeep Singh[1] In this paper, system was based on recognition of offline handwriting numerals. The main aim of the proposed work in this paper was to efficiently recognize the offline handwritten digits with a higher accuracy. But a difficult problem in this field was the recognition of completely touching handwritten digits and in this paper the proposed system focused on segmentation for isolating the digits so multiple images can be recognize.

Gunjan Singh, Sushma Lehri [2] Handwritten characters was a difficult task because characters are written in various ways, so they could be of different sizes, orientation, thickness and dimension. An offline HCR(Hindi) system using neural network is presented in this paper. Neural networks were good at recognizing handwritten characters as these networks are insensitive to the missing data.A Backpropagation neural network is used for classification. Experimental result of this system shows that results 93%.

S S Sayyad, Abhay Jadhav, Manoj Jadhav, Pradip Bele,Smita Miraje,Avinash Pandhare [3] In this paper A neural network approach is proposed for automatic offline character recognition system. In this paper, work has been performed to recognize Devanagari characters using multilayer perceptron.Various patterns of characters were created in the matrix with the use of binary form and stored in the file.This sysrem used the back

propagation neural network for efficient recognition and neuron values were transmitted by naive bay's method in the neural network.

Shabana Mehfuz, Member IEEE, 2 Gauri katiyar[4] This paper provides review of existing works in HCR based on soft computing technique during the past decade

Prof. Swapna Borde, Ms. Vinaya Patil, Ms. Ekta Shah, Ms. Priti Rawat [5] This paper presents a fuzzy approach to recognize characters. Fuzzy sets,fuzzy logic were used as bases for representation of fuzzy character and for recognition.Fuzzy-based algorithm which first segments the character and then using fuzzy system gives the characters that match the given input and then using defuzzication system finally recognizes the character. No training is needed by this system for

recognition.

Fatos T. Yarman-Vural and Nafiz Arica [6] The rapidly growing computational power enables the implementation of the present Character Recognition methodologies and creates an increasing demand on many emerging application domains, which require more advanced methodologies. The available Character Recognition techniques with their superiorities and weaknesses are reviewed. The Character Recognition is discussed, and directions for future research are suggested. Special attention is given to the off-line HCR since this area requires more research.

Ms. Seema A. Dongare, Ms. Snehal V. Waghchaure, Prof. Dhananjay B. Kshirsagar [7] proposed system deals with development of grid based method which is combination of image centroid zone and zone centroid zone of individual character or numerical image. Use of feed forward neural network for recognition. Complete process of Devangiri character recognition works in stages as document preprocessing, segmentation, feature extraction, classification using grid based approach followed by recognition using naive bay's NN.

Mitrakshi B. Patil , Vaibhav Narawade [8] This paper interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices for recognizing Handwritten Marathi Characters.In this paper,method for recognition of offline handwritten devnagari characters using segmentation and Artificial neural networks.

Mandeep Kaur, Sanjeev Kumar[9] This paper represent Handwritten Gurmukhi Character Recognition system using some statistical features like zone density, projection histograms , 8 directional zone density features in combination with some geometric features like area, perimeter, eccentricity, etc.Techniques like binarization, morphological operations applied to remove noise and then segmented into isolated characters. Miroslav NOHAJ[10] Created a theoretical and practical basis of preprocessing of printed text for optical character recognition using forward-feed neural networks. Demonstration application was created and its parameters were set according to results of realized experiments.

Nisha Sharma, Tushar Patnaik, Bhupendra Kumar[11] In this paper major steps of an OCR system was discussed like preprocessing, segmentation ,feature extraction, classification, postprocessing.This paper gives an overview of research work carried out for recognition of hand written English letters. Hand written letters are difficult to recognize due to diverse human handwriting style, variation in angle, size and shape of letters, also various feature extraction technique and classification method result was discussed.

Pradeep, E.Srinivasan and S.Himavathi[12] Off-line handwritten alphabetical character recognition system using multilayer feed forward neural network is given. Method like diagonal based feature extraction is

introduced for extracting the features of the handwritten alphabets. Shabana Mehfuz, Member IEEE, 2 Gauri katiyar, [13] The available Character Recognition techniques with their superiorities and weaknesses are reviewed. The Character Recognition is discussed, and directions for future research are suggested. Special attention is given to the off-line HCR since this area requires more research. Pradip Bele, Avinash Pandhare, [14] Neural networks were good at recognizing handwritten characters as these networks are insensitive to the missing data.A Backpropagation neural network is used for classification. Experimental result of this system shows that results 93%.

Fifty data sets, each containing 26 alphabets written by various people, are used for training the neural network and 570 different handwritten alphabetical characters are used for testing. Ms. Priti Rawat, Ms. Vinaya Patil[15], This system performs quite well yielding higher levels of recognition accuracy compared to the systems employing the conventional horizontal and vertical methods of feature extraction. Suitable for converting handwritten documents into structural text form and recognizing handwritten names.

3.Proposed Methodology

This section contains the block diagram and the details about the modules we will be going to use. General Procedure: Handwriting recognition is a difficult problem which includes the recognition of input is given in form of image, scan paper. The handwritten character recognition generally involves the following Modules:-

A. Image Acquisition: In the image acquisition the images for HCR system are acquired by appropriate scanning of handwritten documents, books or by capturing photographs of document. The input image is obtained by camera or through some scanner. The input image may be in gray, color.

B. Preprocessing: The method of extraction of text from the document is called preprocessing. The preprocessing Consists of a series of operations performed on the scan input image, which include background Noise reduction,image restoration, filtering etc.This system assume that the character segmented is made thin to a unit pixel thickness. Various algorithms may be used for this purpose.

C. Segmentation: This step deals with breaking of the lines, words for getting all the characters separated. This module involves the identification of the boundaries of the character and separating them for further processing for further steps. In this algorithm we will assume that this step was already done. Hence the input to our system will a single character.

D. Feature Extraction: To find a set of parameters that uniquely defines the character is called feature extraction. The feature extraction technique should be such that the features of characters should enable clear discrimination of one character from others. To distinguish a class from other class a set of features is extracted for each class. The types of feature may be of statistical, syntactical/structural or hybrid in nature.

E. Classification: This stage represents the decision making part of a recognition system and it employs the features extracted in the previous stage as inputs to the classifiers. The classifiers compare the input features with the stored features to assign a class for the input. The character recognition task is based on four approaches as template matching; statistical techniques; structural techniques and neural network. Here we are

going to use ANN, an Artificial Neural Network usually called "neural network". It is a supervised learning method.

F. Post-processing: The goal of post processing phase refers to detect and correct linguistic misspellings in the Offline HCR output text after the input image has been completely processed. Post processing steps are used to improve the accuracy of Offline handwritten character recognition system.

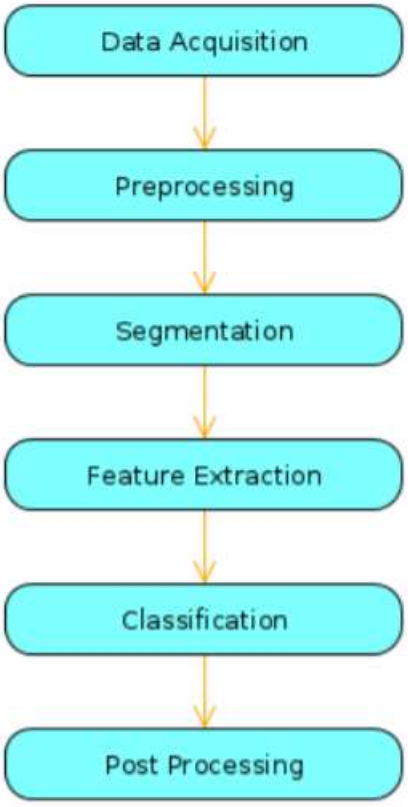


Fig1: Block diagram of the system

Naïve Bayes model

The first dataset will be trained using the Naïve Bayes training model. It is a technique which works on Bayes Theorem with a prerequisite being that the predictors are independent of each other. Thus, it can be said that the Naive Bayes classifier assumes that a particular feature in the system has nothing to do with another feature present in the system.

For example, a ball can be considered to be a cricket ball if it is red and has 3 inches of diameter approximately. If these features depend on each other they depend on any other feature. Still they contribute to the fact that the ball is cricket ball thus, it derives its name.

It is easy to build and is very useful for very large datasets.

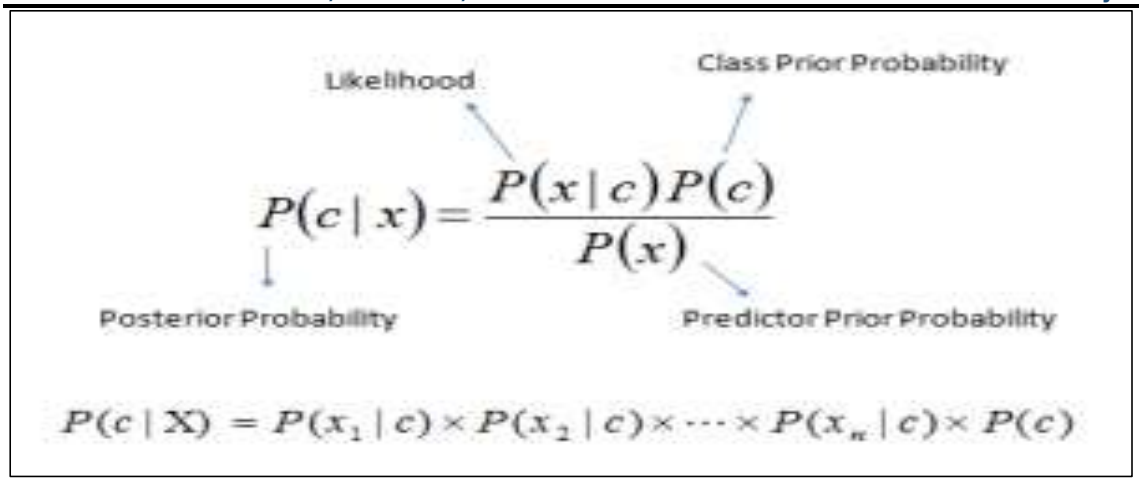


Fig:2 Naives Bayes model

where P(c|x) is the probability of target when the predictor is given

P(c) refers to the prior probability of the class

P(x|c) refers to the probability of predictor class

P(x) refers to the probability of the predictor

Random Forest Classifier model

The random forest classifier consists of individual decision trees, in which each of the individual tree make a prediction, based on which the prediction of the model is based on.

Thus, the reason of the high accuracy of this classifier is that there is majority prediction is always preferred over the individual contribution of the classifier.

Thus, in this method, the prediction of the group of the decision tree has more accuracy over the individual decision tree.

And it can be used for classification as well as regression methods.

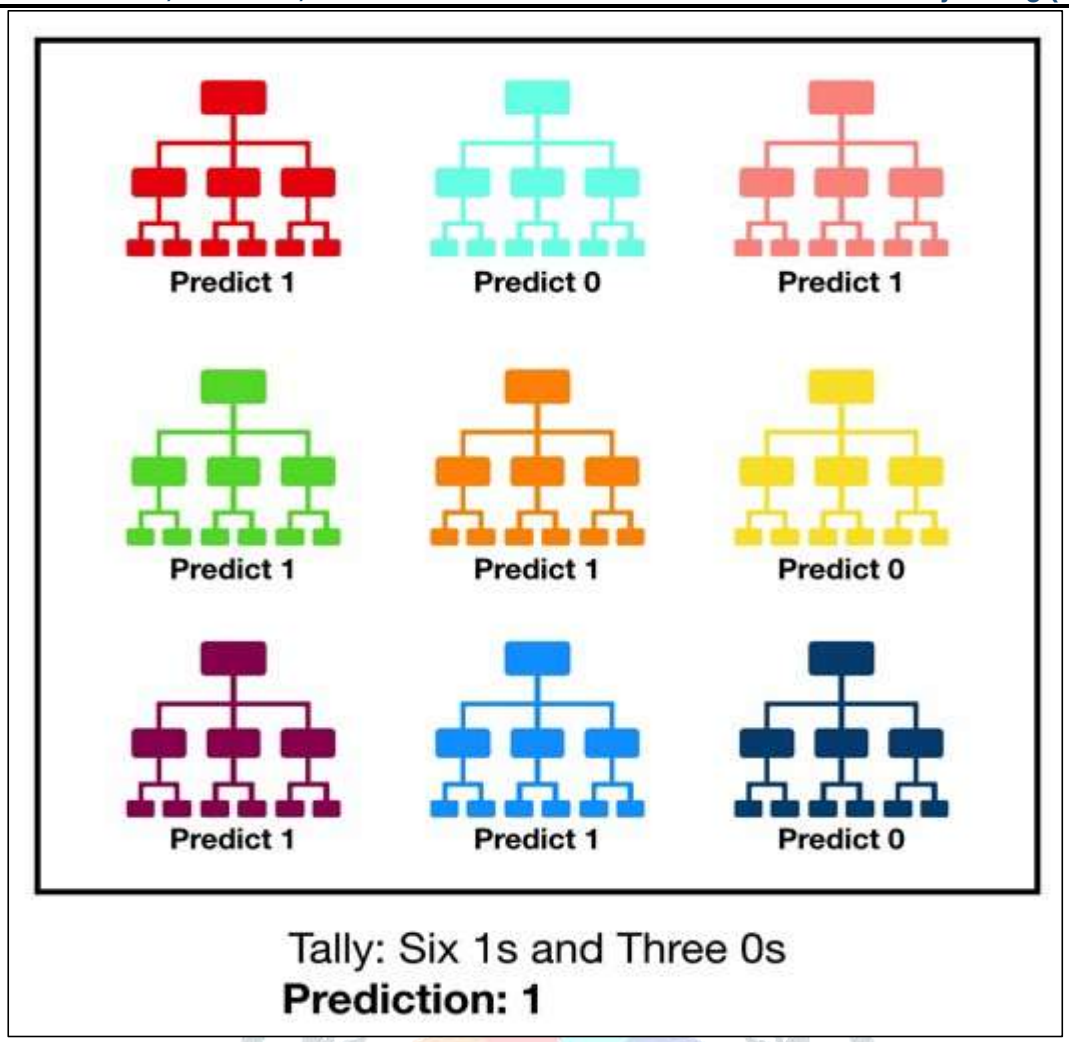


Fig 3: Random forest classifier model

4.CONCLUSIONS

Many regional languages throughout world have different writing styles which can be recognized with HCR systems using proper algorithm and strategies. We have learning for recognition of English characters. It has been found that recognition of handwritten character becomes difficult due to presence of odd characters or similarity in shapes for multiple characters. Scanned image is pre-processed to get a cleaned image and the characters are isolated into individual characters. Preprocessing work is done in which normalization, filtration is performed using processing steps which produce noise free and clean output. Managing our evolution algorithm with proper training, evaluation other step wise process will lead to successful output of system with better efficiency. Use of some statistical features and geometric features through neural network will provided better recognition result of English characters. The existing handwritten has very low accuracy. We need an efficient solution to solve this problem so that performance can be increased. This work will be helpful to the researchers for the work towards other script.

6.Result

Recall

The term recall is one of the values which help in determining the performance of the model and is one of the important characteristics which help in comparing the classifiers.

It refers to the measure of the model correctly identifying the true positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive+False\ Negative}$$

Accuracy

	precision	recall	f1-score	support
0	1.00	0.97	0.99	38
1	0.98	0.98	0.98	43
2	0.95	1.00	0.98	42
3	0.98	0.96	0.97	46
4	0.97	1.00	0.99	37
5	0.98	0.96	0.97	49
6	1.00	1.00	1.00	52
7	1.00	0.96	0.98	50
8	0.94	0.98	0.96	46
9	0.98	0.98	0.98	47
accuracy			0.98	450
macro avg	0.98	0.98	0.98	450
weighted avg	0.98	0.98	0.98	450

The term accuracy is one of the values which help in determining the performance of the model and is one of the important characteristics which help in comparing the classifiers.

It gives the ratio of the number of correct predictions to the total number of predictions. Generally, accuracy is taken as the most important factor which is used to compare the performance of the classifiers.

But, taking the other parameters such as recall, precision into account is the best way to compare the performance of the classifiers.

$$\text{Accuracy} = \frac{True\ Positive+True\ Negative}{True\ Positive+False\ Negative+False\ Positive+False\ Negative}$$

F1 score

The term accuracy is one of the values which help in determining the performance of the model and is one of the important characteristics which help in comparing the classifiers.

It involves the relationship between the recall and the precision.

$$\text{f1 score} = 2 * \frac{Precision * Recall}{Precision+Recall}$$

Now, the performance of the model depends on factors other than accuracy, such as the precision and recall. Thus, having a high f1 score automatically ensures a high precision and recall. Thus, it is also one of the most important parameters in determining the efficiency of the classifier

Precision

The term precision is one of the values which help in determining the performance of the model and is one of the important characteristics which help in comparing the classifiers. It is the ratio of the true positive and all the value of true positive.

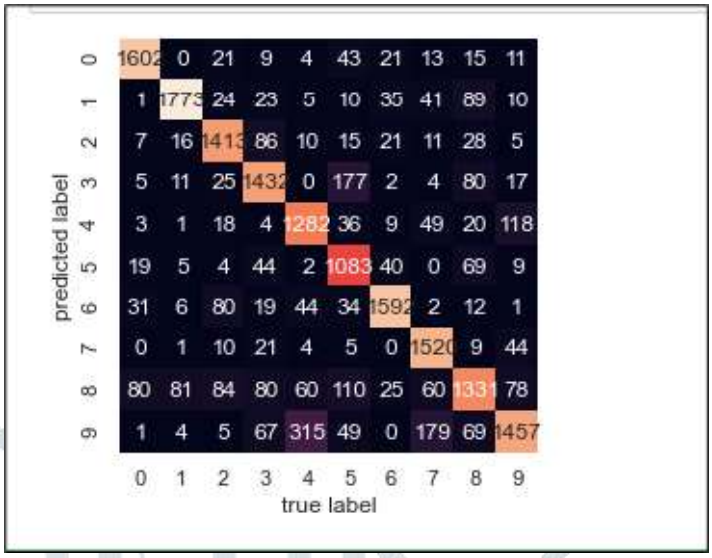
$$\text{Precision} = \frac{True\ Positive}{True\ Positive+False\ Positive}$$

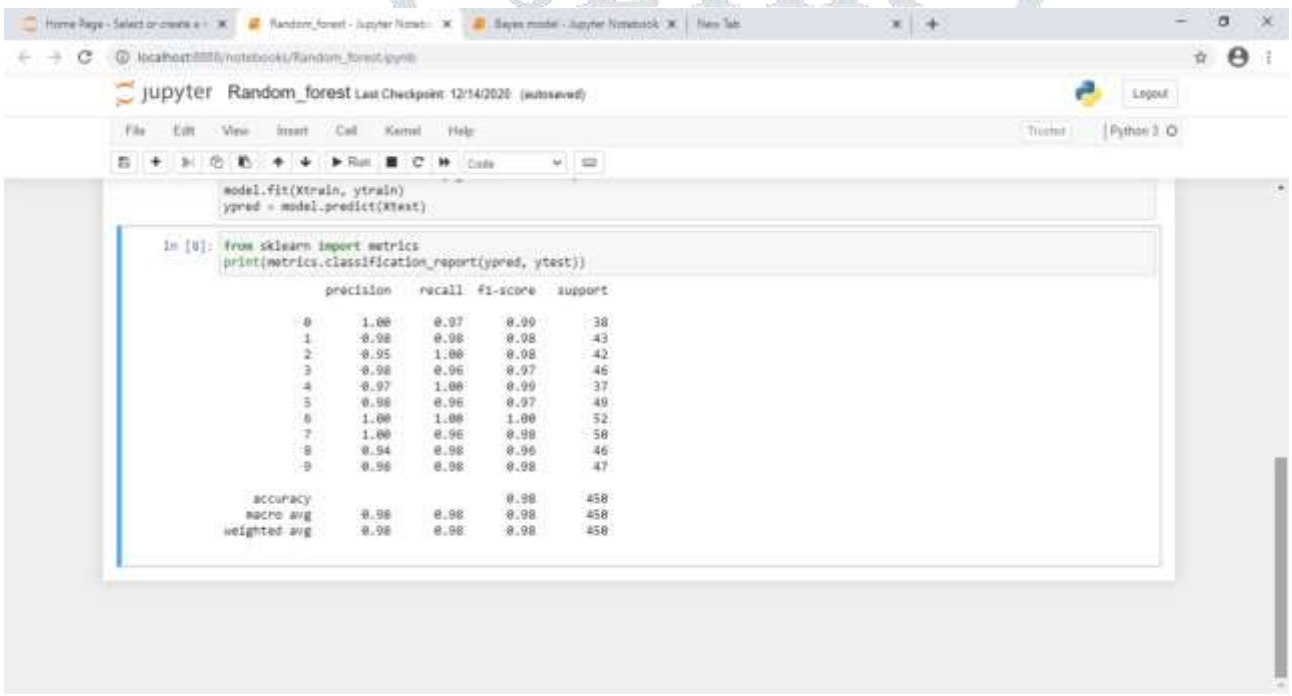
Random Forest Classifier

1. Accuracy: 0.98
2. Precision: 0.98

3. f1 score: 0.98
4. Recall: 0.98
5. Support: 450

Fig 4: Result





jupyter Random_forest Last Checkpoint: 12/14/2020 (autosaved)

```
model.fit(Xtrain, ytrain)
ypred = model.predict(Xtest)

In [8]: from sklearn import metrics
        print(metrics.classification_report(ypred, ytest))
```

	precision	recall	f1-score	support
0	1.00	0.97	0.99	38
1	0.98	0.98	0.98	43
2	0.95	1.00	0.98	42
3	0.98	0.96	0.97	46
4	0.97	1.00	0.99	37
5	0.98	0.96	0.97	49
6	1.00	1.00	1.00	52
7	1.00	0.96	0.98	50
8	0.94	0.98	0.96	46
9	0.96	0.98	0.98	47
accuracy			0.98	450
macro avg	0.98	0.98	0.98	450
weighted avg	0.98	0.98	0.98	450

Fig 5

Naive Bayes Model

1. Accuracy: 0.821
2. Precision: 0.83
3. f1 score: 0.83
4. Recall: 0.82
5. Support: 17500

Fig 6

	precision	recall	f1-score	support
0	0.92	0.91	0.91	1785
1	0.88	0.93	0.91	1927
2	0.88	0.84	0.86	1766
3	0.81	0.81	0.81	1768
4	0.82	0.75	0.79	1667
5	0.84	0.66	0.74	1552
6	0.88	0.91	0.89	1715
7	0.95	0.82	0.88	1887
8	0.66	0.76	0.71	1707
9	0.68	0.83	0.75	1726
accuracy			0.83	17500
macro avg	0.83	0.82	0.82	17500
weighted avg	0.83	0.83	0.83	17500

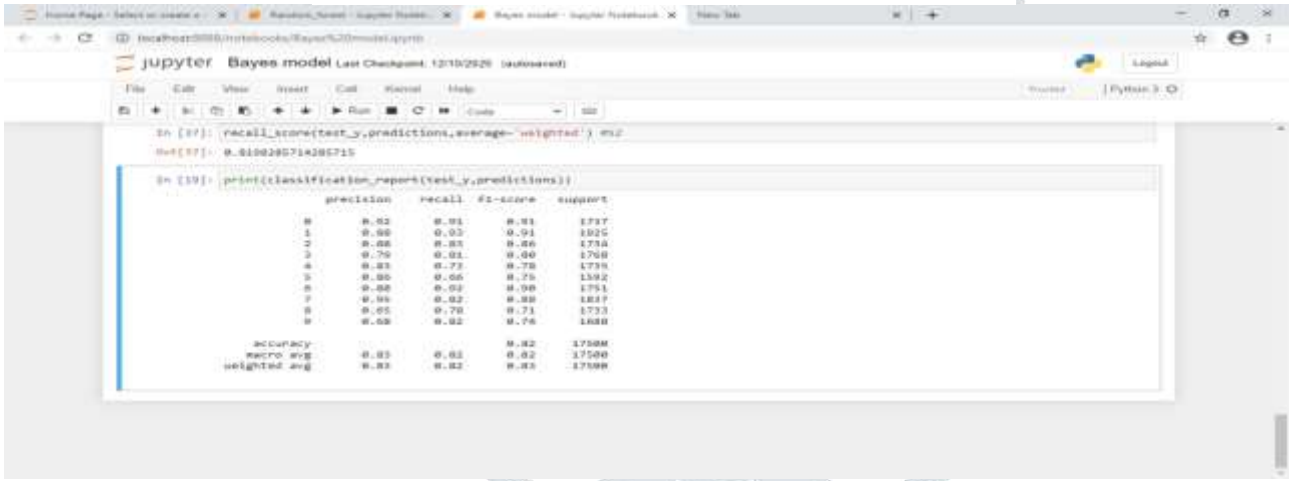


Fig 7: Result

It is clearly visible that the Random Forest Classifier performs better than the the Naïve Bayes classifier.The accuracy produced by the Random Forest Classifier is 98% whereas the accuracy produced by the Naïve Bayes model is 83%. Thus, it can be deduced that the number of correct predictions made by the Random Forest Classifier was higher than those predicted by the Naïve Bayes classifier.In addition to that, another parameter, that is the f1 score, which is one of the most important parameter is 0.98 in both macro and weighted average in Random Forest Classifier, whereas it is 0.82 and 0.83 in macro, weighted average respectively in case of Naïve Bayes classifier. Thus, the f1 score in case of Random Forest Classifier is greater than in the Naïve Bayes.And as the f1 score is determined by the combination of recall and precision, both the parameters will again be greater in the Random Forest Classifier (RFC) as compared to the Naive Bayes classifier.And as can be seen, the precision in case of RFC, the precision is 0.98 whereas it is 0.83 in case of Naïve Bayes classifier. And also, the recall in RFC is greater than in Naïve Bayes classifier.The only parameter where Naïve Bayes has a greater value than the RFC model is the support, which is 17500 and 450 in the case of RFC model**.**

7.REFERENCES

[1] Isha Vats, Shamandeep Singh, "Offline Handwritten English Numerals Recognition using Correlation Method", International Journal of Engineering Research and Technology (IJERT): ISSN: 2278-0181 Vol. 3 Issue 6, June 2014.

[2] Gunjan Singh,Sushma Lehri, " Recognition of Handwritten Hindi Characters using Back propagation Neural Network", International Journal of Computer Science and Information Technologies ISSN 0975-9646, Vol. 3 (4) , 2012,4892-4895.

[3] S S Sayyad, Abhay Jadhav, Manoj Jadhav, Smita Miraje, Pradip Bele, Avinash Pandhare, "Devnagiri Character Recognition Using Neural Networks", International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 1, July 2013.

[4] Shabana Mehfuz, Member IEEE, 2 Gauri katiyar, "Intelligent Systems for OffLine Handwritten Character Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering Volume 2, Issue 4, April 2012.

[5] Prof. Swapna Borde, Ms. Ekta Shah, Ms. Priti Rawat, Ms. Vinaya Patil, "Fuzzy Based Handwritten Character Recognition System", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, VNCET 30 Mar'12.

[6] Rahul KALA, Harsh VAZIRANI, Anupam SHUKLA and Ritu TIWARI, "An Overview of Character Recognition Focused on OffLine Handwriting", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART APPLICATIONS AND REVIEWS, VOL. 31, NO. 2, MAY 2001.

[7] Ms. Seema A. Dongare, Prof. Dhananjay B. Kshirsagar, Ms. Snehal V. Waghchaure , "Handwritten Devanagari Character Recognition using Neural Network", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727 Volume 16, Issue 2, Ver. X (Mar-Apr. 2014), PP 74-79.

[8] Mitrakshi B. Patil, Vaibhav Narawade, "Recognition of Handwritten Devnagari Characters through SegMentation and Arti_cial neural networks", International Journal of Engineering Research and Technology (IJERT) Vol. 1 Issue 6, August - 2012. ISSN: 2278-0181.

[9] Mandeep Kaur, Sanjeev Kumar, "A RECOGNITION SYSTEM FOR HANDWRITTEN GURMUKHI CHARACTERS", International Journal of Engineering Research and Technology (IJERT) Vol. 1 Issue 6, August - 2012 ISSN: 2278-0181.

[10] Miroslav NOHAJ, Rudolf JAK_A A, "Image preprocessing for optical character recognition using neural Networks", Journal of Patter Recognition Research, 2011.

[11] Nisha Sharma et al, "Recognition for handwritten English letters: A Review"International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013. [12] J.Pradeep et al, "Diagonal based feature extraction for handwritten alphabets recognition System using Neural network", International Journal of Computer Science and Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[12] Rahul KALA, Harsh VAZIRANI, Anupam SHUKLA and Ritu TIWARI, "An Overview of Character Recognition Focused on OffLine Handwriting", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART APPLICATIONS AND REVIEWS, VOL. 31, NO. 2, MAY 2001.

[13] Shabana Mehfuz, Member IEEE, 2 Gauri katiyar, "Intelligent Systems for OffLine Handwritten Character Recognition: A Review", International Journal of Emerging Technology and Advanced Engineering Volume 2, Issue 4, April 2012

[14]Pradip Bele, Avinash Pandhare, "Devnagiri Character Recognition Using Neural Networks", International Journal of Engineering and Innovative Technology (IJEIT)

[15] Ms. Priti Rawat, Ms. Vinaya Patil, "Fuzzy Based Handwritten Character Recognition System", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,