*Classifying Churn*

*Panga Tharun Kumar 25th March, 2019*

*CONTENTS*

1. Introduction
   1. [Problem statement 1](#_TOC_250006)
   2. Data 1
2. [Methodology](#_TOC_250005)
   1. pre-processing 4
      1. oulinear analysis 4
      2. feature selection 5
      3. scaling 6
   2. Modeling
      1. Model section 7
      2. classification 8
      3. Regression 9
      4. DecisionTree 10
3. [Conclusion](#_TOC_250004)
   1. [Metrics](#_TOC_250003)
      1. [Accuracy Score 11](#_TOC_250002)
      2. [Confusion matrix 12](#_TOC_250001)
      3. Missclassification 13
4. Extra in python code 13
5. Plotting in python code
   1. Plotting Accuracies 16
   2. [Plotting Error 17](#_TOC_250000)
   3. Plotting Missclassificaton 18
   4. Heatamaps 18

GitHub Link for this Repo : --

https://github.com/Tharun-999/CHURN-PREDICTION/blob/master/Loan.ipynb

# *. Introduction*

## *Problem statement :-*

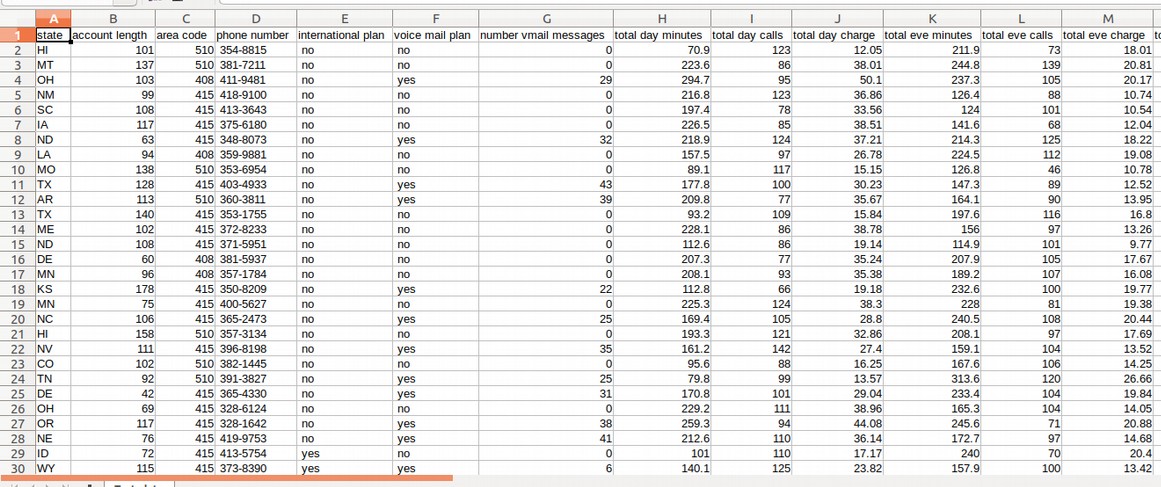
In this problem the main problem statement was the customers usage. company have given data to find the chruns moving or not from the given problem. We have different columns with different data and from the classification of the data it will help company which customers are going to move from company or not. This is done bythe using of machine learning classification problem.

## *About Data :-*

Data was consist of 3333 instances or rows, 21 types of Attributes. 21st column was class column it consists customers will move or not.

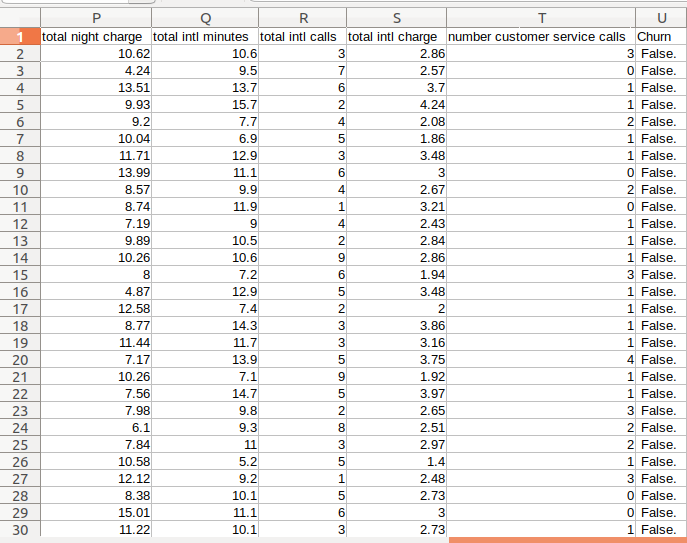
Columns of the test data

columns(1 - 13)



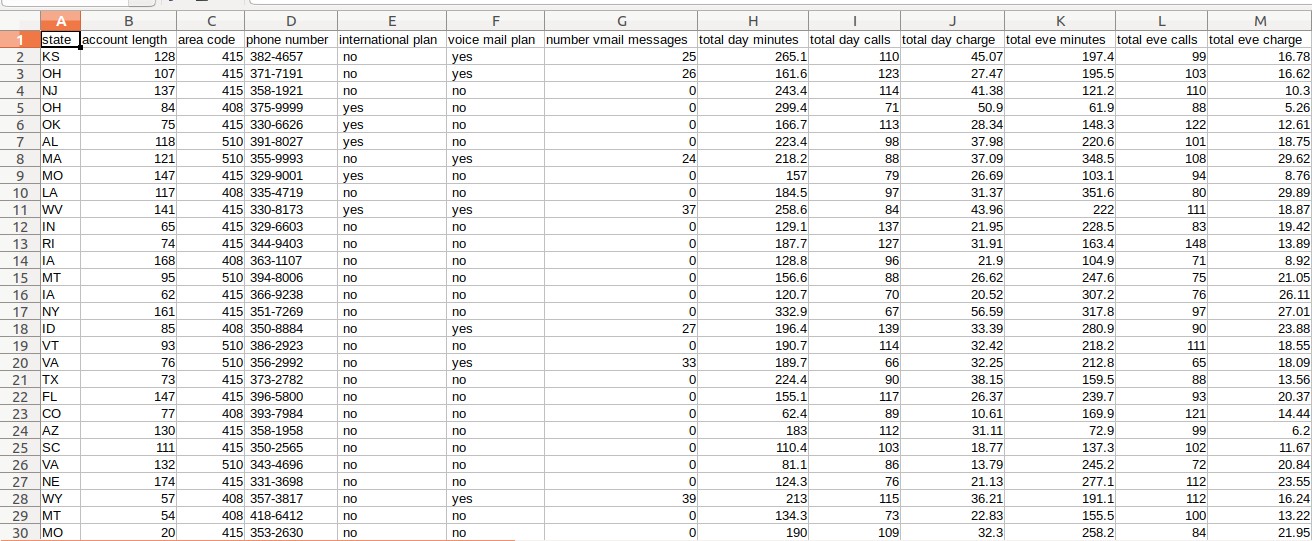
Last column consists of classes

columns(13 - 21)



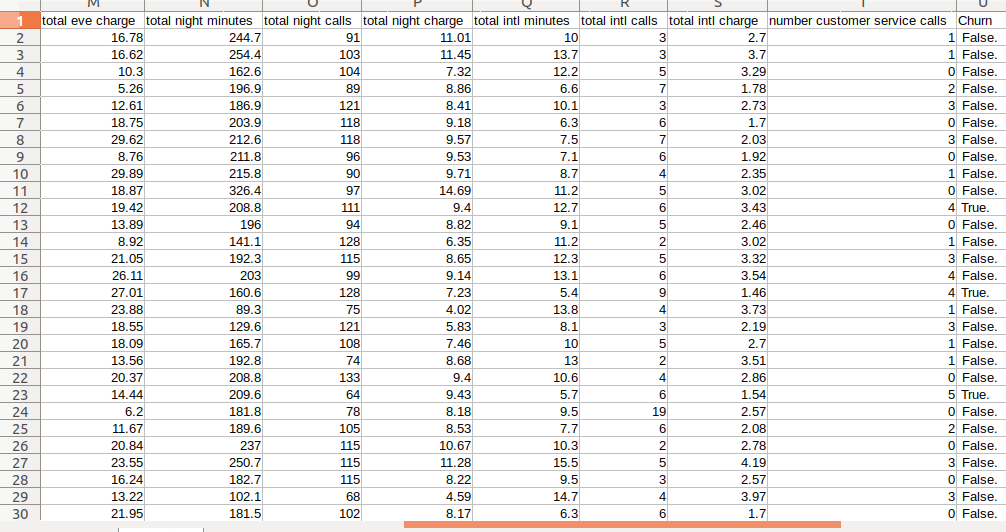
columns of the train data

columns (1- 13)

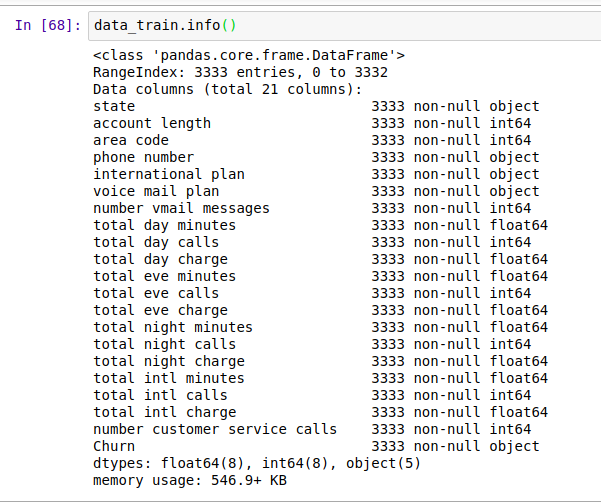


Last column consists of the classes

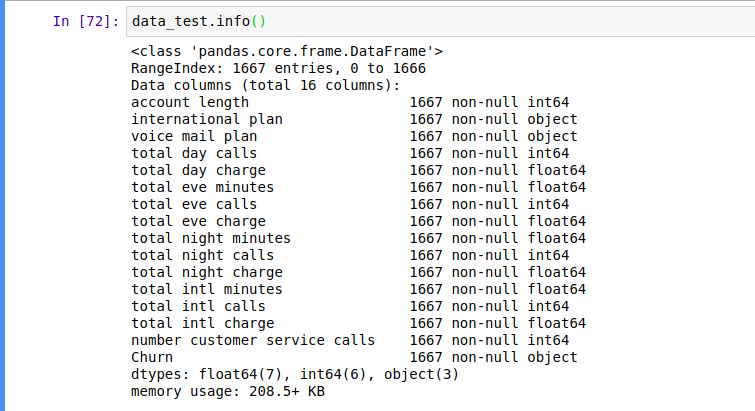
columns(13 – 21)



Datatypes of the columns in training data



Datatypes for the test data



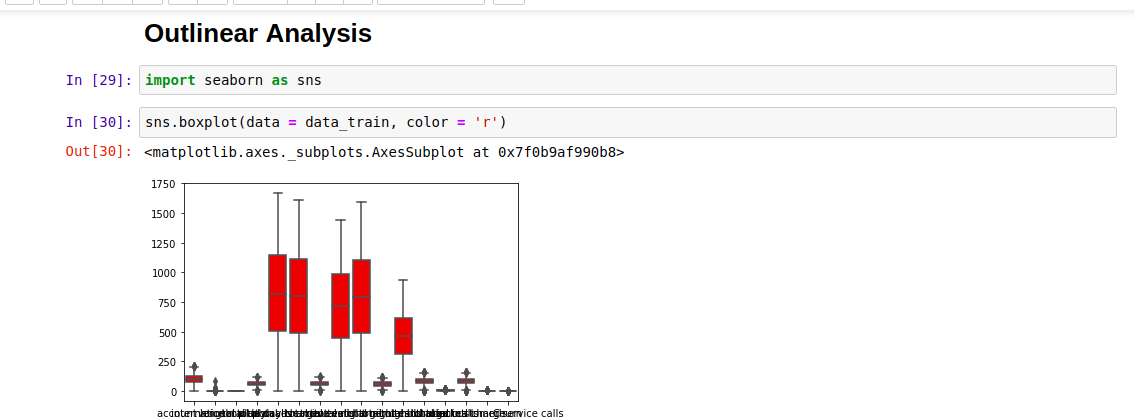
Info function in python will give information of all the data like how many **instances, datetypes**,it contains **null values** or not, **memory usage** of the data

# *Methodology :-*

## *Data preprocessing*

* + 1. *Outlinear analysis:*

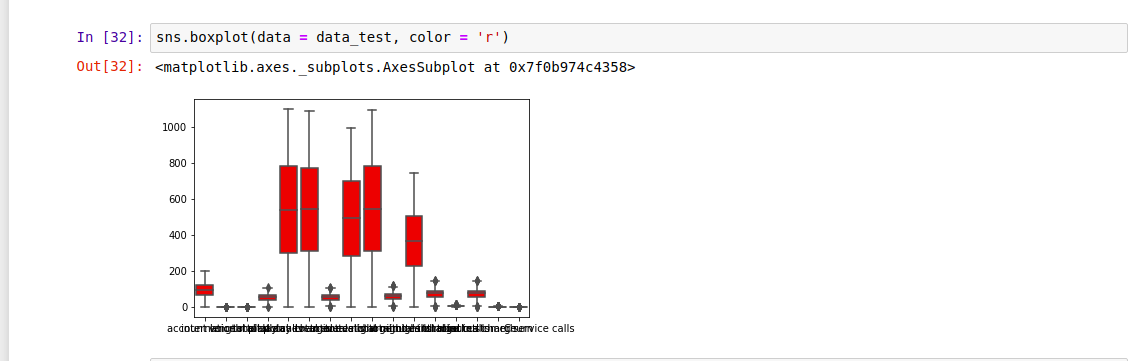
In statistics, an **outlier i**s an observation point that is distant from other observations. An **outlier** may be due to variability in the measurement or it may indicate experimental error the latter are sometimes excluded from the data set. An **outlier** can cause serious problems in **statistical analyses**.*O*utlinear analysis will help to find the which is out side of the box. It help us to which points are away from the region of the data. This can find with help of the Box plot in python it in the library of the seaborn.

Train Data

This analysis for the train data

In train data presence of out linear was not high. So we have no need to drop any values.

Test Data



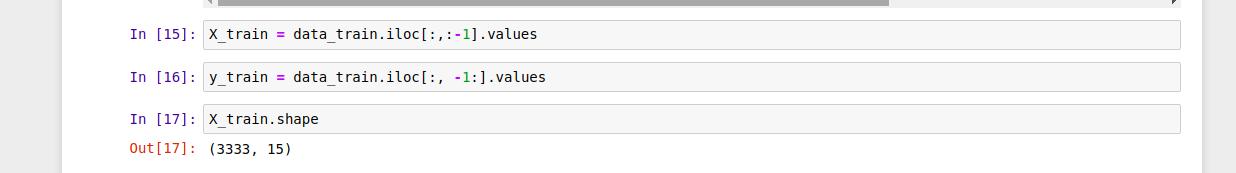
This boxplot is for test data

In this test data outlinears are not very high so there is no need of dropping or any analysis of the data

## *Feature Analysis :-*

In feature selection we are going to divide the data into the

**Features, Labels** . With some size of the data



X\_train says about training features of the data and it has taken all the rows and all the columns **except** last column

y\_train says about training data labels of the data and it has taken all the rows and only with **last** column



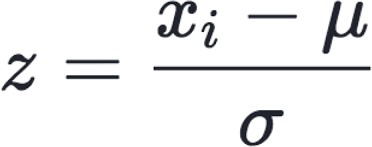
X\_test says about training features of the data and it has taken all the rows and all the columns **except** last column

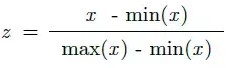
y\_test says about training data labels of the data and it has taken all the rows and only with **last** column

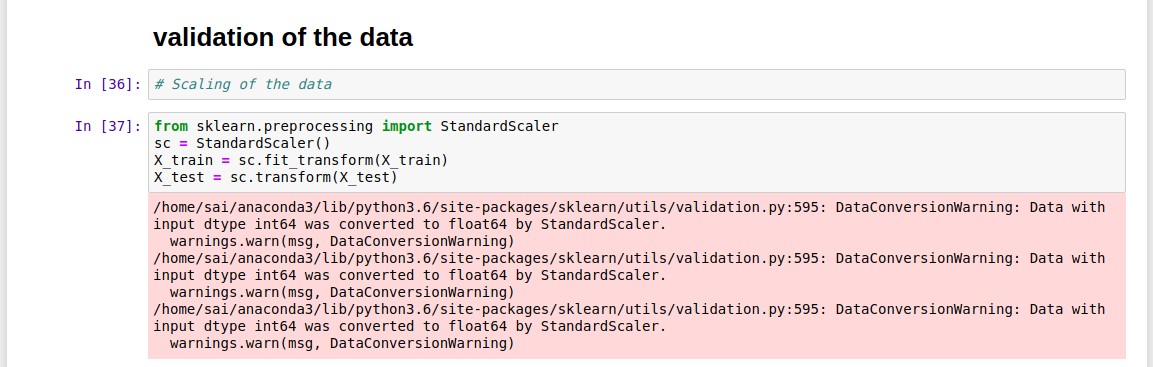
Features are used as the charecters of the data and labels are the drug. In the terms of the Medicine.This are most important to the data which column is the label or not to find the class.

## *Scaling the data :*

scaling of the data was the most important in data analysis because in will have large digit numbers and outlinears to bring all of them to one place we use the technique is called scaling. They have two techniques in scaling there are **StandarScaling** and **Normalization** . Both of this will bring any value in to between (0 – 1) it helps to the algorithms to reduce the complexity and increase to compile fast like KNN, SVM etc

**StandardScaler.**In standard scaler in data that will decrease by both of the **mean** and **Standard deviation** in the each and every value in the data

**Normalization** is used to normalize the data with the **min** and **max** values in the data it will bound in between the (0-1).it is more used for the **Outlinear** data to bring into the **same range** of all data



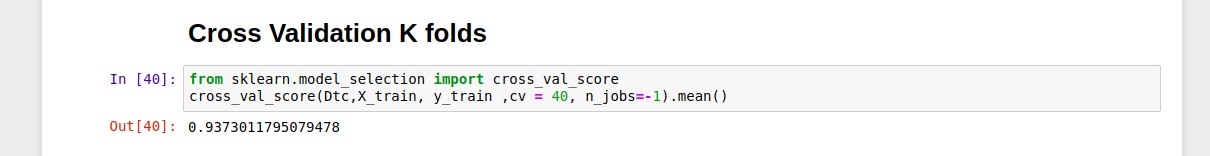
Standard Scaler of the is imported from the preprocessing class in sklearn library

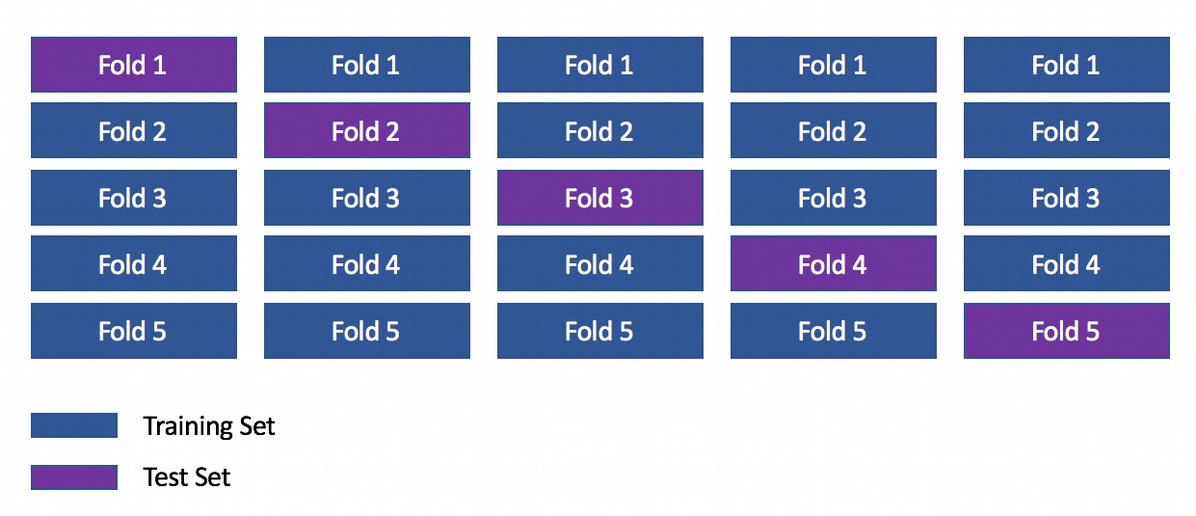
we will fit the data into the formula and then we will transpose it from bigger digits to range of (0-1). we do scaling for only the features of the data because Standard Scaler will only accepts the 2D and more dimensions of the data. But y label was the 1D dimension data.

# *3 Modeling: -*

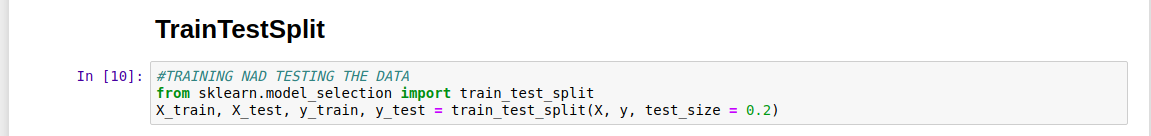
## *Model Selection :*

*Model selection* is class which is used to train the data, test the data, split the data. It has the most popular function **K Folds Cross Score** and **train\_test\_split** data



Here cross validation score used to train the data.in baches with k samples and the mean of the data is equal to the Accuracy of the data. We can get more by using this technique.

Train\_test\_split is another famous function to train and test data it takes features and labels in to the data. And Test or train size of the data to test or train the data



This also a famous function in sklearn library to split data into the tarin and test.

## *Classification:*

classification is one of the model in supervised learning. It mainly used for the classification problems

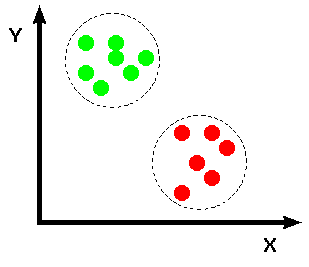
* Classification problems means it will have the classes like the Male,

female , boy, girl, 0, 1, true, false etc..

* Classification algorithms are the SVC, DecisionTreeClassifier, RandomForestClassifier etc..
* In classification there will have metrices mainly. Means the

confusion matrix, Accuracy score, f1 score etc...

* Classification is used to classifiy the data into the different groups and helps to the algorithms which one belongs to which group



## *Regression :*

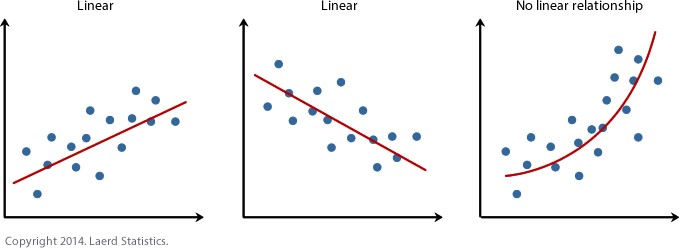
*Regression* is one of the important model in supervised learning

.It i mainly used for the Regression problems.

* Regression is mainy used for the prediction of the data like waether prediction , stockmarket, House rate prediction etc...
* Regression algorithms are the SVR, DecisionTreeRegressor,

LinearRegression etc....

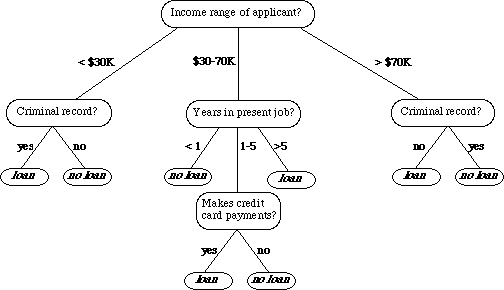
* In regression there will have the metrics class with functions like mean\_square\_error, log\_mean\_square\_error, etc...
* This problems mainly used for the prediction of the data



Mainly this problem was going to discuss about the Classification model because at the last column label is the discrete data, so here DecisionTreeClaasifier is using to find the Accuracy or to classifiy the data

## *Decision Tree :*

Decision tree is one of the most popular machine learning algorithms used all along, This story I wanna talk about it so let’s get started!!! Decision trees are used for both classification and regression problems, this story we talk about classifications



Decision tree will work like the decision making it help us to find which label is for the give feature

1. Decision tress often mimic the human level thinking so its so simple to understand the data and make some good interpretations.
2. Decision trees actually make you see the logic for the data to interpret(not like black box algorithms like SVM,NN,etc..)

For example : if we are classifying *bank loan* application for a customer, the decision tree may look like this

Decision tree are very powerful there are used for the Bagging and Boosting algorithms they work like Human brains. They are good decision makers.

*Information gain* **:-**

****

****

*GiniIndex :-*

**

**

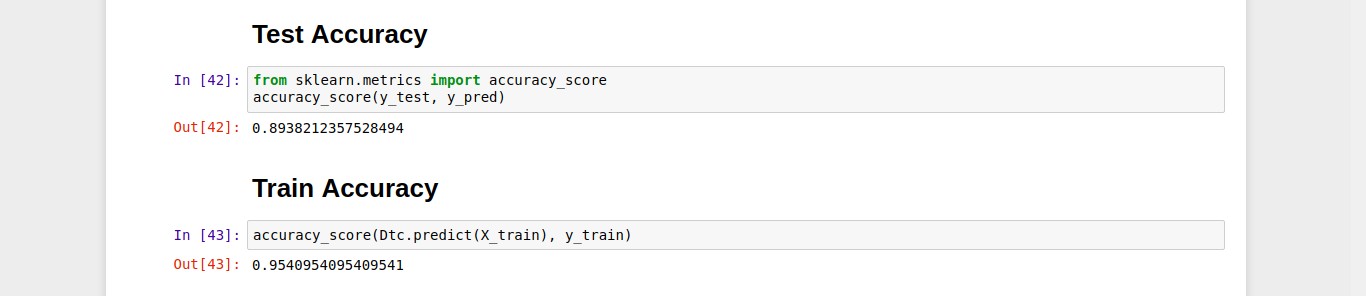
# *CONCLUSION*

## *Metrics :*

The most important thing data science when we are working with the data is Accuracy Score, Confusion Matrix this functions are in the scikit learn library in Metrics class

## *Accuracy Score :*

Accuracy score was the most important in algorithm it say about the how the algorithm was trined with given data. It says about how much learned from the tested data

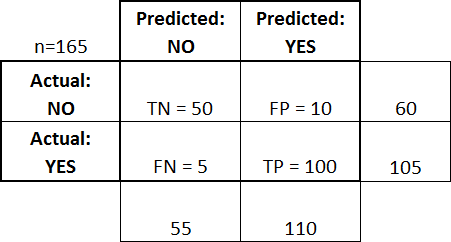


This figure will says about the how much data was trained and tested by the algorithm.The accuracy score on Test data is 89.38 and on trained data

95.48. We can find the algorithm was overfitted or not by seeing the accuracy scores.

# *Confusion Matrix :-*

A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The **confusion matrix**itself is relatively simple to understand, but the related terminology can be **confusing**.

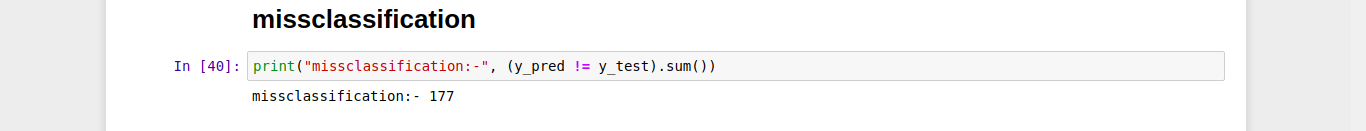


This says the comparision between the true predicted values and the flase predicted values



## *Missclassification values :-*

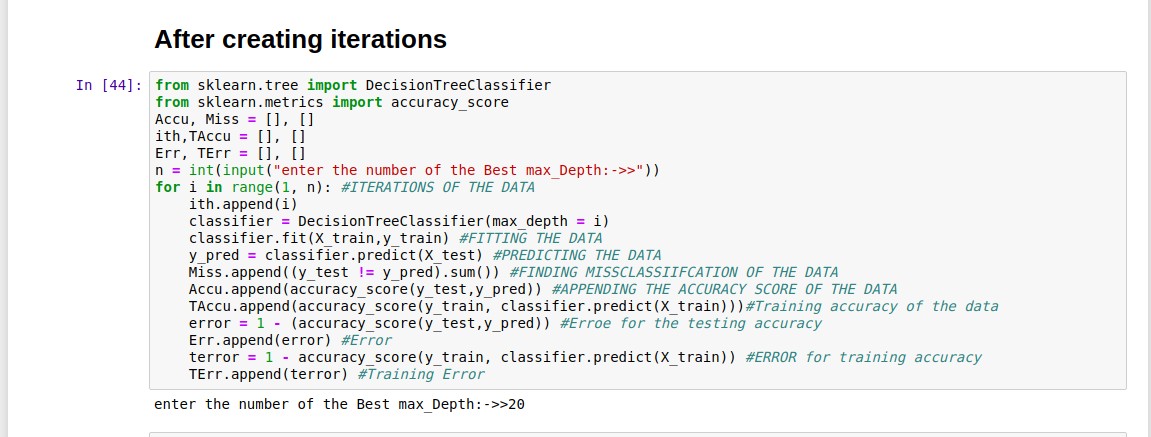
Missclassification says about the how many Actual values are Not equal to the Predicted values

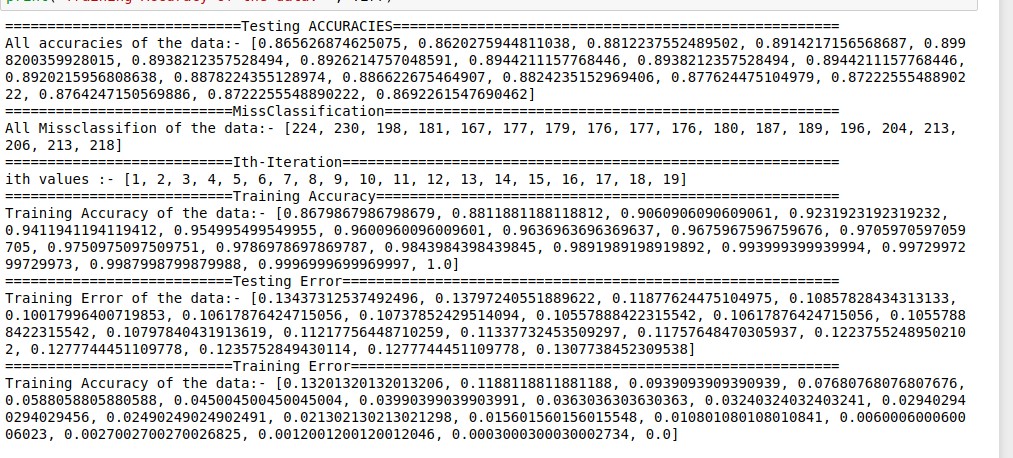


Missclassification in this data was the 177 .It will say about the and its accuracy

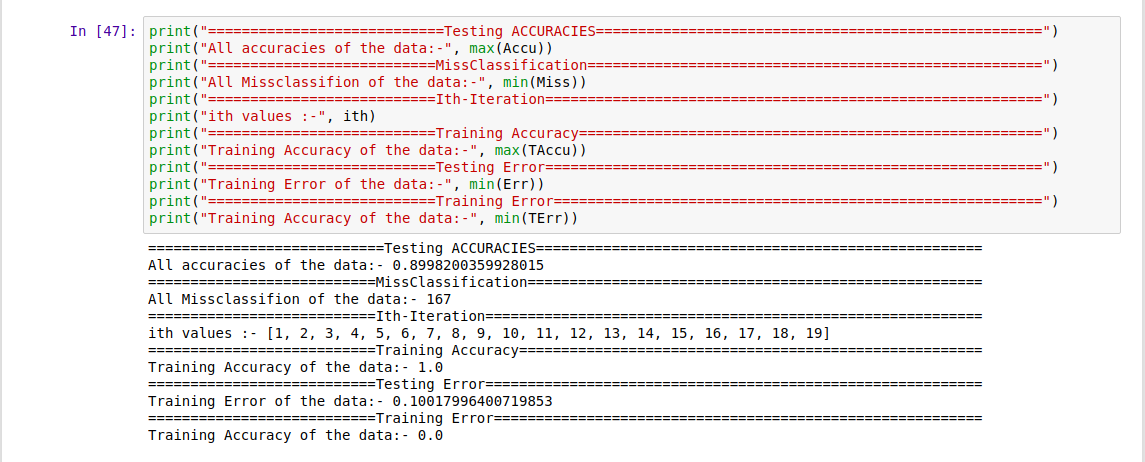
# *Extra in the python code for the data :-*

I tried to increase the accuracy in the data with the help of the values are passing in to the lopp of data





This will give all the information about the missclassification, error, Accuracy

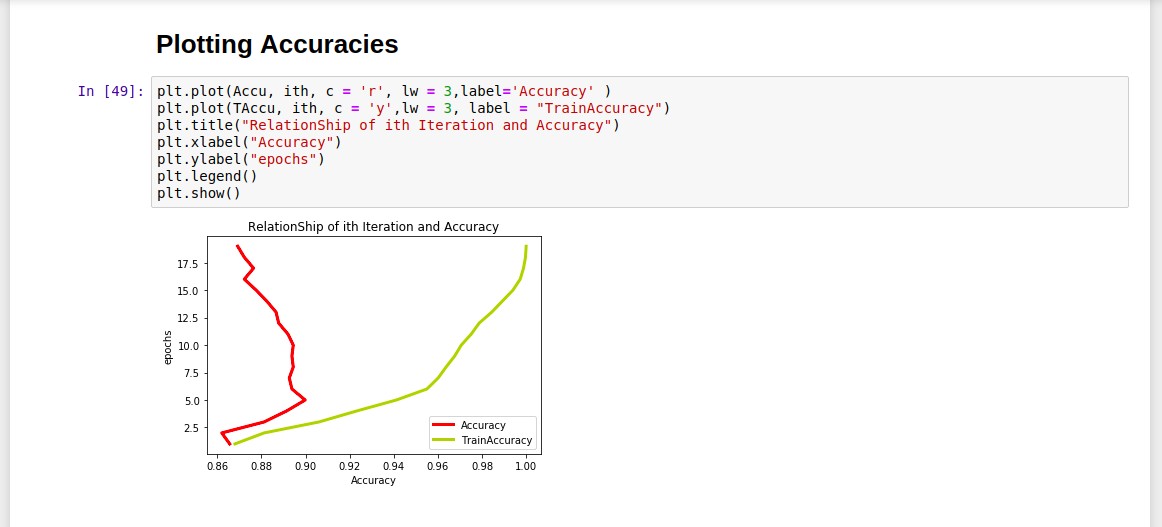


It Gives All the best values from the data

# *PLOTTING OF THE DATA*

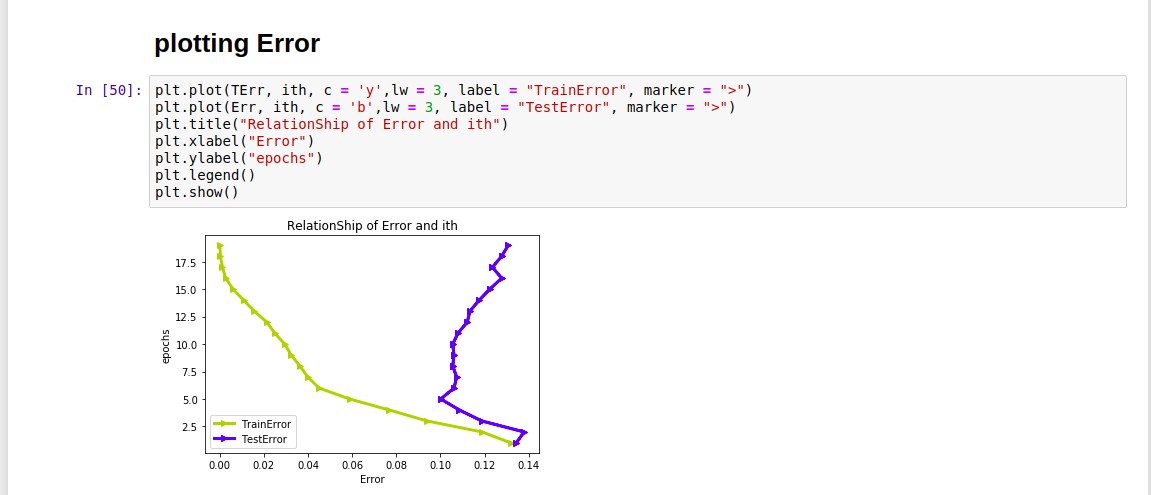
## *Accuracies Plotting :-*

Plotted with the help of matplotlib in Sklearn Library



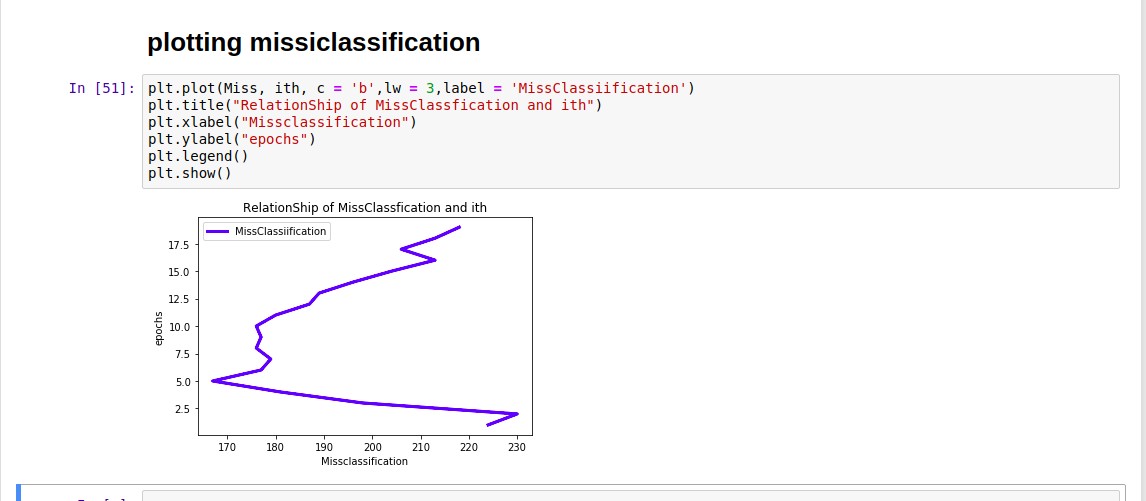
*Accuracy plotting* The diffrence between the train and the test Accuracy on the data in 20 epochs

## *Plotting Error :-*

**

The difference between the train and test in 20 epochs of the data

## *Plotting for MissClassification :-*

**

*Missclassification* says about the data missclassified wile testing with actual values

## *Plotiing for HeatMaps :-*

Heat maps are mostly drawn with the help of the Seaborn which called as the advance matplotlib library. It helps us to find the correlation between two variables

