# Digit Recognition ensemble learning

**Your Name here**

**Your Student ID**

**Semester**

Contents

[Digit Recognition ensemble learning 1](#_Toc101385423)

[Introduction: 3](#_Toc101385424)

[Objective: 3](#_Toc101385425)

[Methodology: 4](#_Toc101385426)

[Data Description: 4](#_Toc101385427)

[Analysis of Results/Findings: 4](#_Toc101385428)

[Data Splitting: 6](#_Toc101385429)

[Training models without parameters: 6](#_Toc101385430)

[Accuracy Graph: 8](#_Toc101385431)

[Ensemble the Models: 9](#_Toc101385432)

[Conclusion: 9](#_Toc101385433)

# Introduction:

The MNIST database is a huge data set of transcribed digits that are generally utilized for preparing different picture handling frameworks. The data set is additionally generally utilized for preparing and testing in the field of AI and machine learning. It was made by "re-blending" the examples from NIST's unique datasets. The MNIST dataset is an acronym that stands for the Modified National Institute of Standards and Technology dataset. It is a dataset of 60,000 small square 28×28-pixel grayscale images of handwritten single digits between 0 and 9.

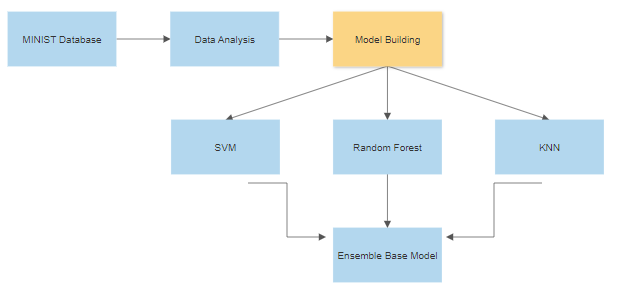
We downloaded the dataset from the internet, the goal is to train three different machine learning classification models to see their accuracies.

# Objective:

This appraisal project looks at the introduction of three AI classifier models, explicitly, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Random Forest with propensity help. These different systems give full outcomes yield with worked on computational benefit. In this paper, we are utilizing the going with three strategies classifiers, for example, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Random Forest. After this, we get to the game-plan of preparing information that further structures the exactness of the procedure with regarded classifiers. We mark straightforwardly off the bat with SVM classifiers; its outcomes appear with a better than normal presentation result of 97.27%. By then, we check with Random-Forest classifiers in this way shows the result with a customary showcase outcome of 91.80% and KNN gave us the accuracy score of 96.95%. The goal after training the models is to gather all of them into one base model and train the base model again to see if we can get the accuracy higher than the previous models gave. We got an accuracy of 97.79, which is slightly higher than SVM. So we can say that the ensemble model gave us the highest accuracy.

# Methodology:

The overall methodology of this project is given below:



Flow chart of the process

The overall process is to perform analysis on the MINIST dataset, then we build three machine learning models and gathered them into one base model.

# Data Description:

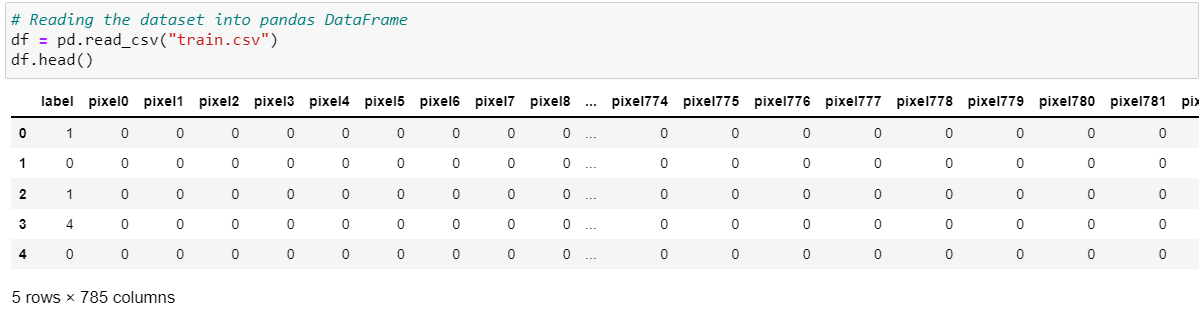
The MNIST database is a huge data set of transcribed digits that are generally utilized for preparing different picture handling frameworks. The data set is additionally generally utilized for preparing and testing in the field of AI and machine learning.

# Analysis of Results/Findings:

In this section, we will look at all the code we have performed and show the results we get. As the goal of this project is to perform the data analysis on the MINIST dataset and train three different ensemble models, after that we will use the ensemble learning technique to gather all the trained models into one base model.



The first step is to import all the important libraries and the required models. We are importing NumPy, Pandas, and our three ensemble models and a base Voting classifier.



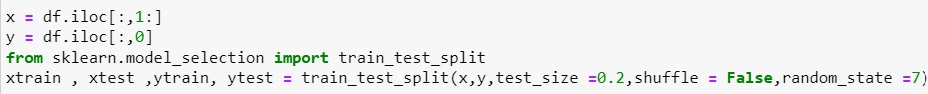
The shape of the data is shown above, we have columns as pixels and rows are the observations. We are storing the data in a Pandas Data Frame.



After loading the data into pandas Data Frame, we are visualizing the first digit graphically, we can the number 4 in the above graph.

## Data Splitting:

The next step is to split the data into testing and training sets. The training set will be used to train the models and the testing set will be used to test our models and perform the predictions on our data. We are splitting the data by using the train set split method.



Here we are first splitting the independent and dependent variables into X and Y variables then splitting them further into test and train sets. We are splitting the data in the ratio of 80:20, which means that 80% data will be used to train the models and 20% will be used to test the models.

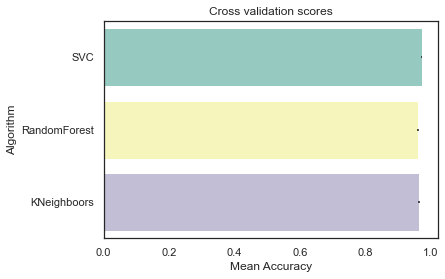
## Training models without parameters:

We will train our models on the training data and perform the predictions using the testing dataset, we will first train the models using no parameters and then we’ll perform the parameter tuning technique to train the following models on their suitable parameters.

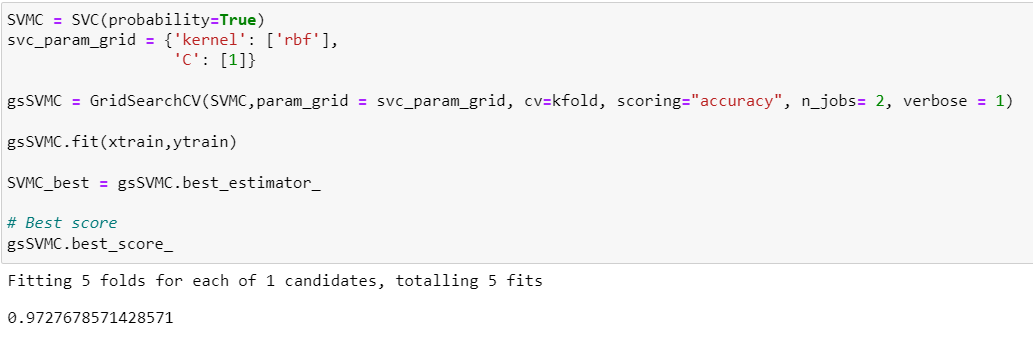
* SVM
* Random Forest
* KNN



We are using the K-Fold technique to train the three models and plot their accuracy graph.



The above is the accuracy graph of our selected models, we can see that the accuracy of these three models is above 90%. This is the accuracy without any parameter tuning. We will train these models again by using the GridSearchCV to find the suitable parameters and will train the Random Forest and SVM again on the suitable parameters.

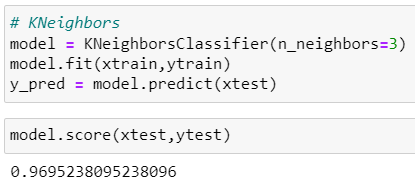


In Machine learning, support vector machines are managed and supervised learning models with related learning calculations that investigate information for grouping and relapse examination. We performed parameters tuning on SVM by using the GridSearchCV and got an accuracy of 97.27%.

Now let’s train the Random Forest again.



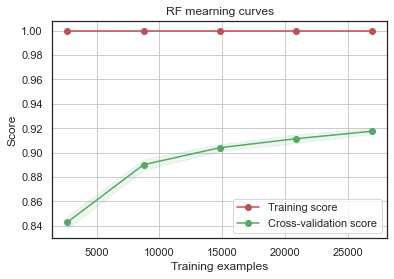
The accuracy of Random Forest after GridSearchCV is 91.80%.



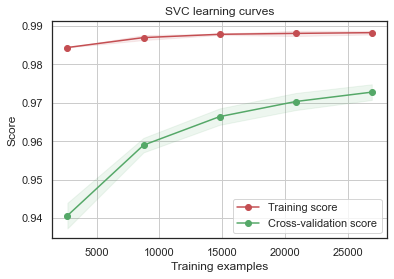
We are now training the KNN manually bypassing the n\_neighbors = 3. This model gave us the highest accuracy of 96.95%. we can say that KNN performed very well on our data.

## Accuracy Graph:

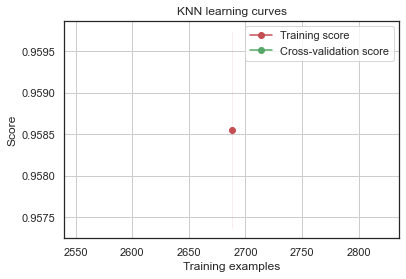
The following is the accuracy curve graph of Random Forest on the training data and validation data.



The following is the accuracy of SVM shown graphically.

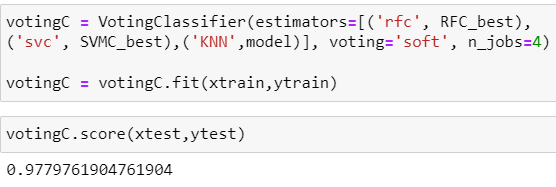


Accuracy curve graph of KNN on training and validation data.



## Ensemble the Models:

We started the project by training different ensemble models like Random Forest, SVM, and KNN, the next step is to ensemble them into a base model and train the base model again. An ensemble is a machine learning model that combines the predictions from two or more models. The models that contribute to the ensemble, referred to as ensemble members, maybe the same type or different types and may or may not be trained on the same training data.



We selected the Voting classifier as the base classifier and train the model again. We got an accuracy of 97.79% which is the highest accuracy we got through the journey.

# Conclusion:

As AI and machine learning estimations are used like K-NN, SVM, and RF are close by different boundaries and feature scaling vectors. We moreover noticed a substitute assessment among the methodologies classifiers in regards to features of precision and timing. Precision can change as it depends upon the use of explicit planning and testing data, and this can moreover be improved if the amount of it is given to get ready and testing data. Every classifier has its precision and time use. In this field, we can track down the way that if the force of the Computer chip changes to GPU, the classifier can perform with better precision and less time and improved results can be noticed. Notwithstanding, not all physically composed digit sets are normalized in size, or centered and set aside progressively as 28x28 pixel pictures in grayscale in the certified cases.