**HANDWRITING LETTER PREDICTION**

A PROJECT REPORT

*Submitted by*

**B Sai Abhishek BL.EN.U4AIE21015**

**B Ruchith Balaji BL.EN.U4AIE21017**

**Chillakuru Hari BL.EN.U4AIE21038**

***for the course***

***21AIE205- Python for Machine Learning***

***Guided and evaluated by***

***Niharika Ma’am***

***Dept. of CSE,***



**AMRITA SCHOOL OF ENGINEERING, BANGALORE**

**AMRITA VISHWA VIDHYAPEETHAM**

**BANGALORE-560 035**

January 2023

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **TOPIC** | **Page No** |
| 1 | Abstract | 3 |
| 2 | Introduction | 4 |
| 3 | Literature Survey | 7 |
| 4 | System Model | 9 |
| 5 | Implementation | 11 |
| 6 | Sample Code | 13 |
| 7 | Sample Output | 15 |
| 8 | Conclusion | 16 |
| 9 | References | 17 |

**ABSTRACT:**

Letter recognition has been an active and challenging area. Letter recognition

system plays a very important role in today’s world. At present time it is very difficult to find the correct meaning of Lettered documents as different people have different ways of writing digits or texts.

There are many areas where we need to recognize the words, alphabets and digits like postal addresses, bank cheques etc.

The main aim of this project is to predict the Alphabet corresponding to the Attributes of the letter which are given in the Data set.

Output of the project that allows users to predict the digits drawn on the interface using machine learning and shows the results in percentage how likely the prediction is correct. The project will be built using python and the application will be trained with the help of dataset from Kaggle using K-Nearest Neighbour, Support Vector Machines, Decision trees. Upon all which has the Highest Accuracy, then that Algorithm is preferred to predict the Letters.

* ***Purpose of Study:***

The purpose for this project is to allow us to explain the implementation of the Machine learning models and how the models use the dataset to analyze and predict the outcome.

This project allows us to study the K-Nearest Neighbor, Support Vector Machines,

Decision trees of the Letter Recognizer and the working of the model to see how the Machine Learning process works in the background. It shows the working of how the number fed into the model is recognized and displays the number that is recognized after analysis and comparing it with the dataset fed to the model. With the help of the Letter Recognition dataset and the K-Nearest Neighbor, Support Vector Machines, Decision trees. The project can explain the purpose, working and the benefits of Machine learning in our daily lives.

* ***Algorithms and Methods in Letter Recognition***
* Convolutional Neural network.
* Semi Incremental Recognition.
* Incremental Recognition.
* Line and word segmentation.
* Part-based method.
* Slope and Slant correction method.
* Ensemble method.

**INTRODUCTION:**

To make this model, the data set needs to be imported which is very important in the analysis and recognizing the digits which are fed into the model. The dataset used in the project is called the Letter Recognition dataset.

The Letter Recognition is a large dataset of Letters with different attributes like size of letter, height of the letter, how much space needed to fit, Width of the letter, space needed for edge and etc… Machine Learning models for training various Recognition systems. This dataset contains 20,000 letters and each letter has its own attributes as mentioned above.

***Making a Model:***

**Training Data**

**Training for ML Algorithm**

**Model Input**

**Input Data**

**ML Algorithm**

**Prediction**

**Successful Model**

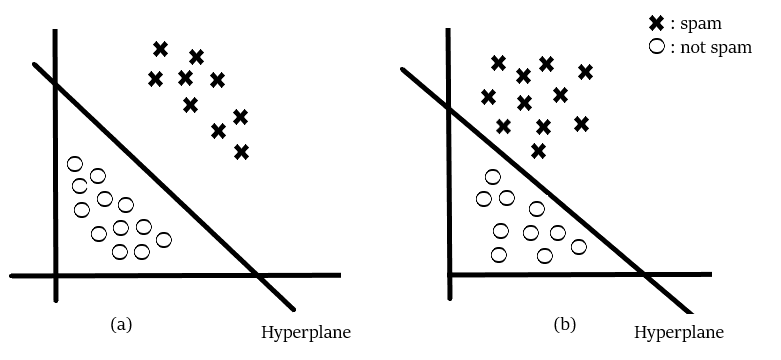
***The Algorithms that are going to use are:***

1. K-Nearest Neighbor.
2. Support Vector machines.
3. Decision Trees.
4. **K-Nearest Neighbor:**

* From the given Data set find dependent and independent Variables.
* Divide the variables to training and testing data, it means Data pre-processing.
* Fit the train and test data.
* Using the KNN Classifier taking n neighbors find the predicted for test data
* From predicted data find Accuracy, Precision, Recall, F1-Score etc.
* Specifically, four different distance functions, which are Euclidean distance, cosine similarity measure, correlation, and Chi square, are used in the k-NN classifier respectively

*Fig-1*

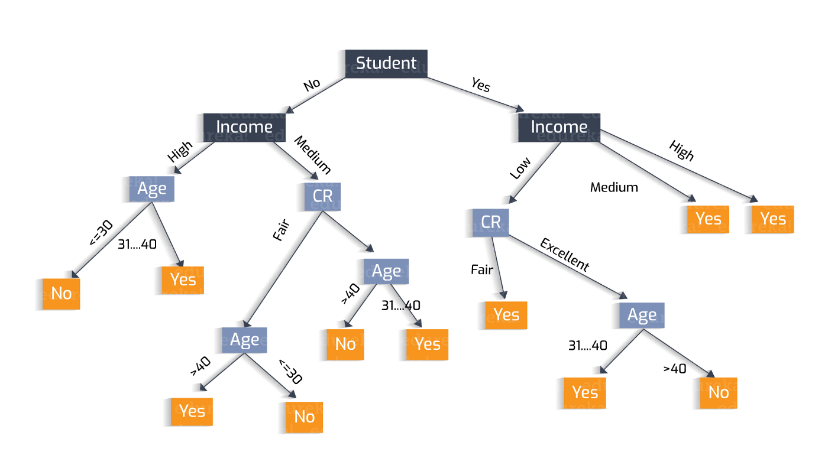
1. **Support Vector Machines:**

* From the given Data set find dependent and independent Variables.
* Divide the variables to training and testing data, it means Data pre-processing.
* Fit the train and test data.
* Using the SVC Classifier taking path as linear, curve find the predicted for test data
* From predicted data find Accuracy, Precision, Recall, F1-Score etc.
* Three different types of SVM-Kernels are displayed below. The polynomial and RBF are especially useful when the data-points are not linearly separable.

*Fig-2*

1. **Decision Trees:**

* From the given Data set find dependent and independent Variables.
* Divide the variables to training and testing data, it means Data pre-processing.
* Fit the train and test data.
* Using the Decision Tree Classifier taking Criterion as entropy find the predicted for test data
* From predicted data find Accuracy, Precision, Recall, F1-Score etc.



*Fig-3*

**LITERATURE SURVEY:**

In this paper[1] we will using three (3) classification to recognize the handwritten which is SVM, KNN and Neural Network. There are many research have been done regarding the handwriting recognition in various field.

Youssouf Chherawala, Partha Pratim Roy and Mohamed Cheriet in their paper “Feature Set Evaluation for Offline Handwriting Recognition Systems: Application to the Recurrent Neural Network” stated that handwriting recognition system is dependent on the features extracted from the word image. There are various method to extract the features but there are no method that have been proposed to identify the most promising of these other than a straightforward comparison based on the recognition rate. So they propose a framework for feature set evaluation based on a collaborative setting.

They use a weighted vote combination of recurrent neural network (RNN) classifier. They quantify the importance of feature sets through the combination weights, which reflect their strength and complementarity.

* <https://gurus.pyimagesearch.com/lesson-samplek-nearest-neighbor-classification>

This paper[2] proposes machine learning and deep learning techniques for recognizing Arabic letter. Finally, this paper introduces a comparative study between them in term of their performance. Actually, the classification of Arabic handwritten text plays a vital role in the computer vision domain, where traditional machine learning techniques and deep learning techniques are commonly used by researchers. In this paper, both machine learning and deep learning techniques are proposed and evaluated for recognizing Arabic handwritten text. Several experiments were carried out using both machine learning and deep learning on two different databases the AHCR and ADBase. The AHCR contains 28000 images of handwritten Arabic alphabet letters written by 100 writers. While the ADBase contains 70,000 images of handwritten Arabic digits written by 700 writers. The experimental results on both databases have demonstrated that the performance of the deep learning outperforms machine learning.

From the above paper we have got idea about this model. With the dimensions of letters written by different persons (20000). Creating a model and implementing the algorithm gives you the correct prediction of letter as output with given dimensions.

* [2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)](https://ieeexplore.ieee.org/xpl/conhome/9350668/proceeding)

As we can see from the result, KNN and SVM predict correctly the dataset but for MLP Neural Network that is some mistake to predict the number 9. This is because for KNN and SVM it predict directly from the feature extraction. But for MLP, it is a non-linear function. So it more suitable for learn non-linear models.

And MLP with hidden layers have nonconvex loss function where there exists more than one local minimum. Therefore different random weight initializations can lead to different validation accuracy. But it can improve by using Convolutional Neural Networks with Keras.

This paper[3] presents a comprehensive review of Letter Recognition (HCR) in English language.The handwritten character recognition has been applied in variety of applications like Banking sectors, Health care industries and many such organizations where handwritten documents are dealt with.

Letter Character Recognition is the process of conversion of handwritten text into machine readable form. For handwritten characters there are difficulties like it differs from one writer to another, even when same person writes same character there is difference in shape, size and position of character.

Latest research in this area has used different types of method, classifiers and features to reduce the complexity of recognizing handwritten text.

* [Recognition System Integrating Feature Extraction and Classification for English Handwritten](https://www.researchgate.net/publication/284612377_Neural_Network_Based_Recognition_System_Integrating_Feature_Extraction_and_Classification_for_English_Handwritten).

In this paper[4] a detailed review of Handwritten Character Recognition is presented. Some features of human beings are unique to individuals like iris, fingerprint, DNA etc. Handwriting is one such feature which is different for each human being and it has been proven scientifically.

In Handwritten letter Recognition (HCR) the task is to identify the characters written by humans and converting it into digital text. HCR is a field where plenty of research has been done but still there is scope in terms of improving the accuracy and efficiency.

Digitizing manually written text is very useful in today's world as it makes information readily available anywhere and anytime. Digitized text can be used for commercial purposes and it is more safe and environment friendly as compared to manual text.This review paper will shed some light on various methodologies used till now in this field along with their advantages, limitations and accuracy rate.

* [2020 International Conference on Communication and Signal Processing (ICCSP)](https://ieeexplore.ieee.org/xpl/conhome/9177229/proceeding)

**SYSTEM MODEL:**

***Making a Model:***

**Training Data**

**Training for ML Algorithm**

**Model Input**

**K NN**

**Input Data**

**SVM**

**ML Algorithm**

**Decision Tree**

**Prediction**

**Successful Model**

*SYSTEM-MODEL*

***Brief Description:***

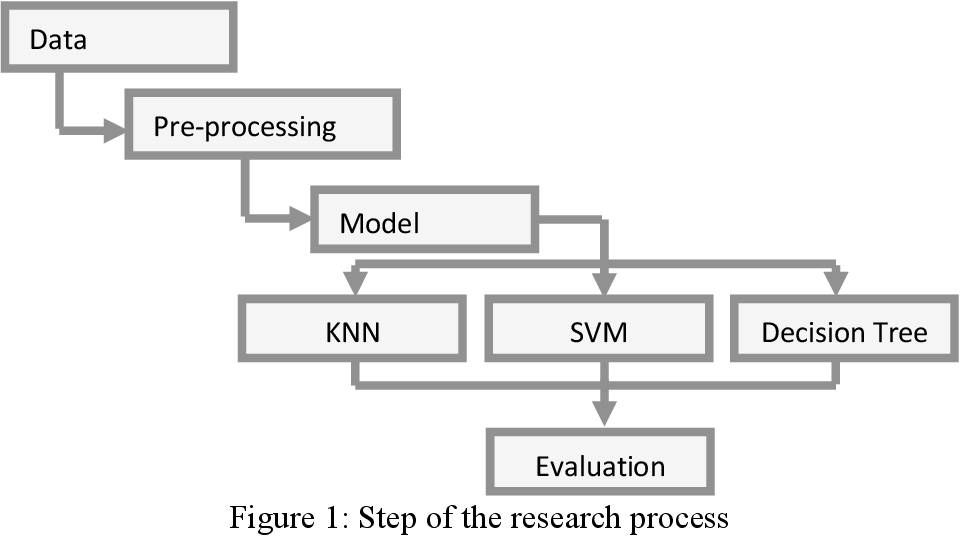
* Firstly, the given Data is cleaned and removing the nan values and zeros to not get errors.
* After the data is divided into two sets

1. Train data set
2. Test data set

* Train data set used to train the model to get accurate results.
* Test data set is used to predict the output.
* Train and test data are divided in the ratio 8:2 or 7:3. (8:2 is recommended)
* Train data sets are preprocessed to fit into array. It is easy for machine to predict.
* Using classifiers the algorithm is implemented. The classifiers are:

1. K Nearest Neighbor
2. SVC
3. Decision Tree classifier

* After the applying of ML algorithm the output array is stored in predicted.
* Using the predict function the test data is passed into it.
* Then our Model is successfully created.
* Confusion matrix can be created.
* With that model we can calculate accuracy, precision, recall, F1 score.



*Fig-4*

**IMPLEMENTATION**

For this model creation we have used Anaconda Navigator. In that Jupyter Notebook is used to create the model. Jupyter Notebook allows users to compile all aspects of a data project in one place making it easier to show the entire process of a project to your intended audience.

***Library that are used:***

1. Sklearn.model\_selection
2. Sklearn.preprocessing
3. Sklearn.metrics
4. Matplotlib.pyplot
5. Sklearn.neighbors
6. Sklearn.svm
7. Sklearn.tree

***Short Description:***

1. ***Sklearn.model\_selection:***

sklearn. model\_selection . train\_test\_split. Split arrays or matrices into

random train and test subsets.

***2)Sklearn.preprocessing:***

The sklearn. preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. In general, learning algorithms benefit from standardization of the data set.

***3)Sklearn.metrics:***

The sklearn. metrics module implements several loss, score, and utility functions to measure classification performance. Some metrics might require probability estimates of the positive class, confidence values, or binary decisions values.

***4)Matplotlib.pyplot:***

Matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

***5)Sklearn.neighbors:***

sklearn. neighbors provides functionality for unsupervised and supervised neighbors-based learning methods. Unsupervised nearest neighbors is the foundation of many other learning methods, notably manifold learning and spectral clustering

***6)Sklearn.svm:***

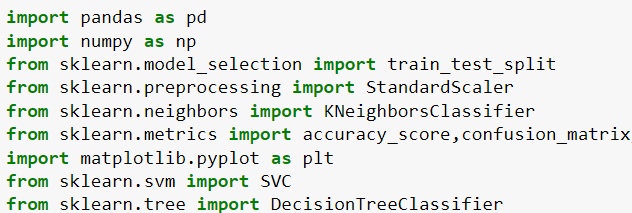
SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin.

***7)Sklearn.tree:***

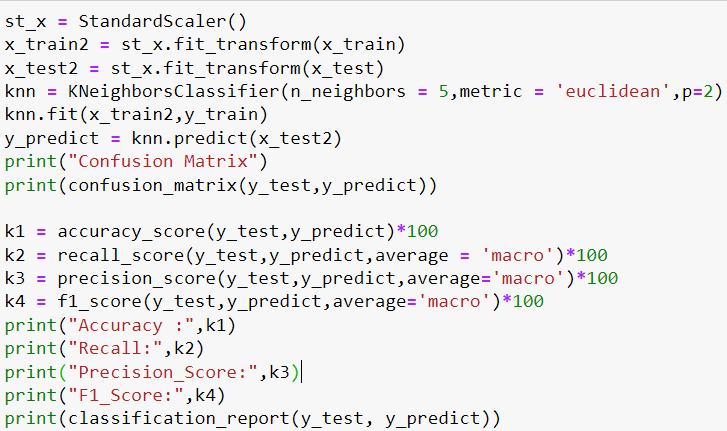
sklearn.tree .DecisionTreeClassifier· The function to measure the quality of a split. · The strategy used to choose the split at each node.

**SAMPLE CODE:**

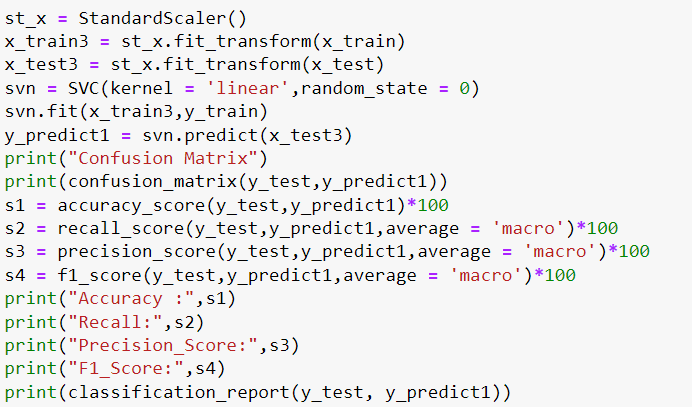
***Importing Libraries:***

****

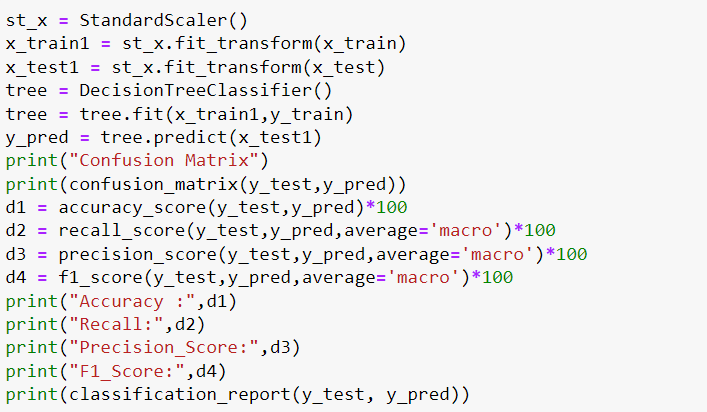
***K-Nearest Neighbor:***

****

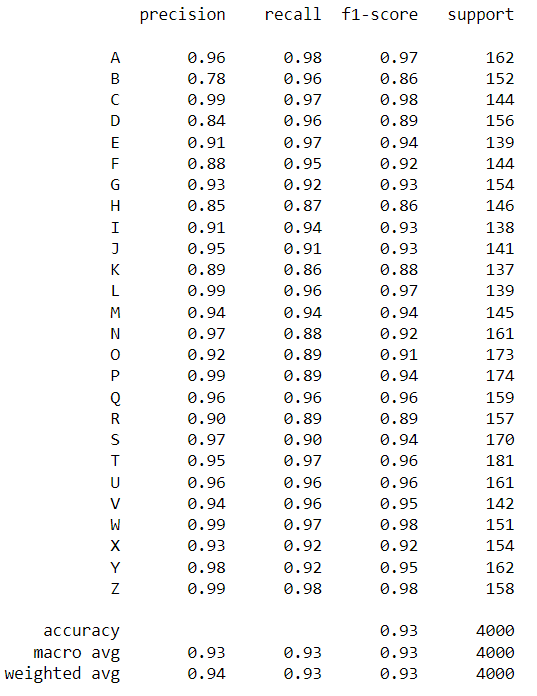
***Support Vector Machines:***

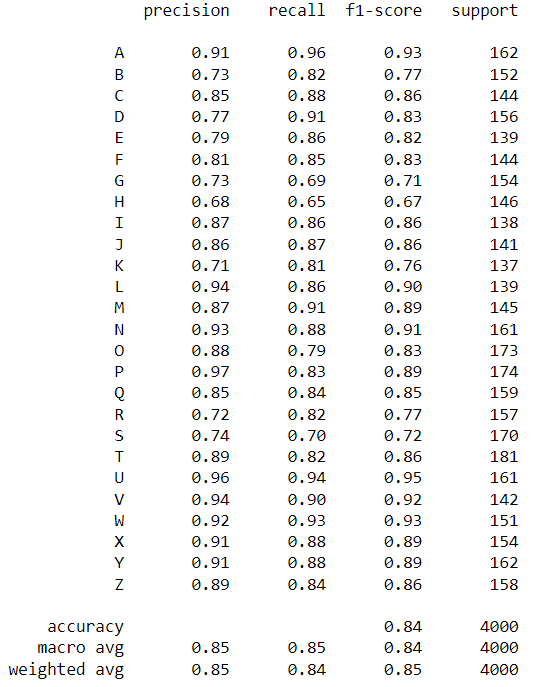
****

***Decision Trees:***

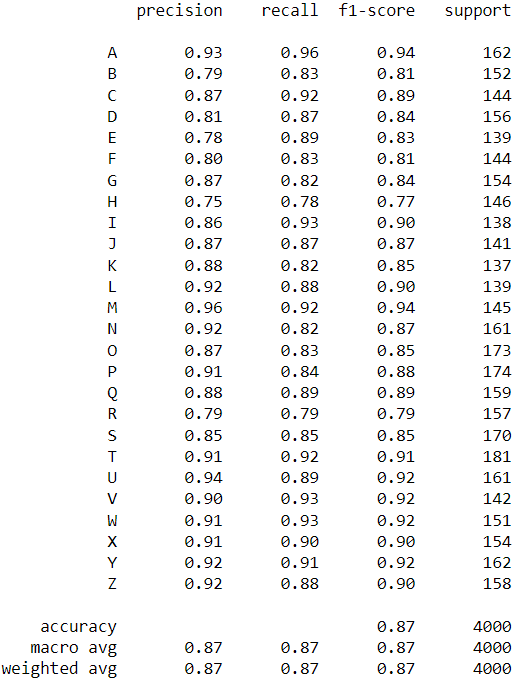
****

**SAMPLE OUTPUT:**

***K-Nearest Neighbor:***

***Support Vector Machines:***

***Decision Trees:***

****

**CONCLUSION:**

**REFERENCES:**

**Conference**

**[1]** [Ayush Kumar Agrawal](https://ieeexplore.ieee.org/author/37088893270), [A.K. Shrivas](https://ieeexplore.ieee.org/author/37088895487), [Vineet kumar Awasthi](https://ieeexplore.ieee.org/author/37088896616), Department of IT & CS,

Dr. C.V. Raman University, Bilaspur, (C.G.), India, A Robust Model for Handwritten

Digit Recognition using Machine and Deep Learning Technique, 2021 2nd International

Conference for Emerging Technology (INCET).

**[2]** [Anupama Sahu](https://ieeexplore.ieee.org/author/37088763331);[S. N. Mishra](https://ieeexplore.ieee.org/author/37086219659) [2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)](https://ieeexplore.ieee.org/xpl/conhome/9358465/proceeding)

# **[3]** [Najoua Rahal](https://ieeexplore.ieee.org/author/37086038137), [Maroua Tounsi](https://ieeexplore.ieee.org/author/37085748211), [Tarek M Hamdani](https://ieeexplore.ieee.org/author/37597160600), [Adel M Alimi](https://ieeexplore.ieee.org/author/37270046700), Handwritten Words and Digits Recognition using Deep Learning Based Bag of Features Framework, [2019 International Conference on Document Analysis and Recognition (ICDAR)](https://ieeexplore.ieee.org/xpl/conhome/8961318/proceeding)

**Journal**

**[3]** *Alkhateeb Jawad et al.,*Performance of hidden Markov model and dynamic Bayesian network classifiers of handwritten Arabic word recognition. Knowledge-Based Systems*, vol. 24, pp. 680-688, 2011.*

**[4]** H Alkhateeb Jawad, "Off-Line Arabic Handwritten Isolated Character

Recognition", International Journal of Engineering Science and Technology (IJEST), vol. 7, no. 7, pp. 251-257, 2015.