**Loan Application Status Prediction**



**Problem Statement:**

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

**Independent Variables:**

- Loan ID

- Gender

- Married

- Dependents

- Education

- Self Employed

- Applicant Income

- Co-applicant Income

- Loan Amount

- Loan Amount Term

- Credit History

- Property Area

**Dependent Variable (Target Variable):**

- Loan Status

You have to build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.

**1.Problem definition:-**

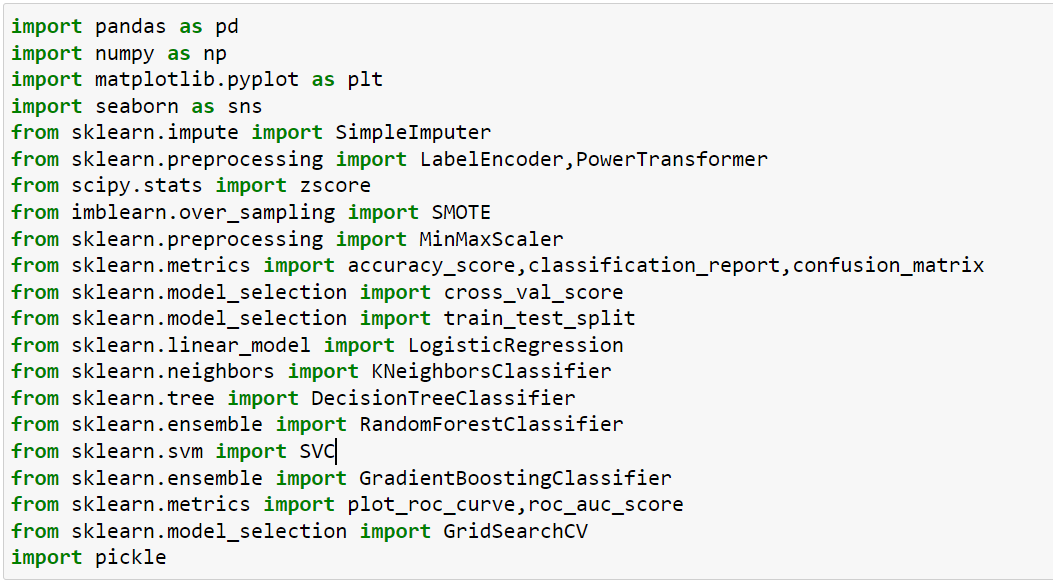
Now a days apply for the loan is very easy it takes minutes, but it is hard for the banks to filter applicant’s who are eligible and who are not eligible, after spending some time on to check applicants details the result is he/she not eligible for the loan, here precious time of the banks is becomes waste, we have a dataset that includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

Objective:- here with the help of details provided by the applicant, we have to predict he/she eligible

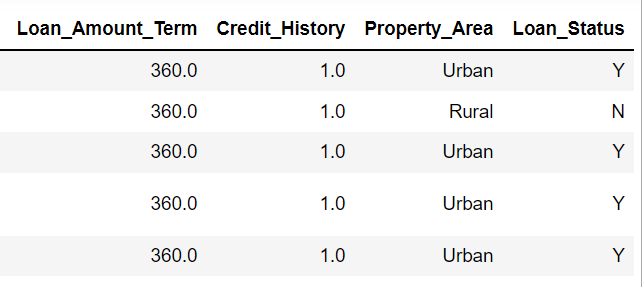
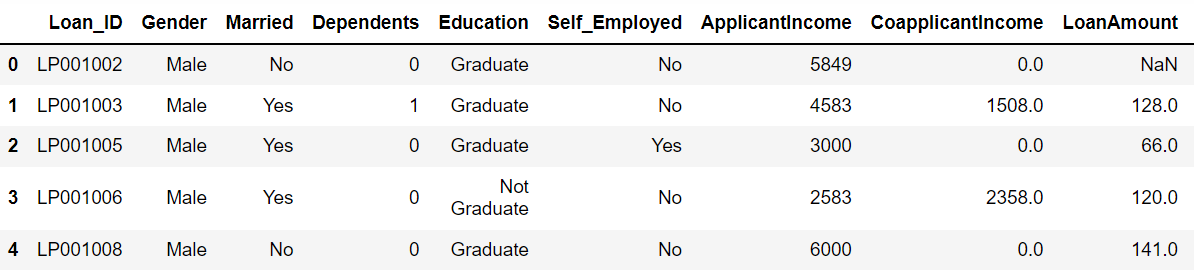
Type of problem:-it is classification problem our target variable has two categories yes and No

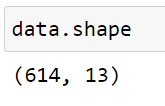
**2.Data Analysis:-**

Importing required libraries:-



Import the data set:-





**Dataset description:-**

we can clearly see that we have only one data set and it contains 614 rows and 13 columns, we have to split the data set into train and test sets after some preprocessing steps, training set is used for training the model I,e our model will learn from the training set and test set is used for how much our model learn from the training set and target variable is Loan\_Status

let’s understand each feature in our data set:-

**Independent Variables:**

- Loan ID:- ID is a unique thing for each applicant

- Gender:- Gender of the applicant

- Married:-marital status of the applicant

- Dependents:-no of persons depends on the applicant

- Education:-education level of applicant

- Self Employed:-applicant is self employed or not

- Applicant Income:-income of the applicant

- Co-applicant Income:-income of co-applicants income

- Loan Amount:-total amount of the loan

- Loan Amount Term:-duration of the loan in months

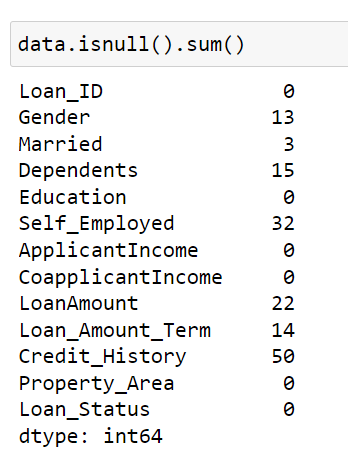
- Credit History:-Applicants previous credit history meeting guidlines

- Property Area:- property belongs to which type of area

**Dependent Variable (Target Variable):**

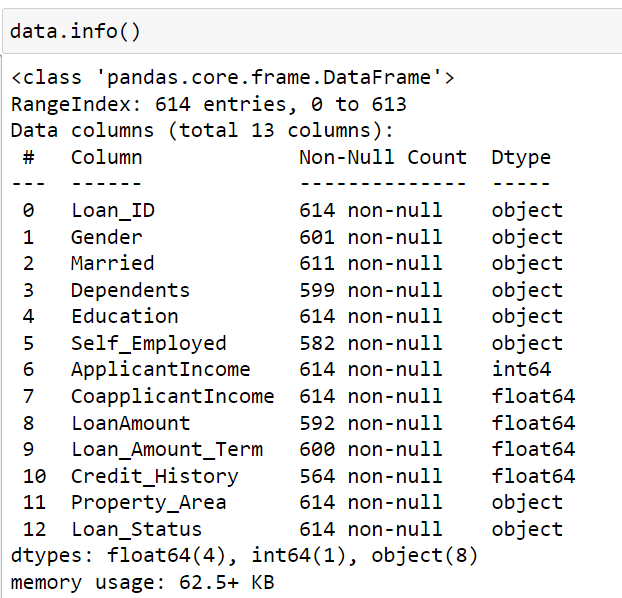
- Loan Status:- ‘y’ indicates yes ‘n’ indicates No

**Let’s check the null values:-**

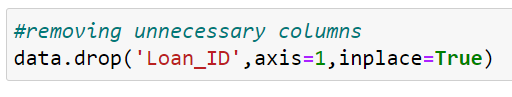


* Most of the columns are hiving null data we have to fill these by appropriate data
* Maximum null values in credit history column and minimum is married

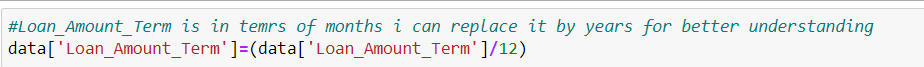
**Let’s check the information of the data set:-**



* Our data set contains multiple data types
* We have object dtypes, machine learning model not understand string data so features(8) which are in object dtype we have to convert them into numerical(integer) type



Loan\_ID is nominal data it is unique for every applicant so it is not helpful to our model prediction



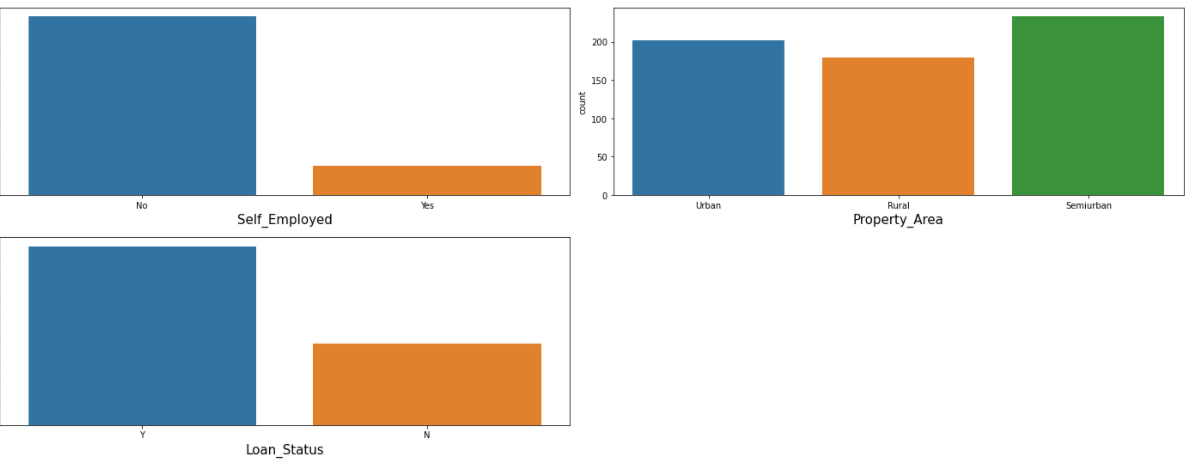
Loan\_amount\_term is in the form of months we can convert it into years for better understanding.

**2.DATA ANALYSIS:-**

**Univariate analysis**

Let’s analyse features through plots:-





**Observations:-**

**GENDER:-**

Most of the applicants are male compared to female count

**MARRIED:-**

Married people are applying loans more compared to unmaried

**DEPENDENTS:-**

Dependents with 0 count is high that means individuals are applying loans more

**EDUCATION:-**

we can see clearly most of the applicants are graduates

**SELF\_EMPLOYED:-**

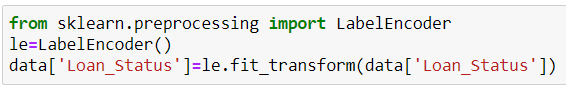
The count is high who are not self employed

**LOAN\_STATUS:-**

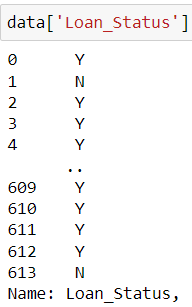
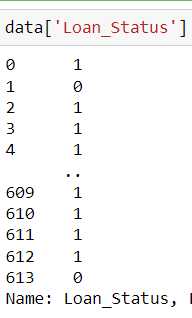
Most of the applicants are got approved loans

**Encode our target variable by using LabelEncoder:-**

Iam using LabelEncoder to encode the target variable

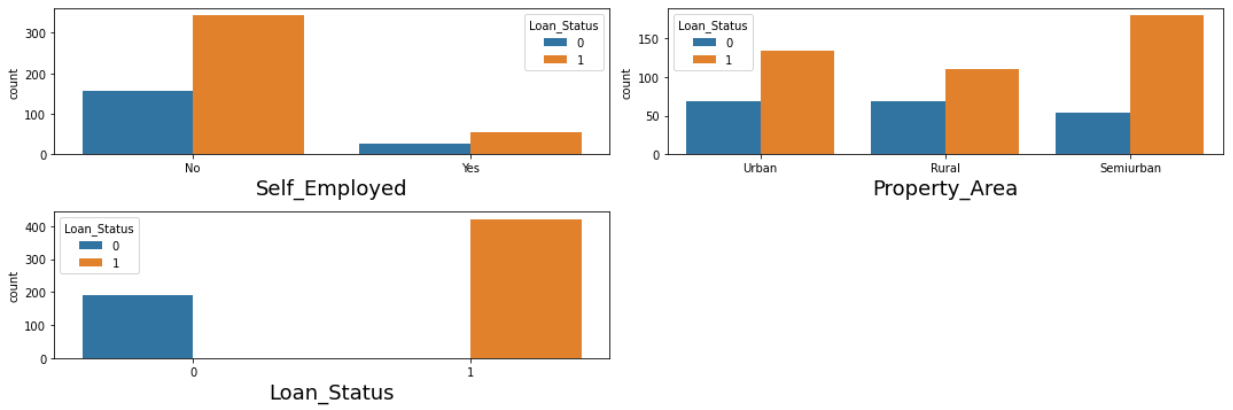
****

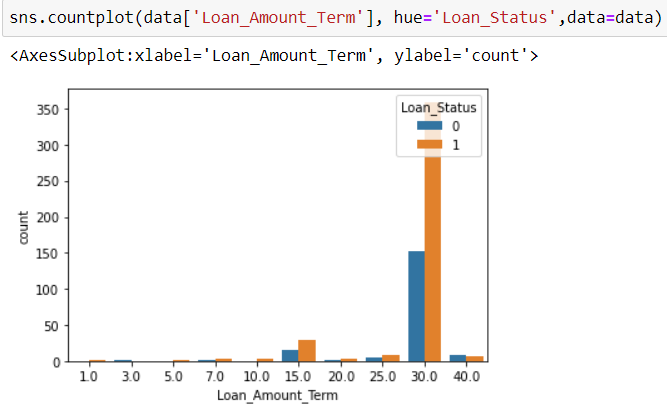
Before encoding after encoding

****

**Bivariate analysis**

****

****

****

****

**Observations:-**Here ‘1’ represents loan approved and ‘0’ represents loan not approved.

**Gender:-**Most of the loans are approved in male category and also female category approve rate is high compared disapprove.

**Married:-**Maximum loans are approved married people only and also who are unmarried the approve rate little bit high compared to disapprove.

**Dependents:-**Maximum loans are approved to who are have dependents with 0 and also approve rate is in other dependents category is high compared to disapprove rate.

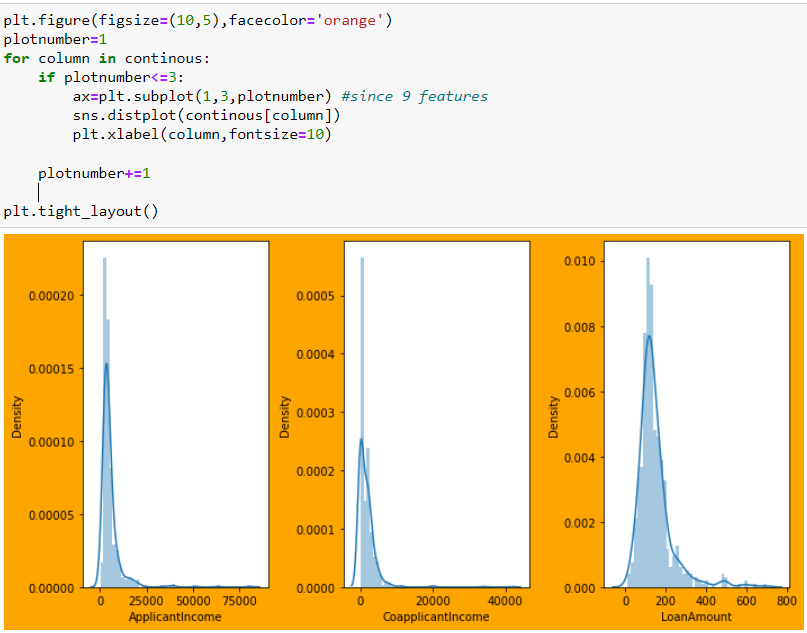
**Education:-**Most of applicants are graduates compared to non graduates.

**Self\_employed:-**Applicants who are not self employed approve rate is high comparing to non graduate.

**Property\_Area:-**loan approve rate is good in all property areas but property in semi urban is high chance to get approval 2nd one is urban and least one is rural.

**Loan amount term:-**As per the given data set loan amount term is in months we can see that there is most of the loans approved for the term of 30 and disapprove rate in this category is also high compared to all other

**Credit\_history:-** Although Credit History is not an object type, it has categorical values, and as we can see people who have 1 credit history are having more chances of getting a loan approved.



**Applicants income :-**

Data is skewed right side due to income range of the applicants are high one on another

**Coapplicants income:-**

This is also data which is skewed right side i,e positive side

**Loan amount:-**

it has skewed data but not much as above features.

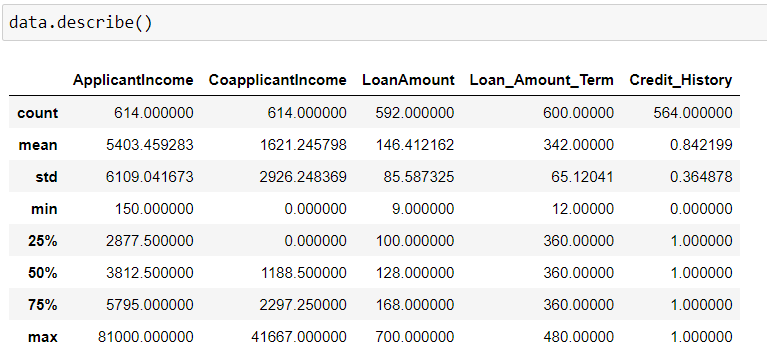
**Note:-if the data is skewed right side it indicates that following conditions**

**1)mean greater than the mode**

**2)median greater than the mode**

**3)mean greater than the median**

**Lets see some statistical information of the data which have:-**

****

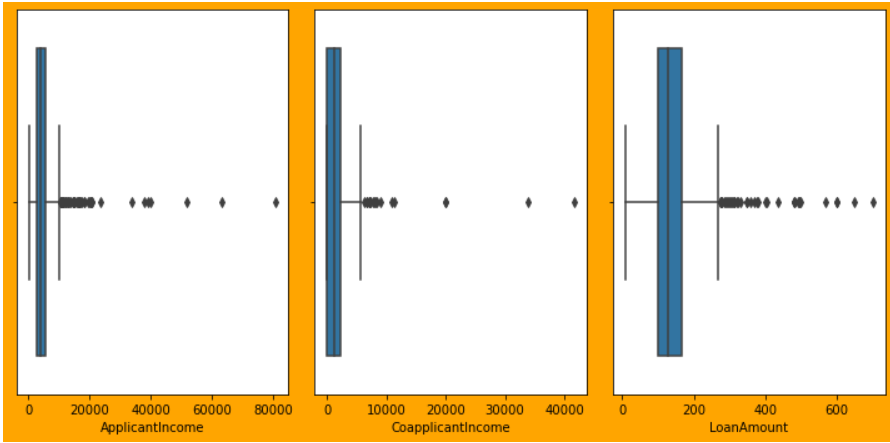
**Obseravation:-**

1)In some features mean is less than standard deviation because data is more scattered

2)minimum income of the applicant is 150 and coapplicant income is 0 which effects loan approve

3)we can observe some of the columns are having outliers because difference between 75% and max is muh more, we can analyse those features by plotting box plots

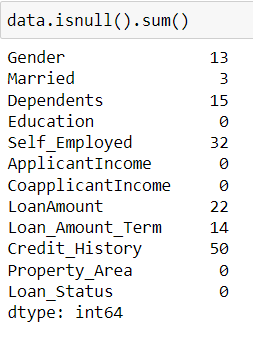
**Lets check the outliers:-**

****

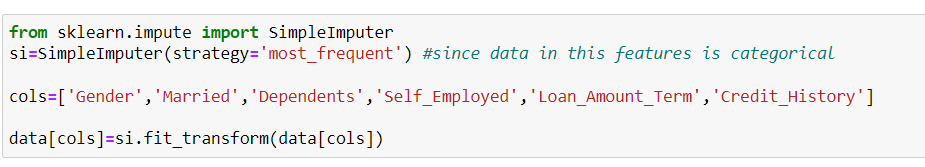
As we seen in the statistical information, yes there are some outliers

We have zscore and IQR methods for treating outliers, in this problem I am using zscore method ,I will apply it later.

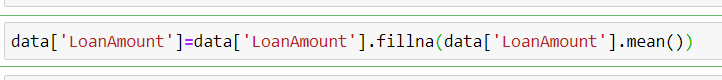
**Fil the null values each feature:-**

****

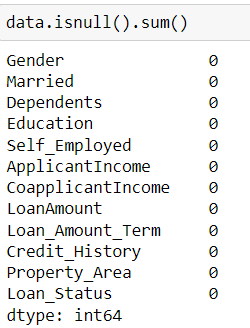
Categorical data:-Features which has categorical data we can use mode to replace the nan values



continuous data:-mean is the best method to fill the nan values in the continuous features

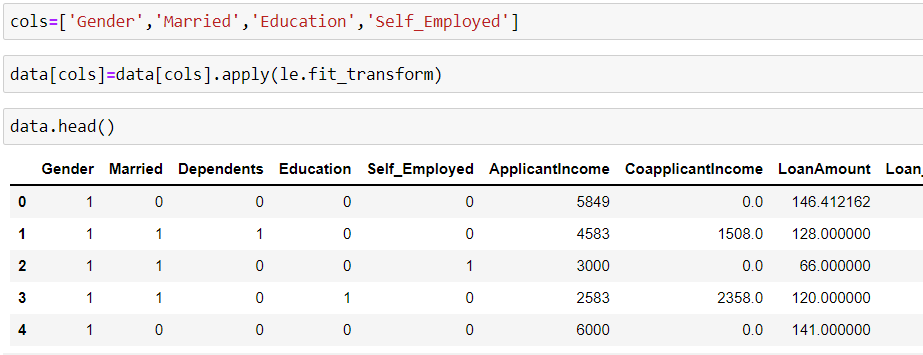


Checking nan values again:-

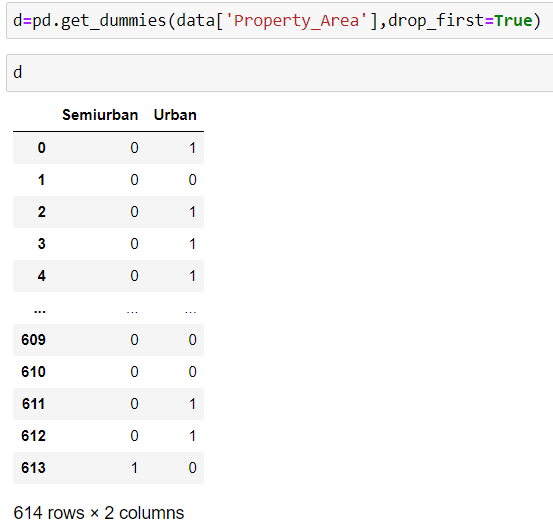


**Encode the object data type:-**

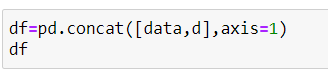
We have 4 features have object data type by using label encoder encode the data



Again we have property area in the form of object and with 3 categories so it is better to use get dummies method



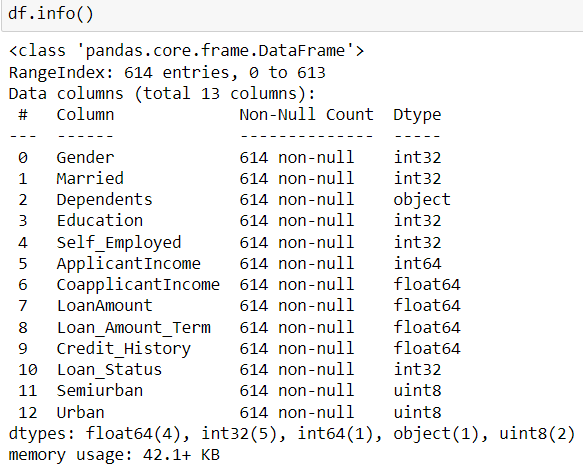
Let’s concat the data frames ‘d’ and ‘data’



After adding two data frames remove the duplicate column I,e Property\_Area

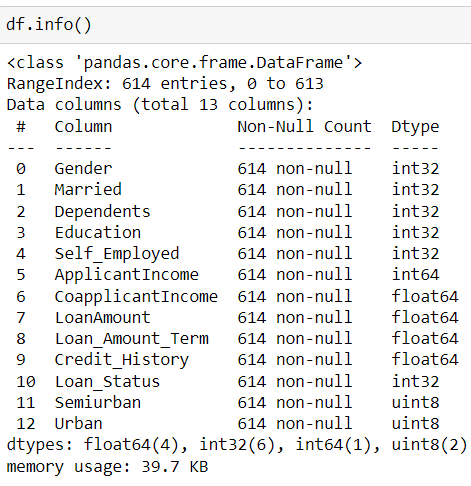


Let’s check the information of data set again

  
Dependents feature is in the form of object dtype because ‘3+’ is one of the category in it .

So we can replace ‘3+’ by 3,since three represents more than two so we can replace by 3

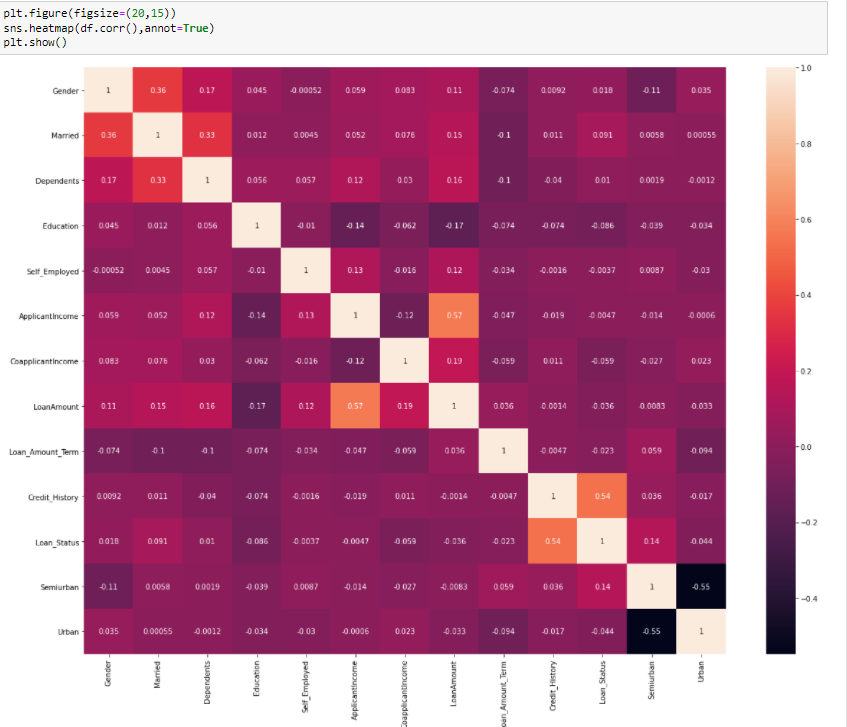




**3.EDA Concluding Remarks**

* Filled missing values with appropriate values
* Converted categorical data into numerical by using LabelEncoder and get\_dummies methods
* Removed special characters from the data

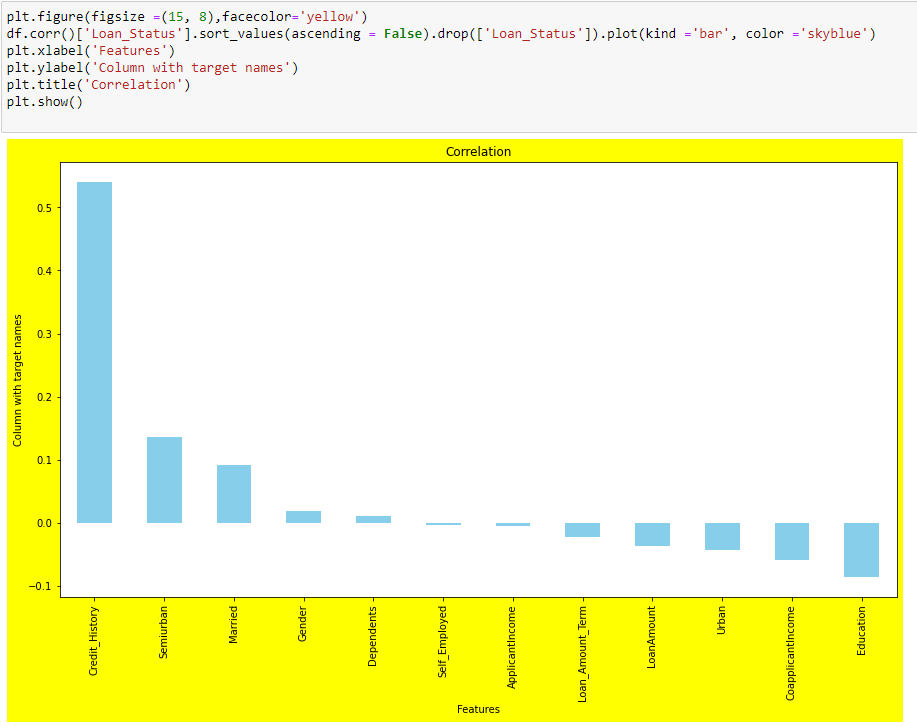
**Let’s check correlation between features by using heatmap**

****

**Observations:-**

* Credit\_history highly correlates with target variable
* Semi urban negatively correlates with target variable
* Applicant income and loan amount are correlates each other but not much
* Finally there is no features with the correlation of more than 90% so we cannot remove any feature

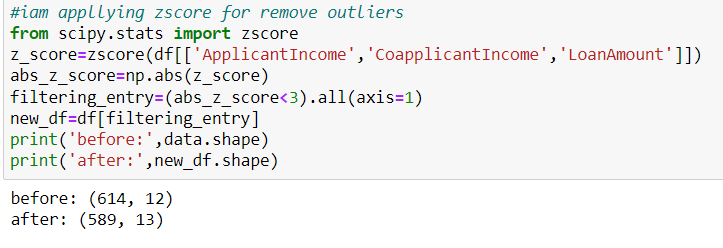
**Correlation with target variable vs features by using bar plot**

****

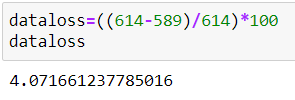
* Applicants income and self employed are less correlated with the loan status

**Removing outliers:-**

* By using zscore

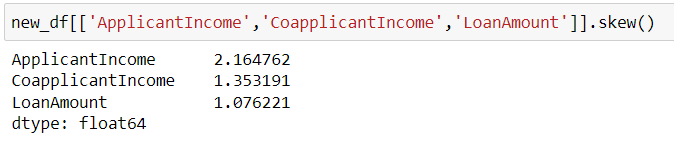


Lets check loss% of the data



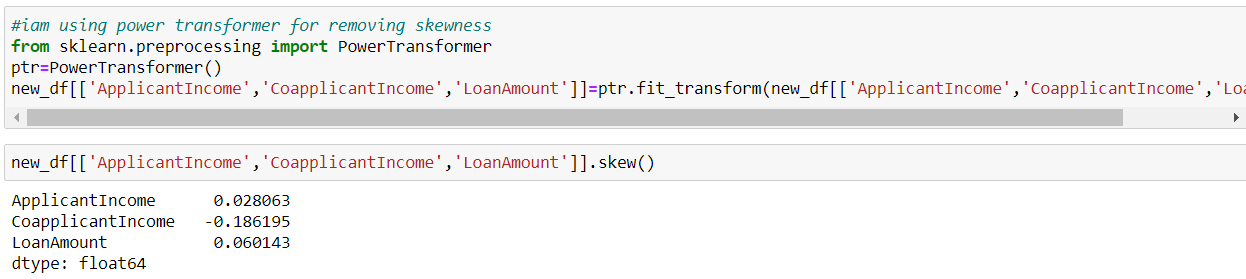
* It is less than 5% so we nee not worry we can proceed further steps.

**Let’check the skewness of the continuous data:-**

****

**Note:-skewness is consider in the case of continuous data only**

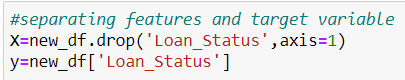
* Accepted range of skewness is -0.5 to +0.5
* But here data skewness range is more than accepted range
* We have multiple transformation techniques to remove skewness
* I am using power transformation technique to deal with skewed data



* Skewness is in the range accepted range

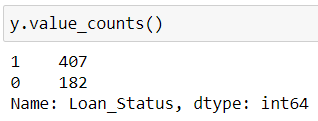
**4.pre-processing Pipeline**

let’s separate the data into dependent and independent



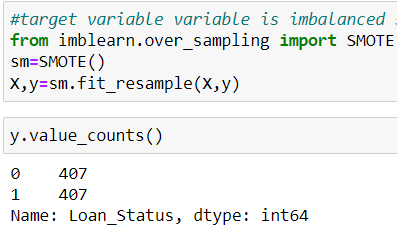
From the univariate analysis I remember that our target variable has imbalanced data

Let’s check again



Yes our target variable is imbalanced

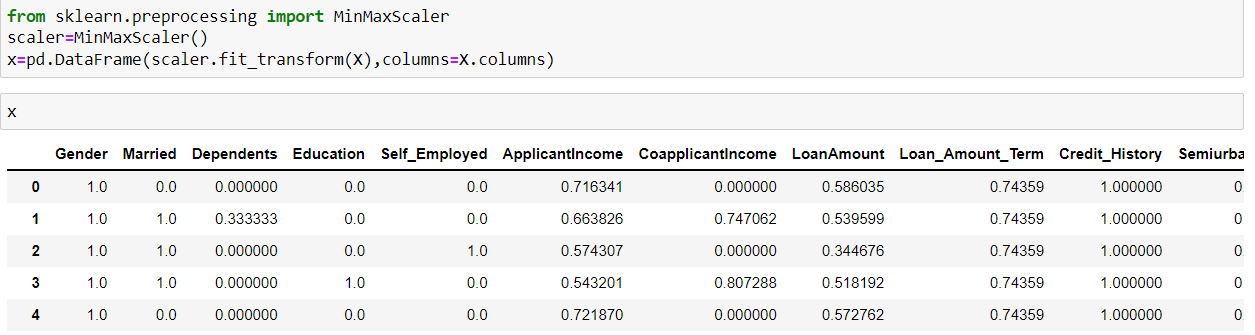
We can use ‘SMOTE’ for balancing the data



Now it looks balanced

**Scaling the data by using Minmax scalar:-**

**Minmax scaler:-**where the minimum of feature is made equal to 0 and the maximum of feature equal to 1. MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.



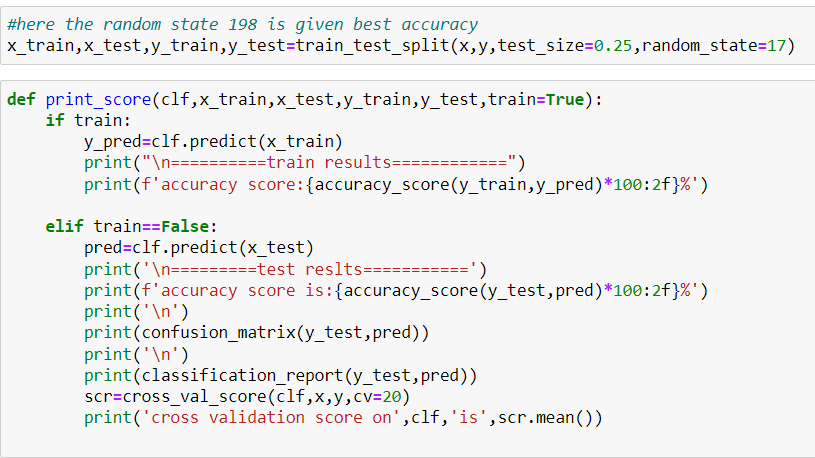
**Finding best random state:-**

****

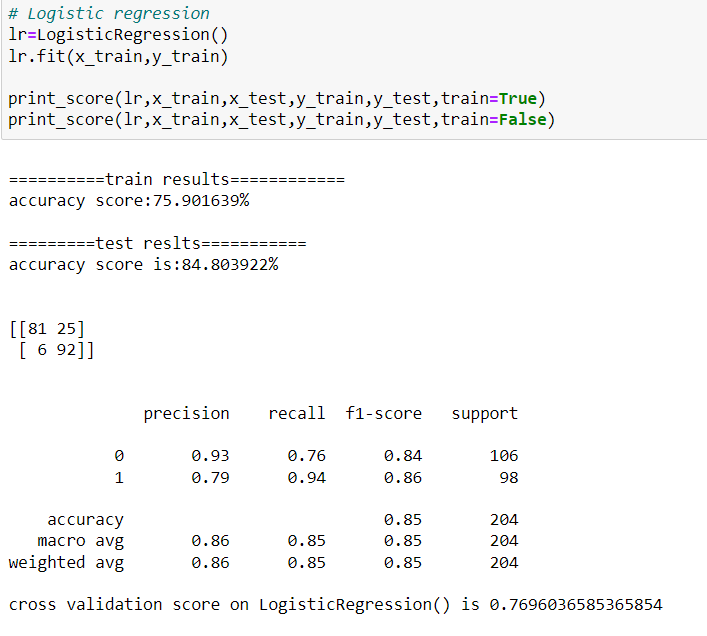
**5.Building machine learning model**

We know that our problem classification type we have to check accuracy score, confusion matrix and classification report for every model and also we check cross validation score for every model then we will select best model apart from all the models

I am creating one function instead of writing the code every time for every model

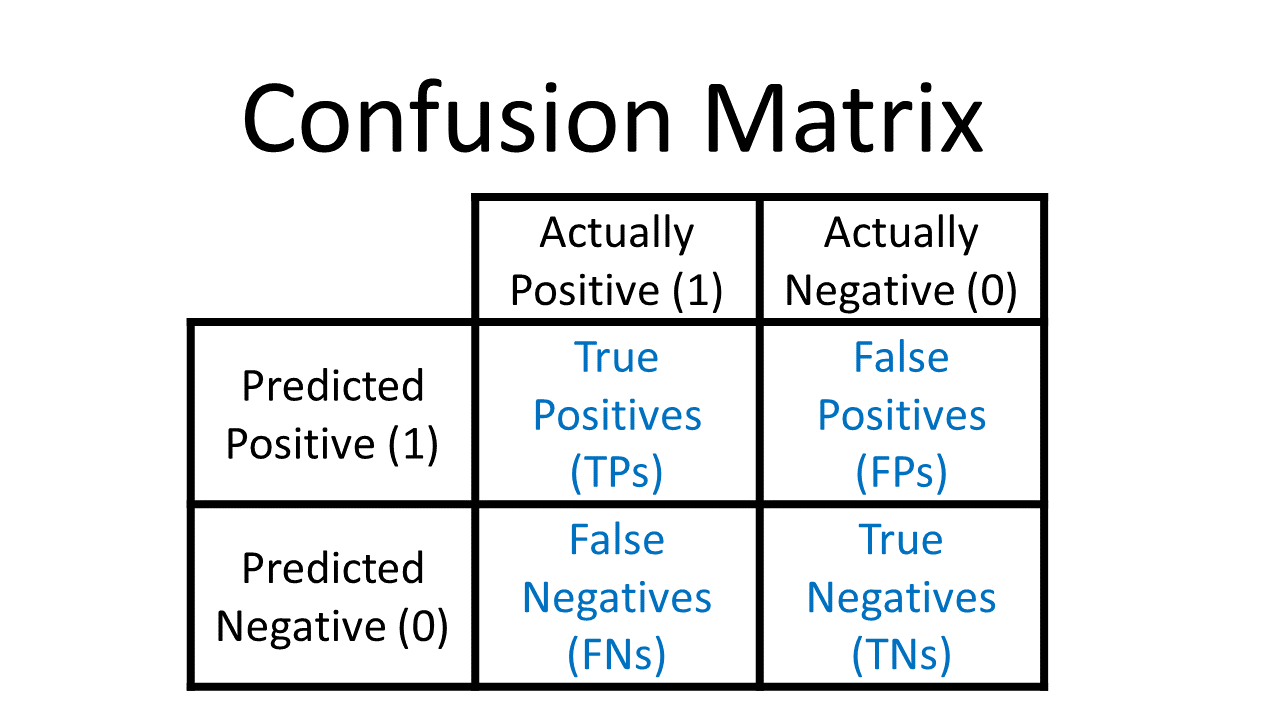
****

* Logistic Regression

****

Now let’s understand some terminologies present in classification report and confusion matrix

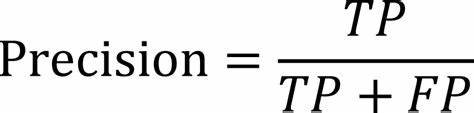
**Confusion matrix:-**

****

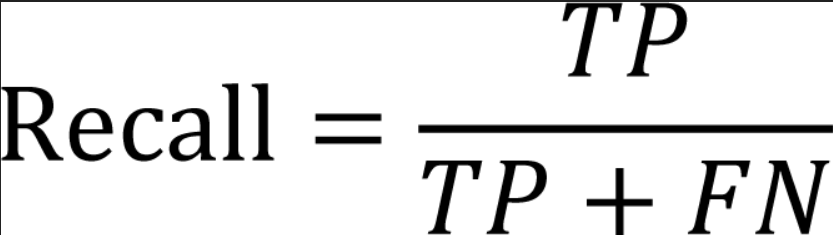
* **True positive(TP):-**The values which were actually positive and were predicted positive
* **False positive(FP):-**The values which were actually negative but predicted as positive it also known as type-1 error
* **False negative(FN):-** The values which were actually positive but predicted as negative it is also known as type2 error
* **True negative(TN):**- The values which were actually negative and predicted values also negative it also known as specificity

**Classification report;-** Let’s understand some important terminologies which are present in the classification report

* **Precision:-** It is a measure of amongst all the +ve predictions how many of them were actually +ve



* **Recall:-**It is measure of from the total no of +ve results how many +ve’s were correctly predicted.

****

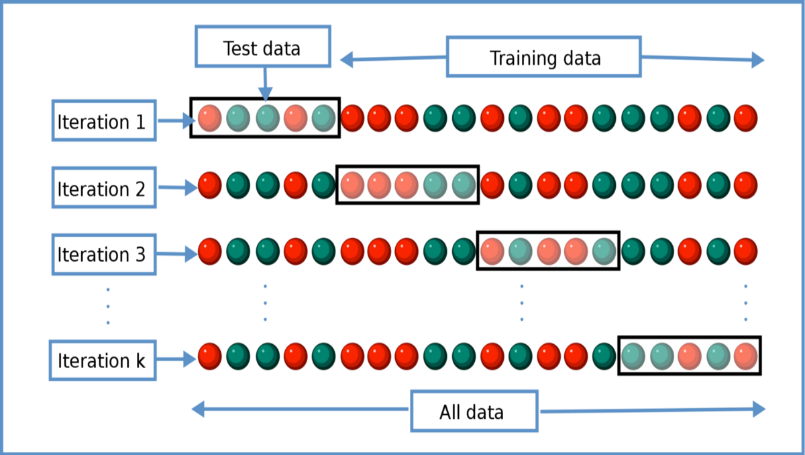
* **F1 score:-**It is hormonic mean of precision and recall



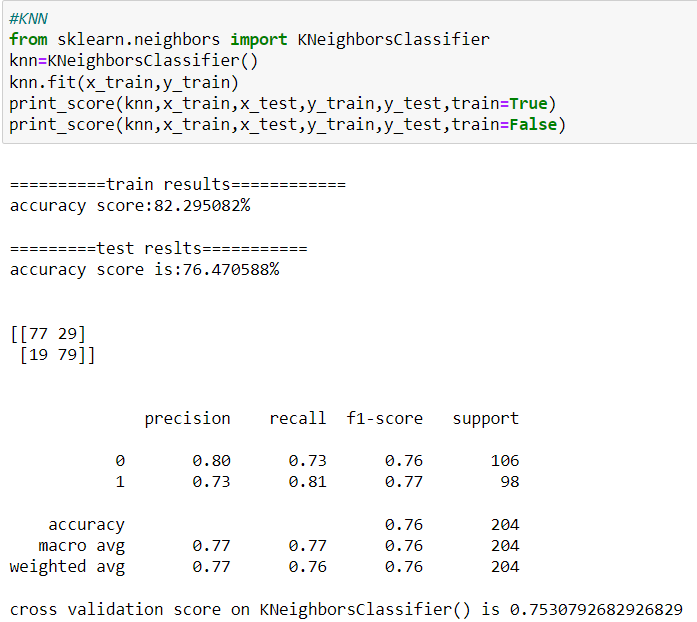
**cross validation:-**in this technique we reserve a particular sample of the data set which was not part of the training data set

let’s see how cross validation works behind the scene step by step

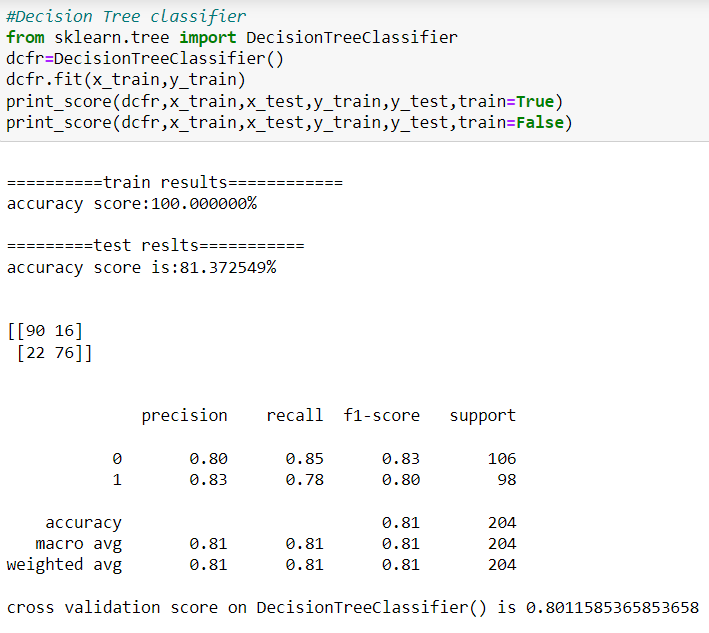
* It holds subset of the data set as validation set
* Using the rest data set train the model
* Test the model using hold subset of the data set



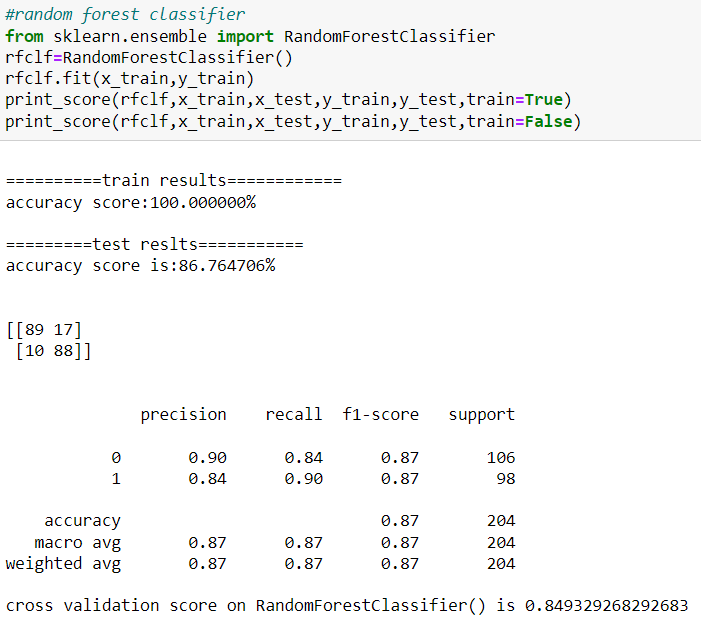
* KNeighborsclassifier

****

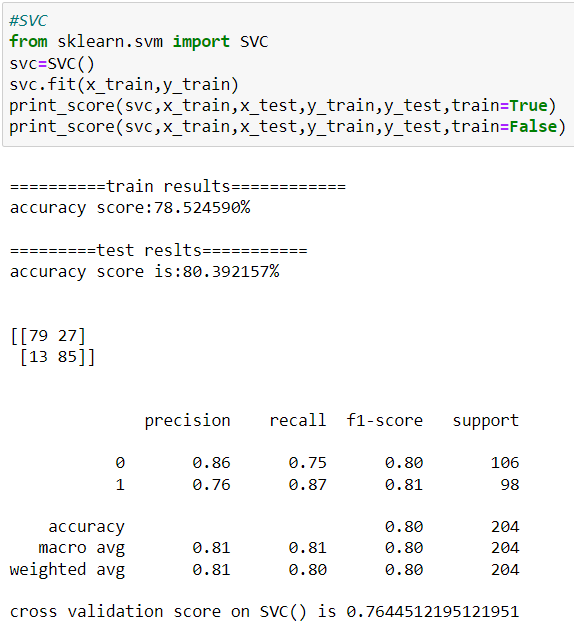
* Decision tree classifier



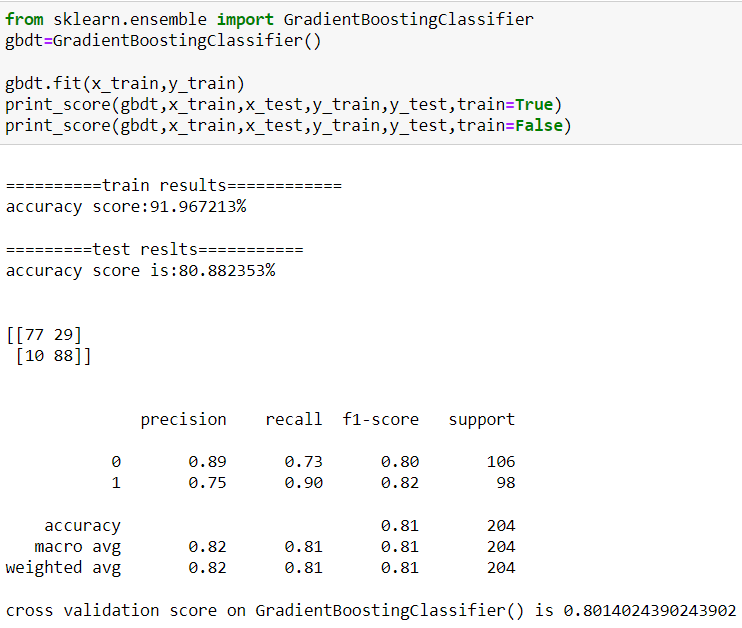
* Random forest classifier



* Svc

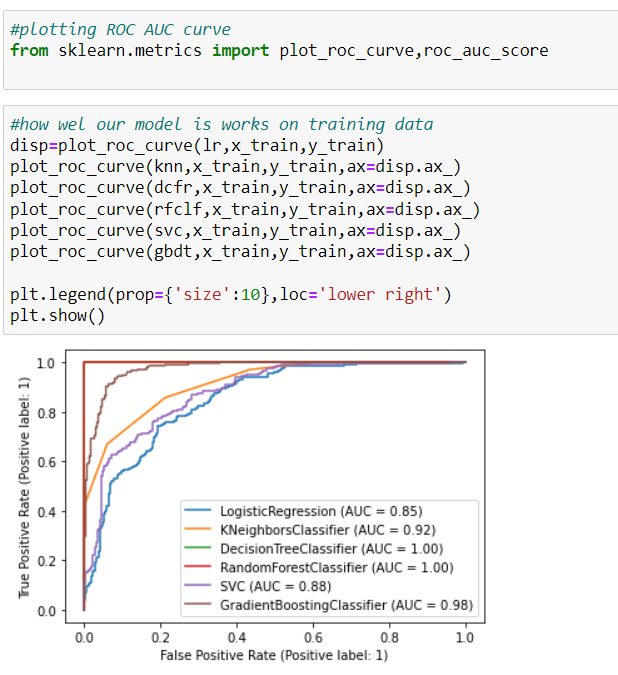


* Gradient boosting classifier

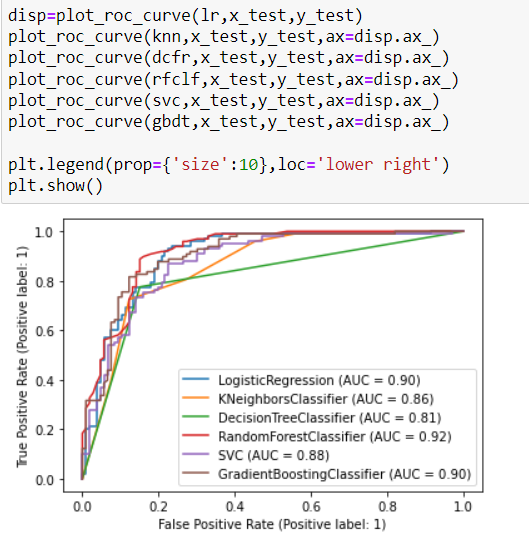


**Plotting ROC AND AUC CURVE :**-

* **ROC** stands for receiver operating characteristic curve it is a graph that showing the performance of a classification model at all the classification thresholds, this curve plots two parameters
* True positive rate
* False positive rate
* for train data

****

* for test data



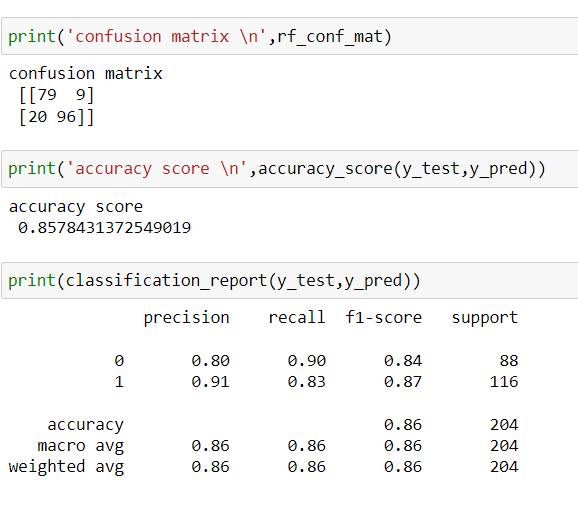
from the above plots randomforestclassifier covers maximum area under the curve so I choose Random forest is the final model

**Hyperparameter tunning:-**

* Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm.

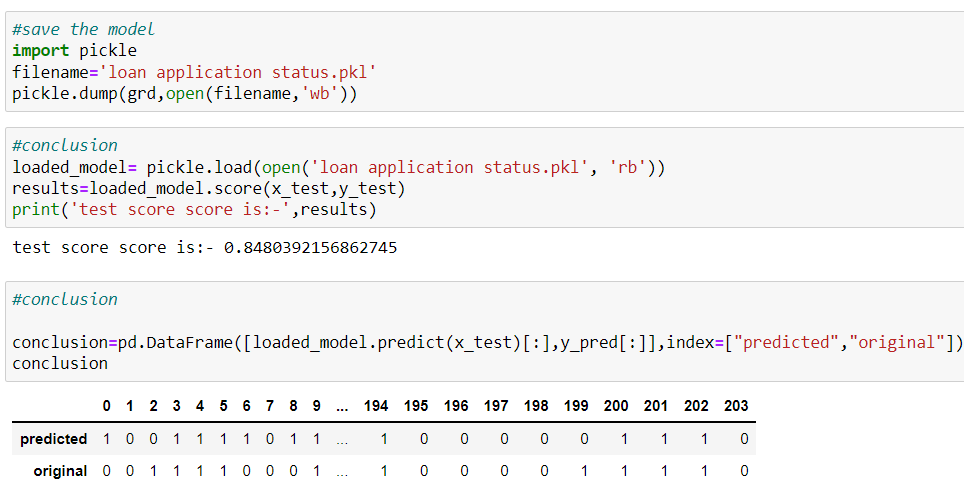
****

After tunning with best parameters let check again with confusion matrix and classification report



**6.Concluding remarks**

* Saving the model by using pickle

****

**Final conclusion**

predicted values shows similar as original values so our model performs well, so it helps banks to save the time and money to decide approve loan or not and also it helps customers also to save their time and hopes, this model reduces the manual interruption between banks employees and customers, Hence by using Machine learning techniques we can solve this problem & reduce manual efforts.