**Report Ensemble Learning**

**Background of Ensemble Learning**

The technique is used to enhance the performance of the individual machine learning models. Initially, at start, if we look back in the history of machine learning that comprises of classification tasks, by Nilsson in 1956. It includes the merging of the individual base classifiers into one strong classifier. Thus, it is known as ensemble learning. This approach is inherited from human survival instincts that whenever there is situation that could decide the future outcome, human directed to different options and consulted different experts on the decision making. For example, when one has to go through a critical surgical/medical procedure, they consult with expert doctors in those fields. Whenever we go for purchasing anything for our daily life needs, we consult and visit various suppliers to sure the guarantee of its genuinity. Below image depicts the general idea of how ensemble learning works in the purpose of achieving the higher accuracy and performance of the individual weak machine learning models.

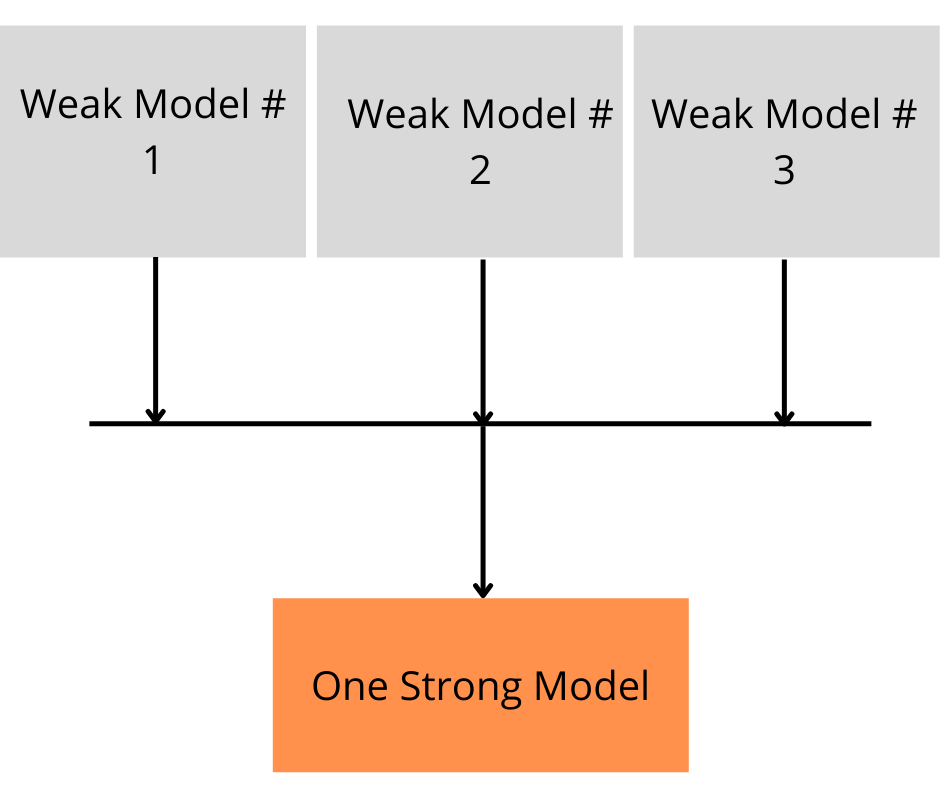


Figure 1 Ensemble learning methodology

Three major categories that are related to the ensemble learning are bagging, boosting and stacking. There are two types of weak learners such as homogeneous and heterogeneous. Bagging and boosting categories deal with homogeneous weak learners. Bagging considers the approach of individually training the base classifiers and then taking an insightful average of the base classifiers results. Boosting uses a different approach than bagging, in a way that it trains one base classifier and then uses that previous classifier’s results for next classifier’s training and evaluation. Whereas, stacking encompasses heterogeneous weak classifiers. It trains and learns them in a parallel way and then trains a one giant classifier to evaluate.

**Literature Review of Ensemble Learning Methods**

Ensemble learning includes various approaches of machine learning classifiers but most of the commonly used and famous are Bagging, Boosting, Majority Voting, Stacking and Averaging. All the methods mentioned are reviewed in-depth along with the advantages and disadvantages of the methods of ensemble learning. The relevant limitations in the working and compatibility of the methods have been disclosed as well.

**Bagging:**

The ensemble technique that is bagging is an acronym of two words “Bootstrapping” and “Aggregating”. Bagging helps in reducing the common machine learning model error, variance. Variance is the value that depicts the sensitivity of the machine learning model due to the minor level of changes in the dataset. If the value of the variance is higher which is due the frequent occurrences of the fluctuations in the dataset, the model will show over-fitting. To tackle this, bagging comes in place.

The main principles of the techniques behind bagging is bootstrapping and aggregating. A large dataset is divided into smaller partial datasets that are trained and tested interchangeably. Aggregating is the technique that is applied on the individual results derived from weak classifiers and then taking an average of the results to deduce a single output. And using majority voting for the training that is based on classification purpose.

**Advantages:**

* Bagging devastatingly improves the model’s accuracy for classification and regression type of weak learners.
* Bagging reduces the variance and hence helping in controlling the over-fitting of the machine learning model.
* One of the major benefits of bagging is, it works as a single strong classifier by using multiple weak classifiers.

**Disadvantages:**

* Bagging works on the foundation of various weak classifiers which eventually increases the cost of computation resources.
* This technique is useful if the model is properly built and compiled but if the opposite happens, then the model may go in under-fitting phenomenon.

**Boosting:**

In order to tackle biasness in the model, boosting method is used. It changes the weak classifiers into strong ones. When the weak base classifiers do not go beyond a specific level of accuracy/confidence then boosting assists in achieving the higher results than weak classifiers. The famous classifiers are Gradient Boosting, XGB and AdaBoost etc.

Under the hood, boosting works in a sequential manner. First, the base classifier is trained a chunk of the dataset and then the remaining dataset is used to evaluate the first classifier. The false predictions of the dataset by first classifier are then passed on to the second classifier which then works on the remaining dataset. So as this cycle continues as long as the models are chosen. Thus, the final strong classifier has the best accuracy results and score than the individual base classifiers.

**Advantages:**

* Boosting helps in tackling the missing values in the dataset.
* The other benefit has been found useful in dealing with binary classification problems.

**Disadvantages:**

* Due to the multiple inclusion of the base classifiers, the model implementation in real life becomes complex and costly.
* Boosting provides flexibility and hence multiple factors are generated that have direct impact on the working of the machine learning model.

**Stacking:**

Stacking or more commonly known as stacked generalization is another ensemble learning methodology that is built upon the concept generalizing the dataset. The base classifiers are trained on the dataset and then those trained classifiers are used again to generate a new dataset that will be used for a new larger (stronger) classifier as an input.

Stacking can include multiple layers of base classifiers in order to achieve a higher value of accurate results. A linear regressor might be used for aggregating the results for linear data and multiclass classification might be dealt with logistic regression approach.

**Advantages:**

* Stacking helps in achieving a higher amount of positive results as it includes multiple machine learning models.
* A large number of models from different categories can result in better evaluation of the features and predictions.

**Disadvantages:**

* A major hindrance is the computation resource consumption of the stacking method. Because multiple machine learning models are put together and thus require higher amount of resources and computation power.
* Inadequately chosen machine learning models may result in loss of useful information and features from the dataset and thus results in poor model performance.

**Majority Voting Method:**

Another simple and not too complex technique of ensemble learning is Majority Voting. The base classifiers are trained either on the same dataset or different instances of datasets. Then a voting is done upon the predictions of individual classifiers and the one that gets most of the votes, is taken as the final prediction.

Each model makes a forecast (votes) in favour of each test occasion and the last result expectation is the one that gets the greater part of the votes. Assuming none of the expectations get the greater part of the votes, we might say that the group strategy couldn't make a steady forecast for this occasion. Albeit this is a broadly utilized strategy, you might attempt the most casted a ballot forecast (regardless of whether that is not exactly 50% of the votes) as the last expectation.

**Advantages:**

* The majority voting method is useful for dataset that is for classification. As it can classify the results more accurately using the voting technique.

**Disadvantages:**

* The majority voting method is not applicable on regression/linear dataset.
* This limits the useability and the diversity of the method.

**Averaging:**

One of the easiest and simplest methods of ensemble learning is averaging. It is the way of deriving the average value of the predictions of various machine learning models used as base classifiers. Then an average is taken out to predict the accuracy or the results of the models.

Averaging uses the same approach for dataset useability as does the Majority Voting method. The datasets can be divided into chunks and then fed to the models or different datasets used for different machine learning models.

**Advantages:**

* The Averaging method is useful for dataset that consists of regression or linear approach. As it can accurately define a fine line along the flow of the fluctuations of the dataset.

**Disadvantages:**

* The Averaging method is not applicable on classification type of dataset.
* Which in case, is a drawback of the averaging ensemble learning method limiting the use of it.

**Experimental design:**

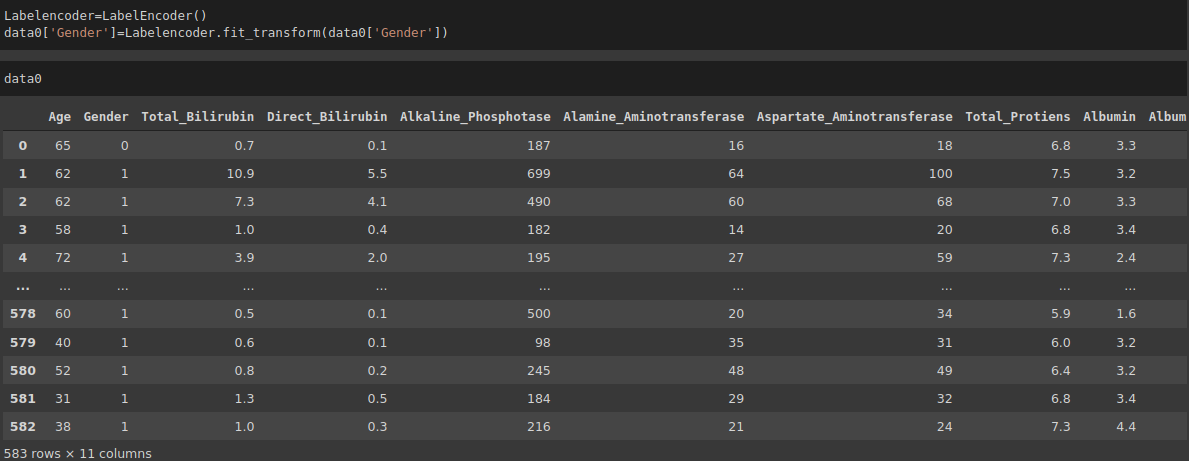
* Use Three Different Type of Datasets in csv Files Formate:

1. liver Patient.csv
2. processed cleveland.data
3. heart.csv

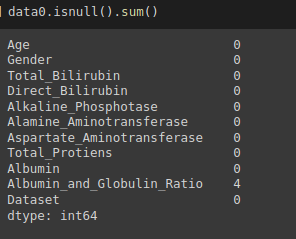
* Add the Columns name in processed cleveland.data



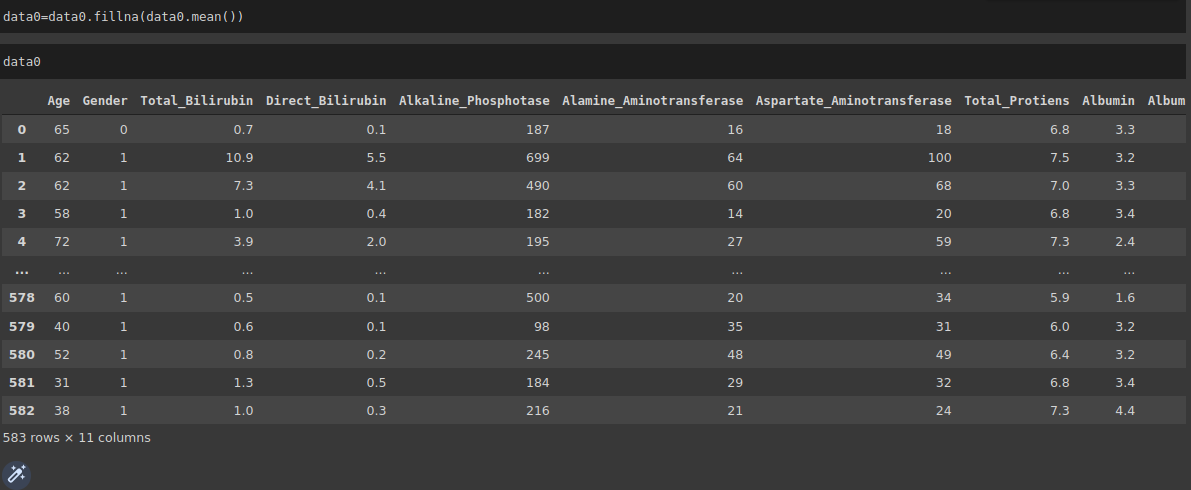
* To Performance LabelEncoder in the specific column (Gender) of the dataset of liver Patient to convert string into integer or float.



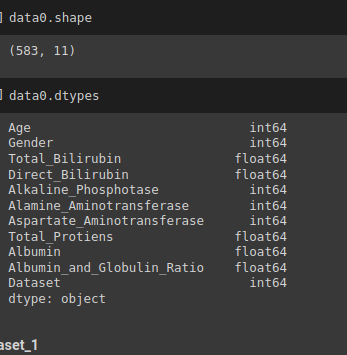
* Check the Missing Value of Each Dataset by using **isnull().sum() Function.**

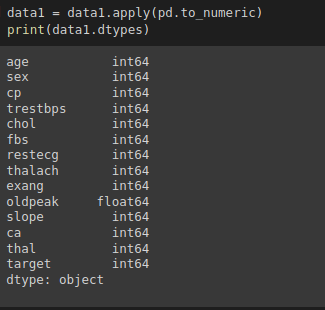


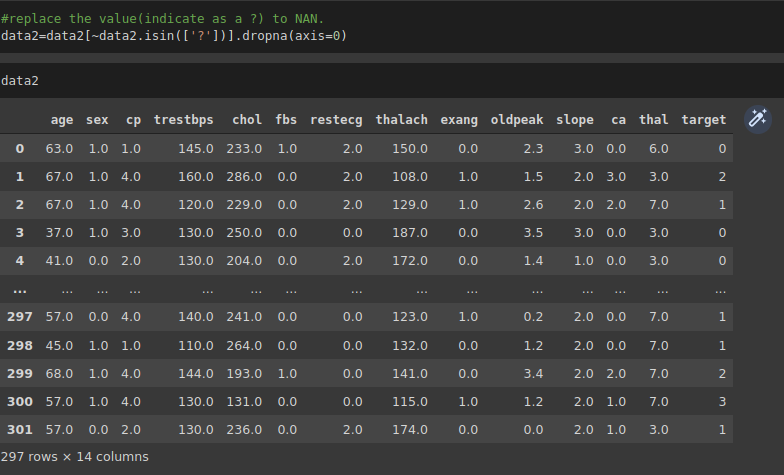
* If dataset contain the null value its replace by mean of this column



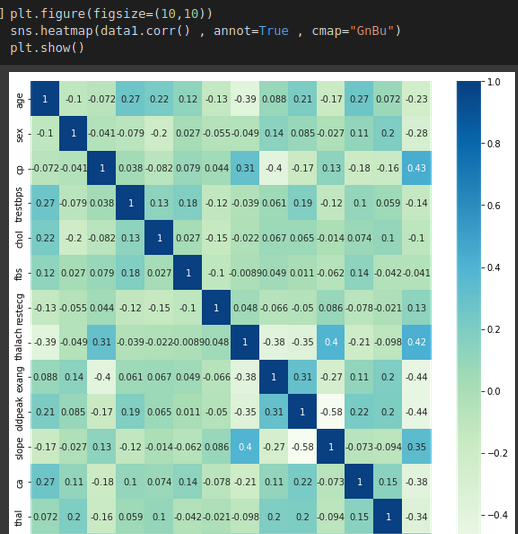
* To identify the shape and type of each dataset.



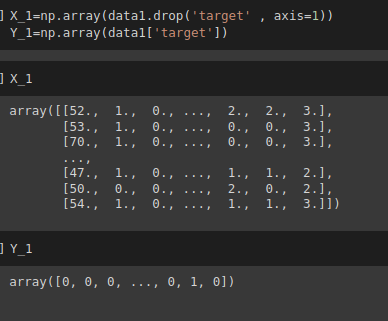
* If dataset has type object ,object type is converted into number by using pd.numeric() function.
* Dataset contains the value in the form ‘?’ .its replace by the null value from the columns.



* Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable.



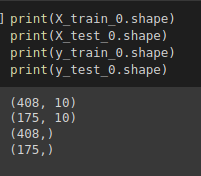
* Drop the target column and save in seperate variable(depended variable) and all dataset without target column save in seperate variable (independed variable).



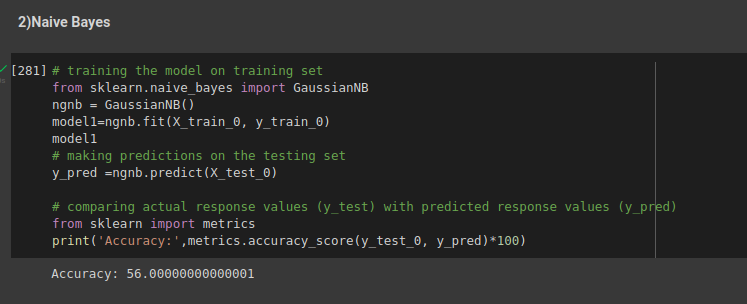
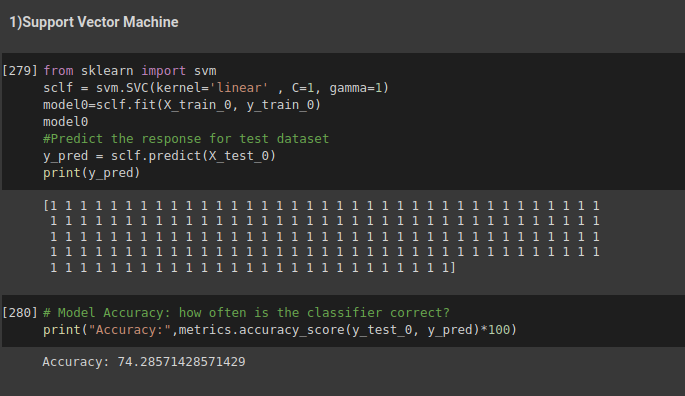
* Split the dataset on the depended and independed variable.

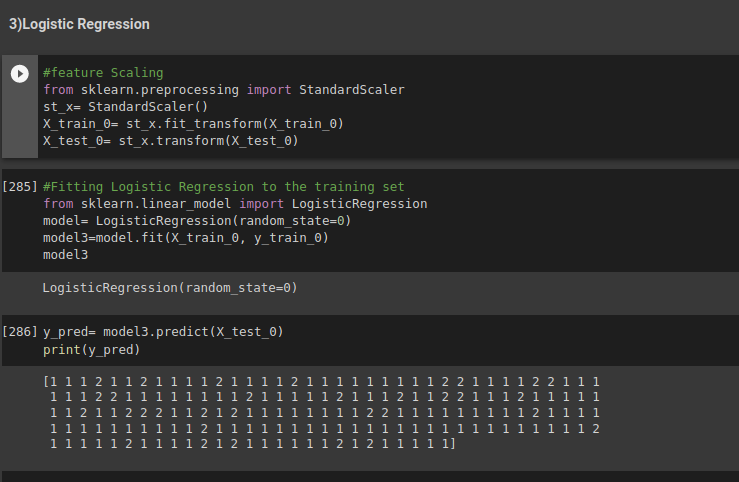


* To identify the shape of train data and test data seperatly

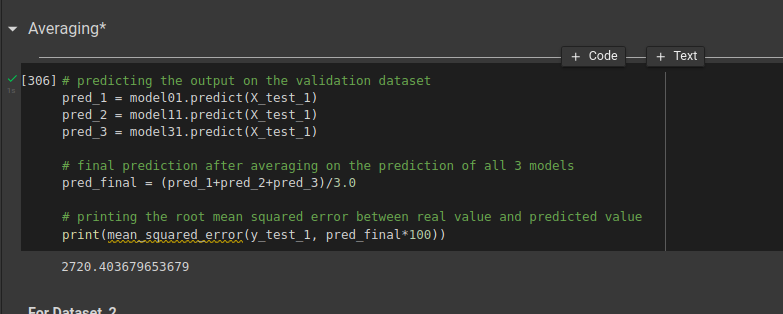


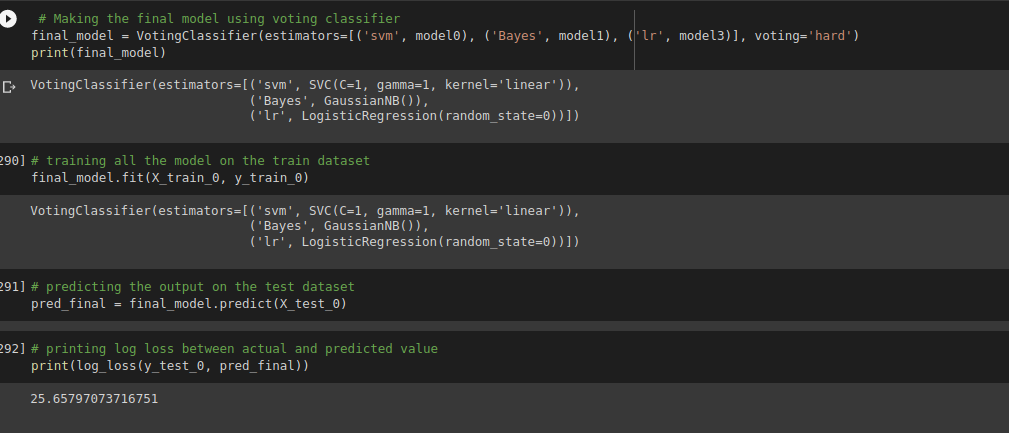
* Touse Three Different Base Classifiers (Logistic Regression , Native Bayes, Support Vector Machine). To fit Each Dataset in three base classifier and identify the accuracy .



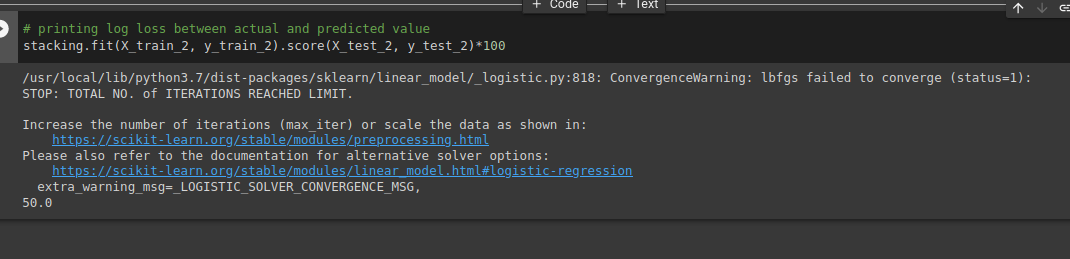


* To Improve The Accuracy By follow Essemble Learning Methods.(Stacking , Voting and Average) .



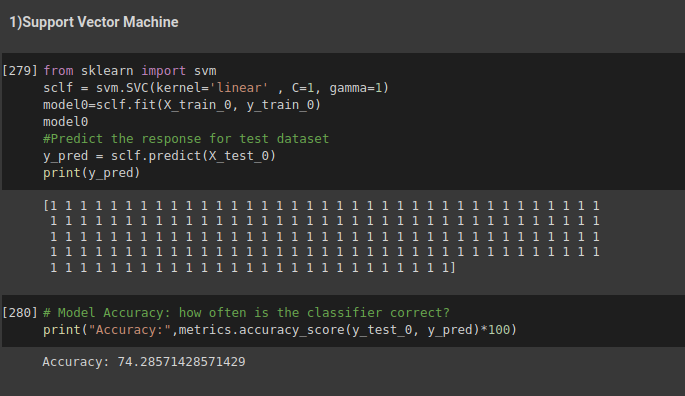


Stacking Essemble Learning Method:

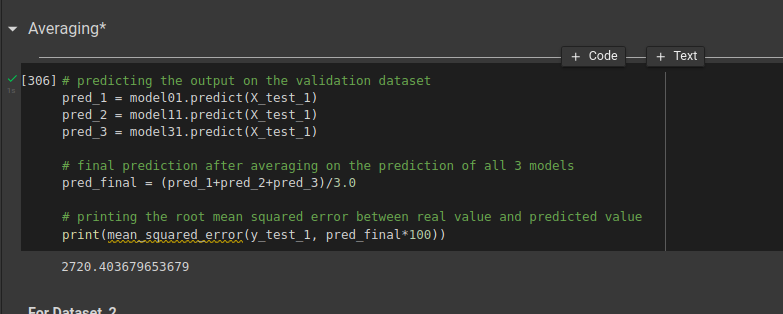


**Experimental results and discussion:**

* In Basic Classifiers, SVM(Support Vector Machine) produce Maximum Accuracy on Each Datasets and Performance Is better as compare the other Basic Classifiers.



* In ensemble learning (EL) methods , we used three different method (Averaging , Stacking and Voting classifier).Averaging Ensemble Learning Method provides Better Performance as Compare to Other Ensemble Learning Methods (Stacking , Voting).

**Conclusion:** Averaging Ensemble Learning Method Produce bether performance As compare to other Ensemble Learning Method..

**References**

[4]https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[5] http://scikitlearn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,than%20the%20number%20of%20samples.

[6] [https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on](https://www.sciencedirect.com/topics/computer-science/logistic-regression" \l ":~:text=Logistic regression is a process,%2Fno%2C and so on).

[7]https://machinelearningmastery.com/voting-ensembles-with-python/#:~:text=A%20voting%20ensemble%20(or%20a,model%20used%20in%20the%20ensemble.

[8]https://towardsdatascience.com/stacking-classifiers-for-higher-predictive-performance-566f963e4840

[9]https://www.geeksforgeeks.org/ensemble-methods-in-python/

[10]https://developers.google.com/machine-learning/crash-course/classification/accuracy#:~:text=Accuracy%20is%20one%20metric%20for,number%20of%20predictions%7D%7D%24%24

[11]https://en.wikipedia.org/wiki/Mean\_squared\_error

[12]https://en.wikipedia.org/wiki/Mean\_absolute\_error#:~:text=In%20statistics%2C%20mean%20absolute%20error,an%20alternative%20technique%20of%20measurement.