

**Loan Application Status Prediction**

Submitted By

Vicky

* **Problem Definition:**

This dataset includes details of applicants who have applied for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, Property Area and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

Using Data Science we have to predict the loan status, whether loan application will have approved or not as per the applicant given details.

* **Data Analysis:**

According to given dataset we have following features:

* Loan\_ID
* Gender
* Married
* Dependents
* Education
* Self\_Employed
* ApplicantIncome
* CoapplicantIncome
* Loan\_Amount
* Loan\_Amount\_Term
* Credit History
* Property\_Area
* Loan Status(Target Variable)

Know we are checking relationship between all feature and target variable, how Loan\_status(target variable) depends on others features.

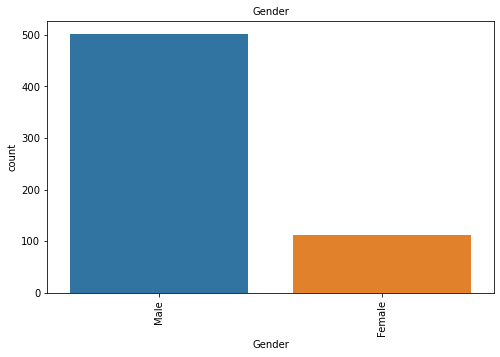
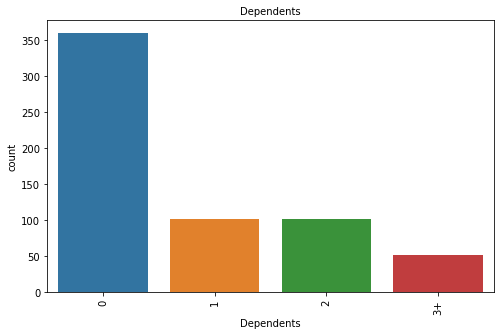
Applicant Income is the most important things to take loan without any trouble from anywhere. If we have permanent income source and we are earning income continuously so we can take loan easily. The Financial institution always takes decision on the bases of application status.

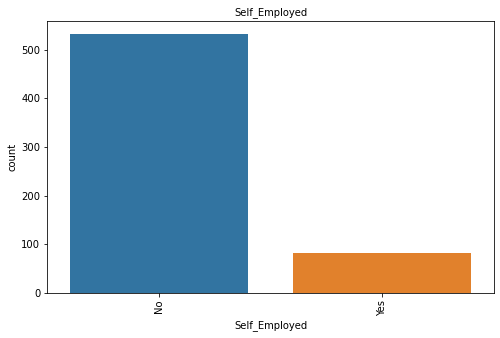
According to our loan application problems, we are trying to analysis data with different-different approches.

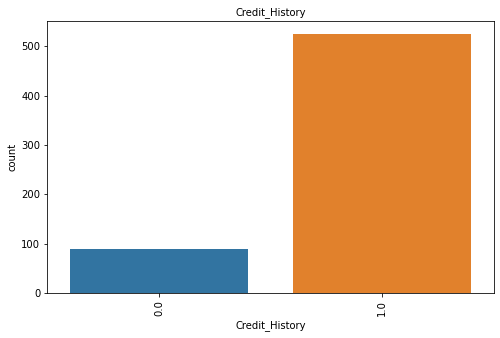
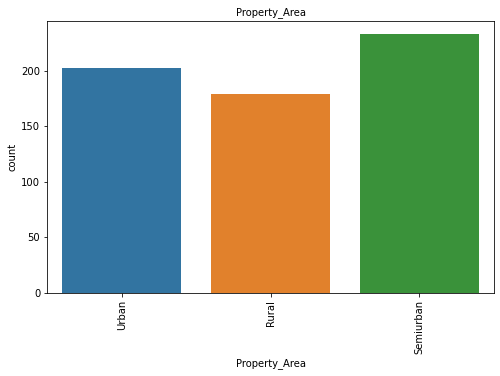
1. If we check shape of our data we have 614 rows and 13 features.
2. All features are not numerical. Applicant Income, Coapplicant Income, Loan Amount, Loan Amount term, credit history are the numerical features and others are categoricals. We will deal with categorical features during Features Engineering.
3. The Null values are present in all columns except Education, Applicant Income, Coapplicant income, Property Area, and Loan status columns.

**Correlation:**

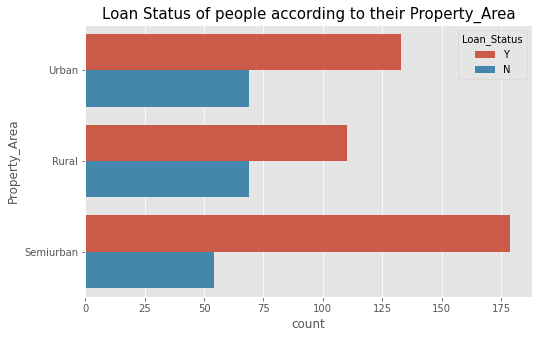
In correlation we are trying to observe which features is highly related with other features. According to above Applicant Income is highly correlated with Loan Amount that means Loan Amount totally depends on applicant income. If applicant earning high income so they can take loan easily. As we are getting others things also have relationship with each other but it is very less.







* On the bases of above graphically analyze the Male applicant are much higher than Female. It means, mostly male are only income source in family. Accordingly Married applicant are higher than Unmarried. It means Married persons needs more loan than unmarried persons. One another factor is also associate with Married applicant he has more responsibilities that’s why married applicant is higher than unmarried.
* As we know on the bases of graphical evaluation. The applicants who have no dependents are much higher. The Banks and Financial Institution evaluate applicant loan application on dependency of family member, that’s why the person who has no dependent have more chance to approved loan.
* If we talk about Self-employed and Salaried applicant. Salaried applicants are higher than self-employed applicants. Any lender before approving any category of loan checks if the borrower has the capacity to repay the loan. This is the thumb rule of any financial or Non-Banking Financial Institution (NBFC) and in most cases, the loan granted by a lender is 5 – 6 times higher than the income of the borrower. However, the amount granted to the borrower depends on the nature of his income, sources of income and the bandwidth of his income. Banks and Financial Institutions have low risk on Salaried person that’s why they have more chance to got loan.
* Whenever you apply for a loan, banks check your CIBIL Score and Report to evaluate your credit history and credit worthiness. The higher your score the better are the chances of your loan application getting approved.

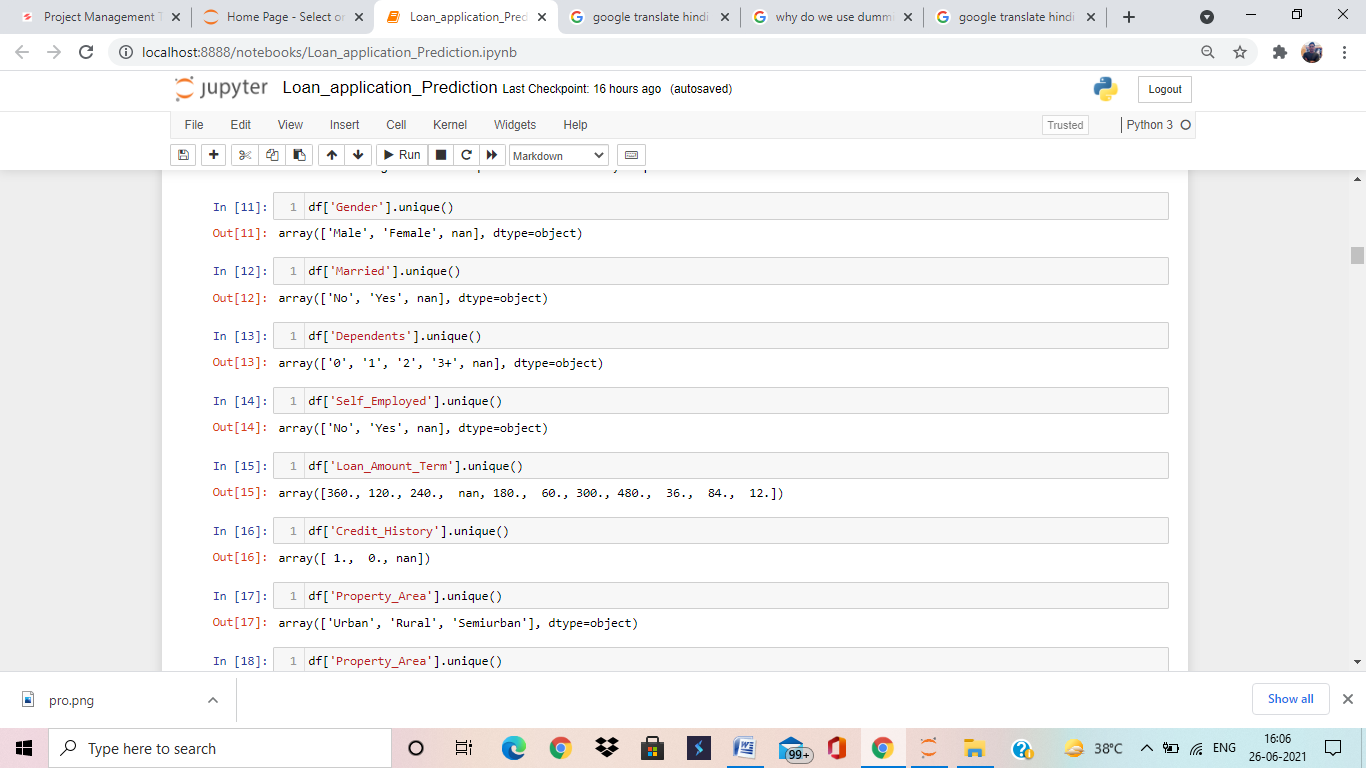


* According to above graphical evaluation the applicants how belongs to Semi-urban Area are higher than other areas.
* **Pre-Processing Pipeline:**

**Feature Engineering:-**

Definition:- *Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.*

If we have more refined data then our model will perform well.

**

We checked the Nan present in our dataset so we have to remove that values from data for better model accuracy. Using “**Simpleimputer most frequent functionality**” for removing nan values with most frequent values.

Some time in single features many categories are present for example: Property\_Area and Dependents columns have other category also present, so here we are extracting those details for better prediction or model building. We are creating new features using Dummies functionality.

Using Z-Score Function to remove outliers from data. As we are getting maximum outliers are present in Applicant\_Income and Coapplicant\_Income columns. It is majorly important to clean data.

Feature Splitting into Independent and Target variable is also part of pre-processing pipeline. We have to divide data into target variable and independent variables. Here Loan\_Status is our target variable.

Using MinMaxScaler to scaling the data into -1 to +1. Some time our features values not in equal scaling (same scale) so that time we have to scale the all features (except target variable) into one scale.

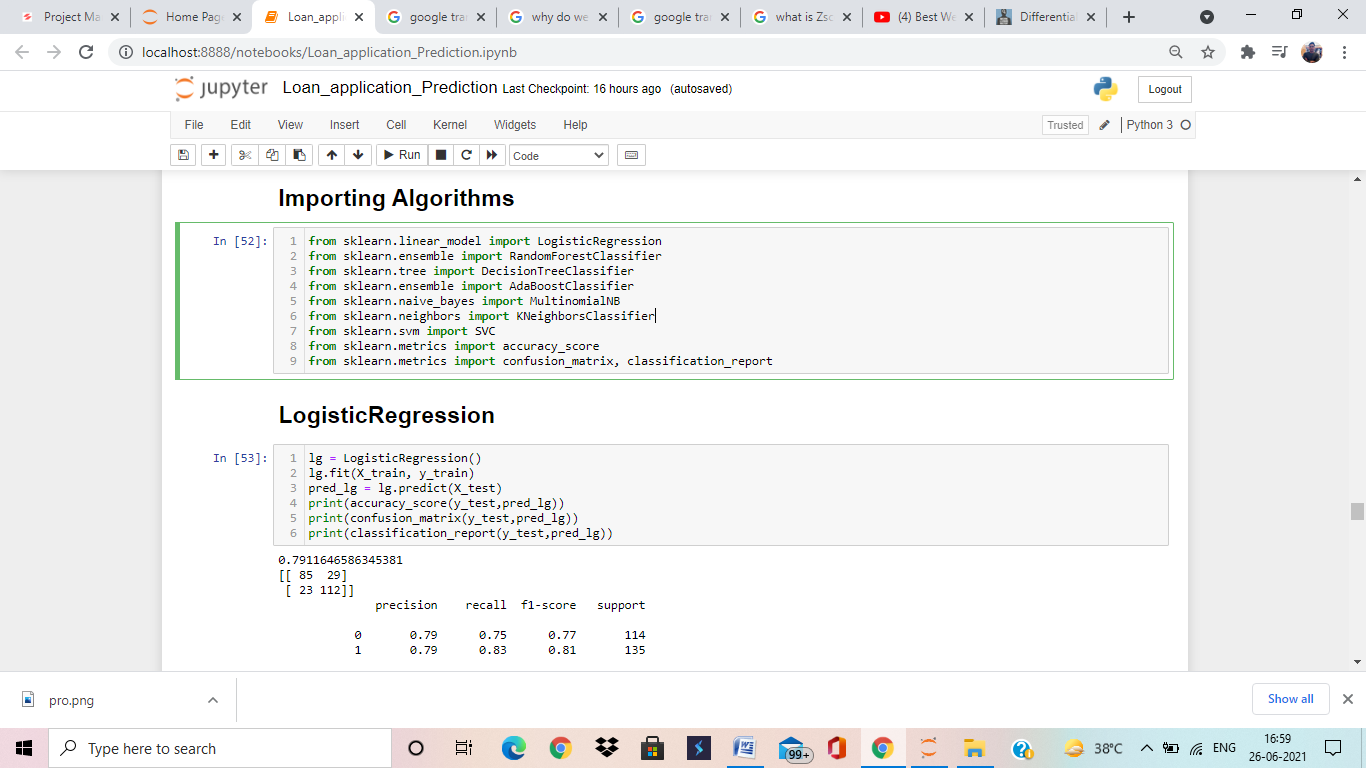
Before going to train model we have to check first target variable data is balance or imbalance. If the data is imbalanced, it can cause the overfitting and bias in the model prediction. In this case our target variable is imbalanced so we are using Over\_Sampling function to balance the data.

After data cleaning, features engineering, encoding and data scaling, know we are entering in Model building phase. We have already done different-different techniques on data.

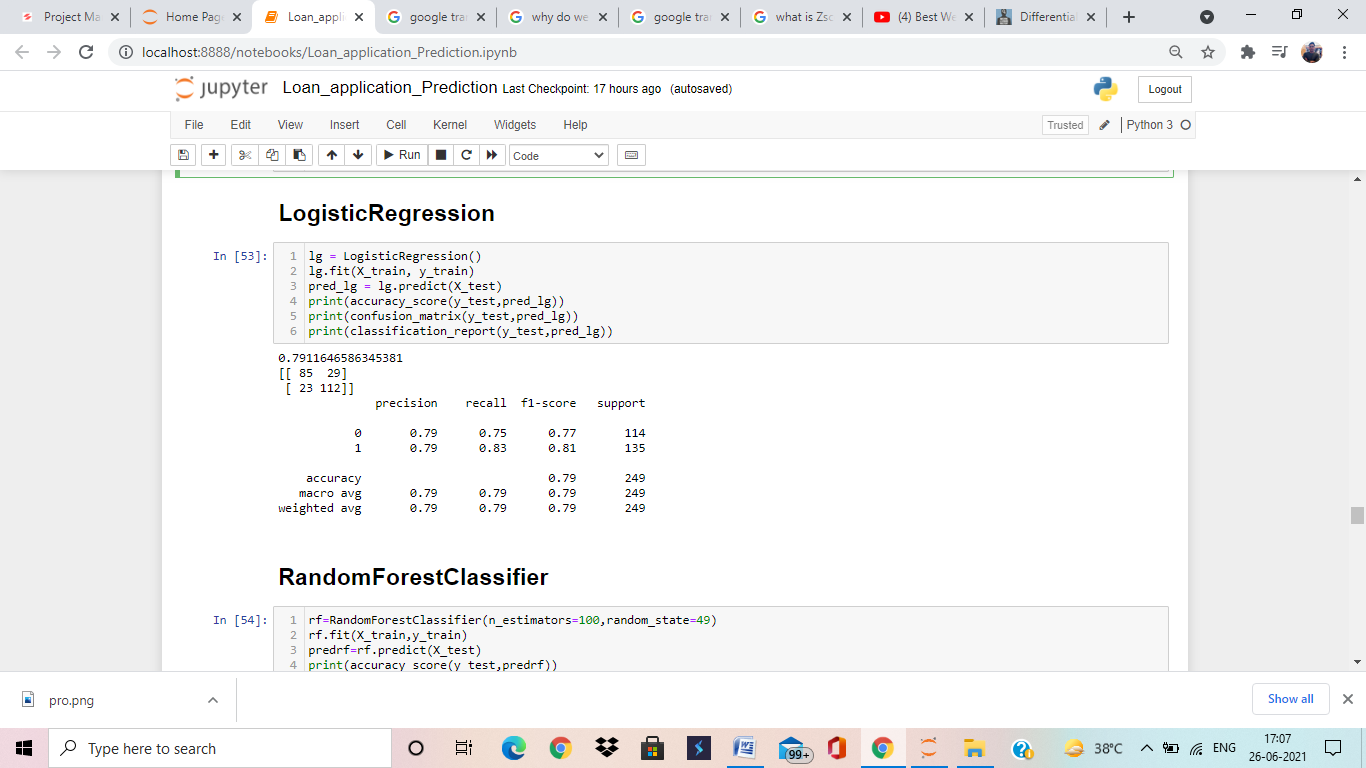
* **Model Building:**

This is Classification problem, so we are applying multiple regression algorithms and Classification algorithms.

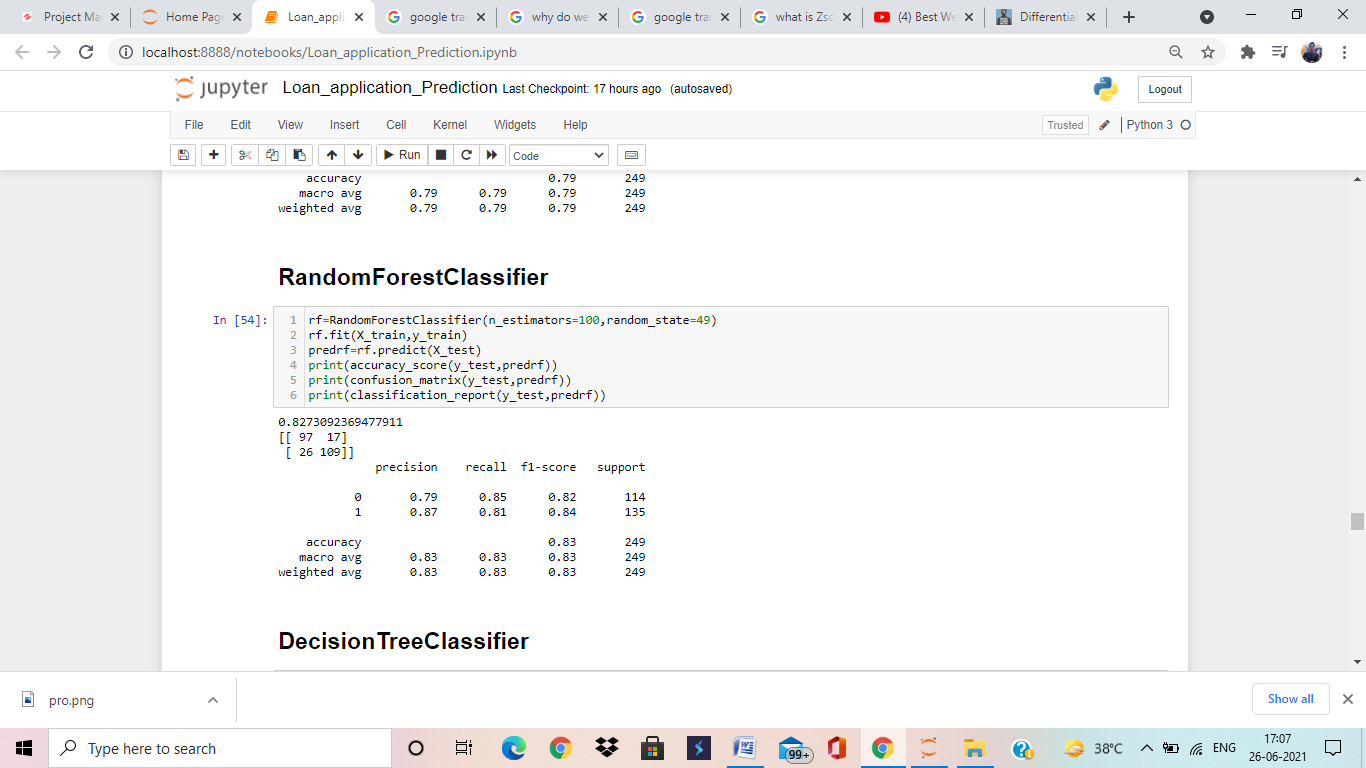
Take a look what algorithms we are applying.



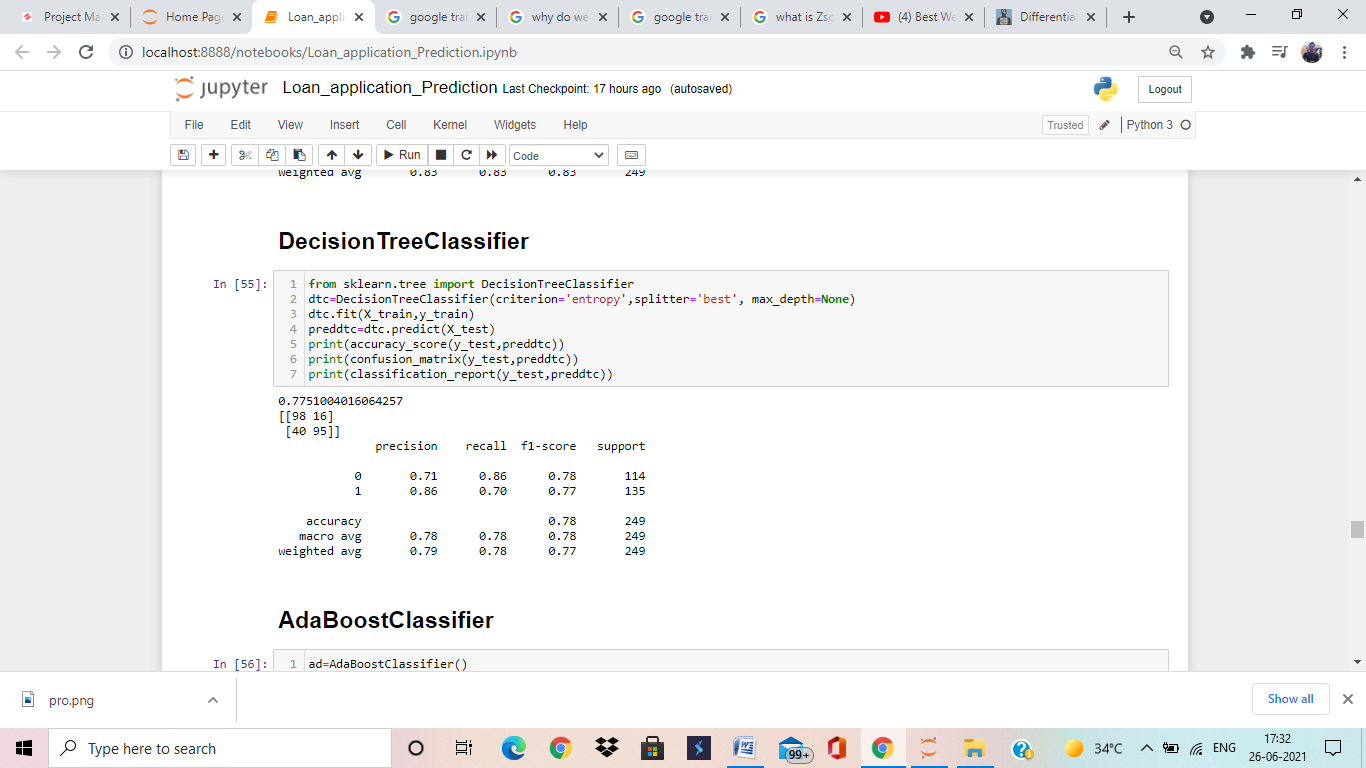
Know we will try to train our model with different-different algorithms for better accuracy.



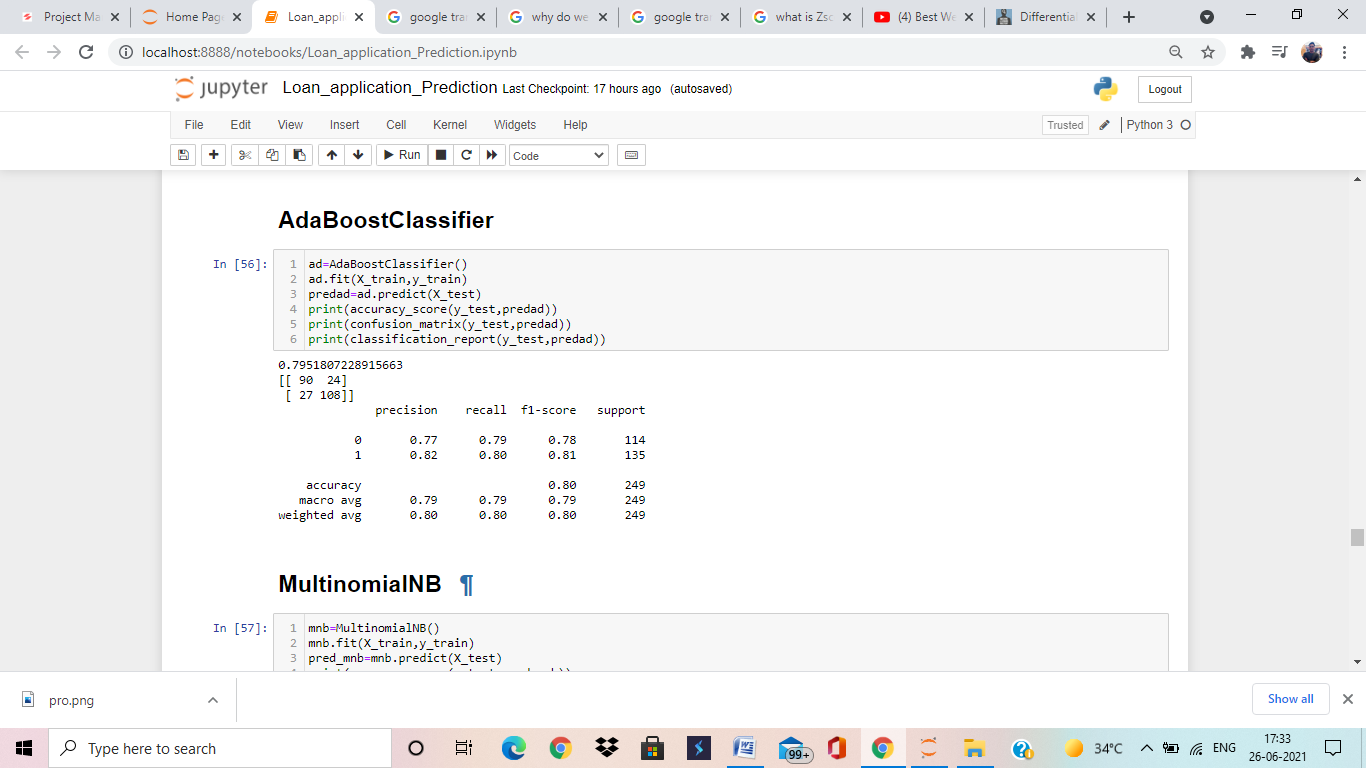
Here, we are applying Logistic Regression algorithms and Logistic Regression is giving us 79% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



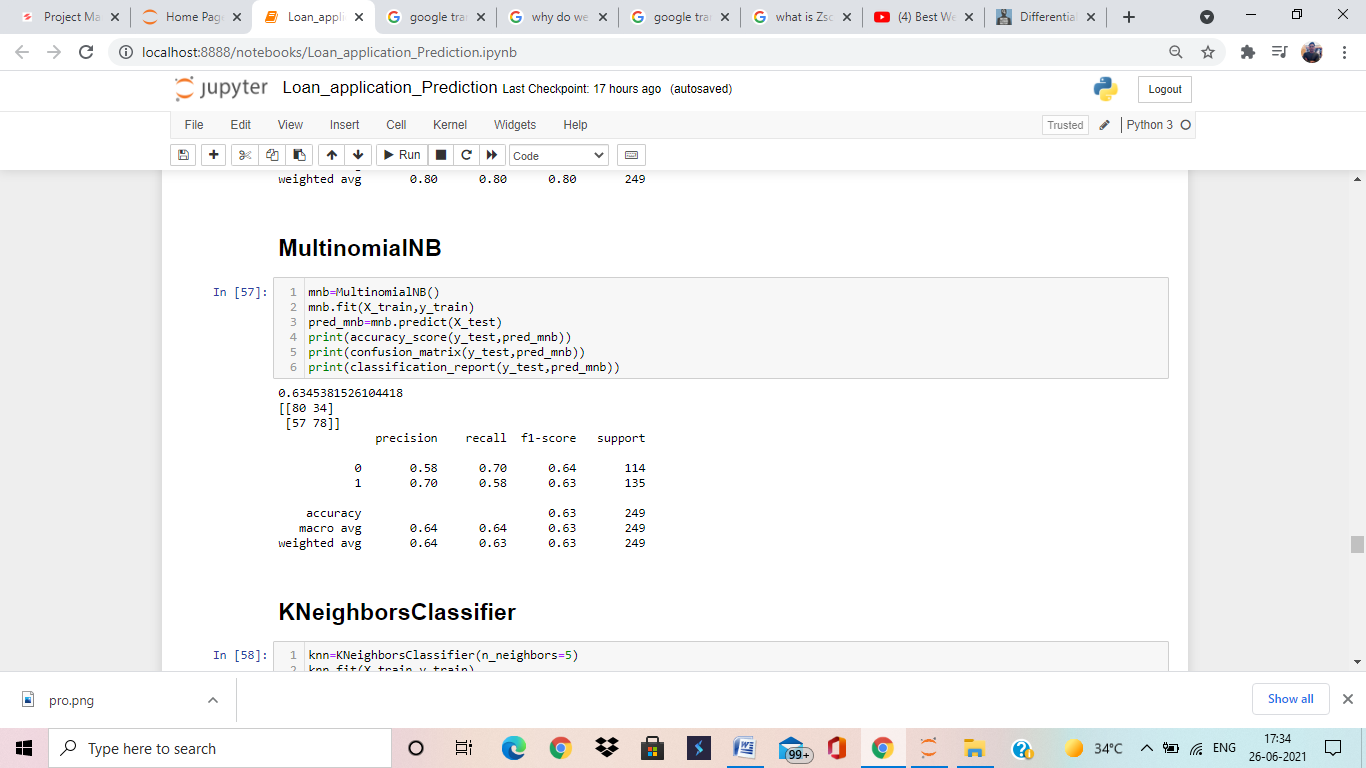
Here, we are applying Random Forest Classifier algorithms and Random Forest Classifier is giving us 82% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



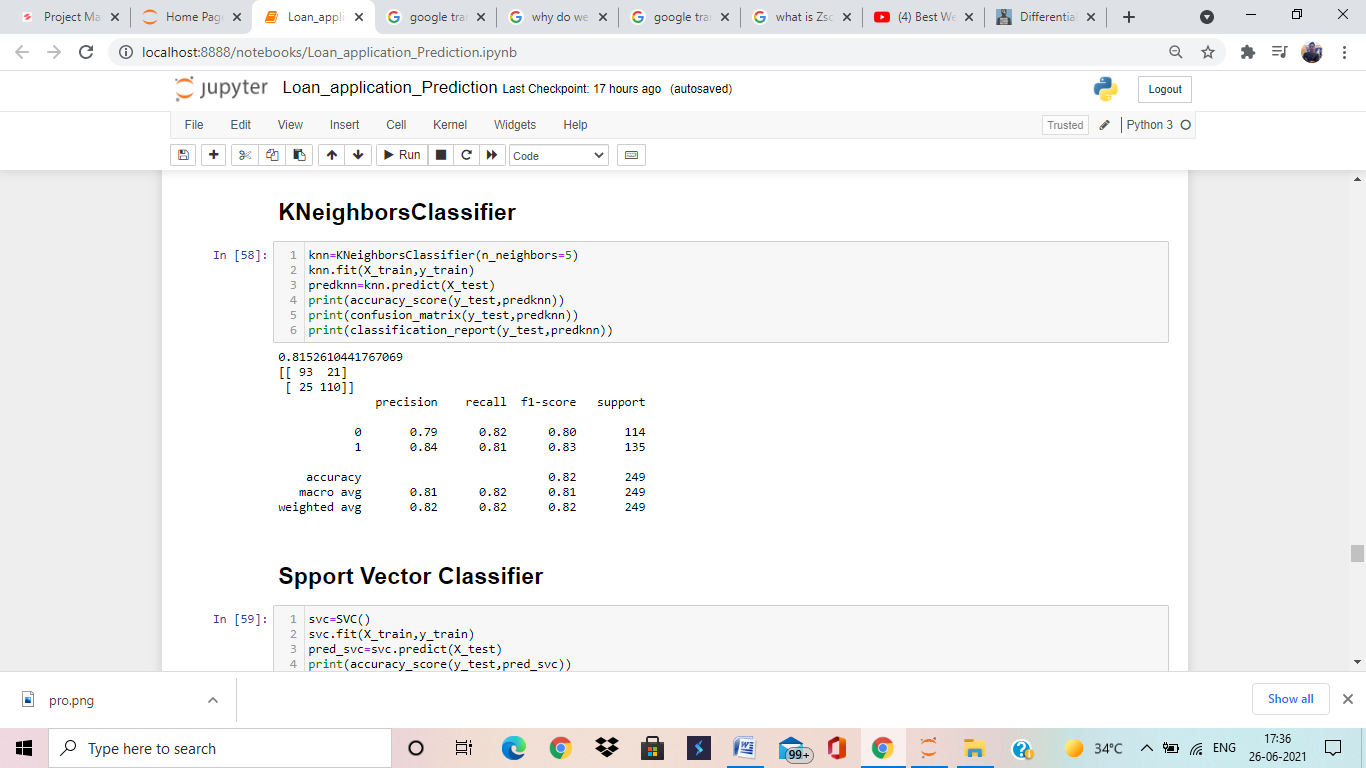
Here, we are applying Decision Tree Classifier algorithms and Decision Tree Classifier is giving us 77% accuracy, which is good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



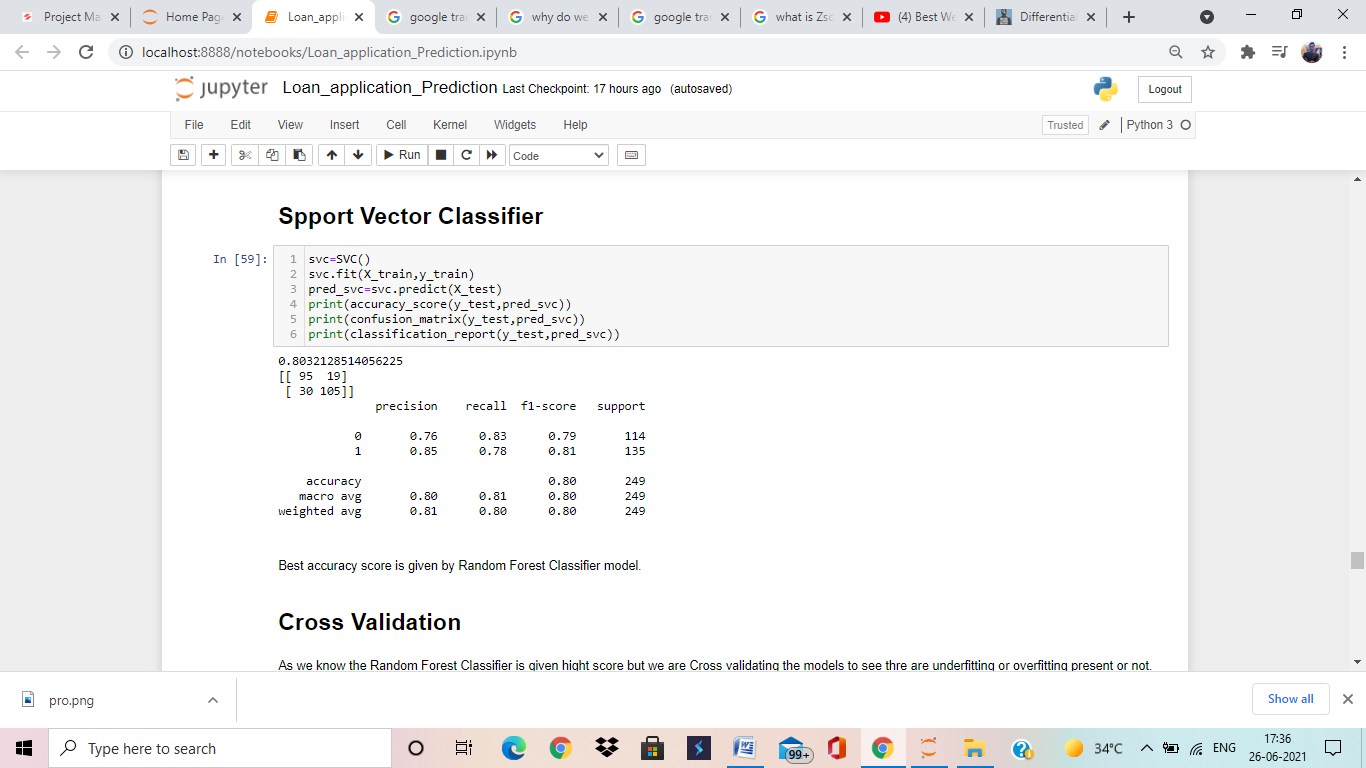
Here, we are applying AdaBoost Classifier algorithms and AdaBoost Classifier is giving us 79% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



Here, we are applying MultinomialNB algorithms and MultinomialNB is giving us 63% accuracy, which is lower than others.

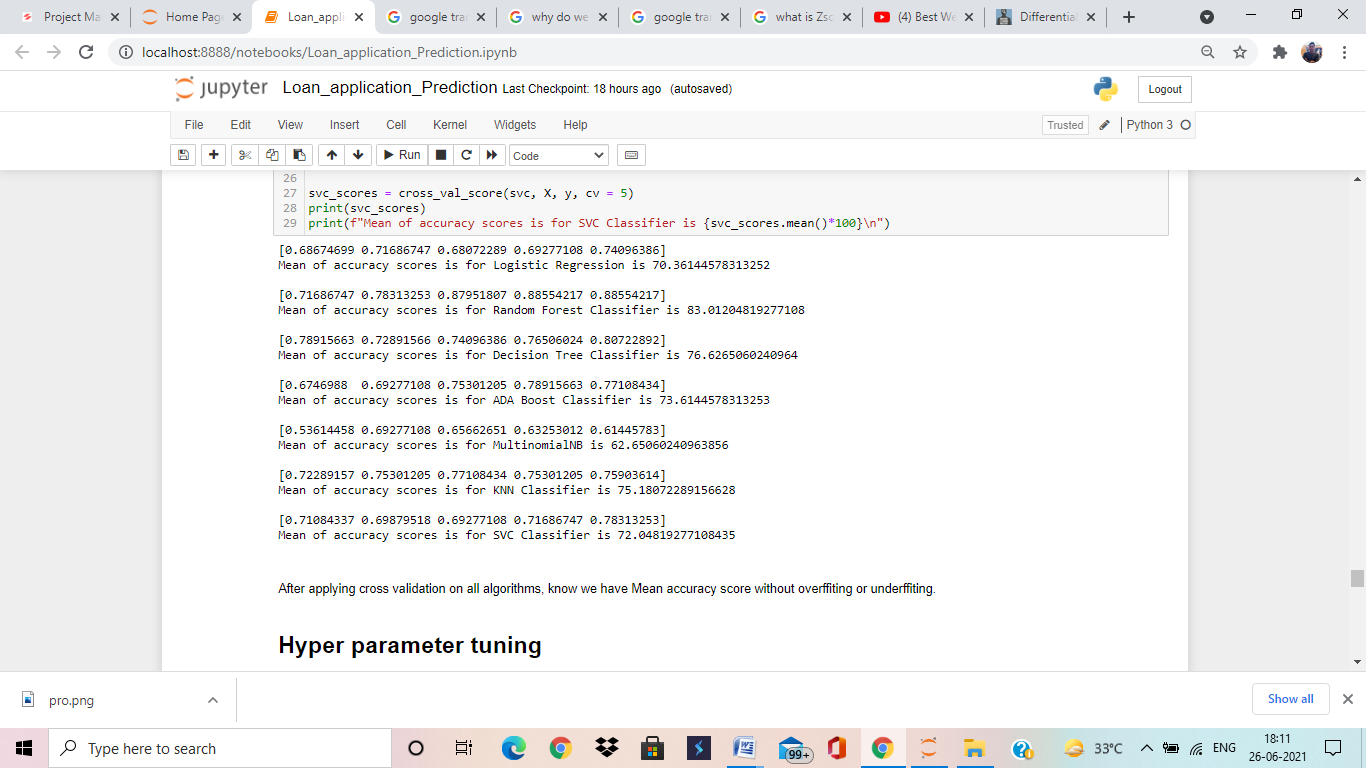


Here, we are applying KNeighbors Classifier algorithms and KNeighbors Classifier is giving us 81% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.

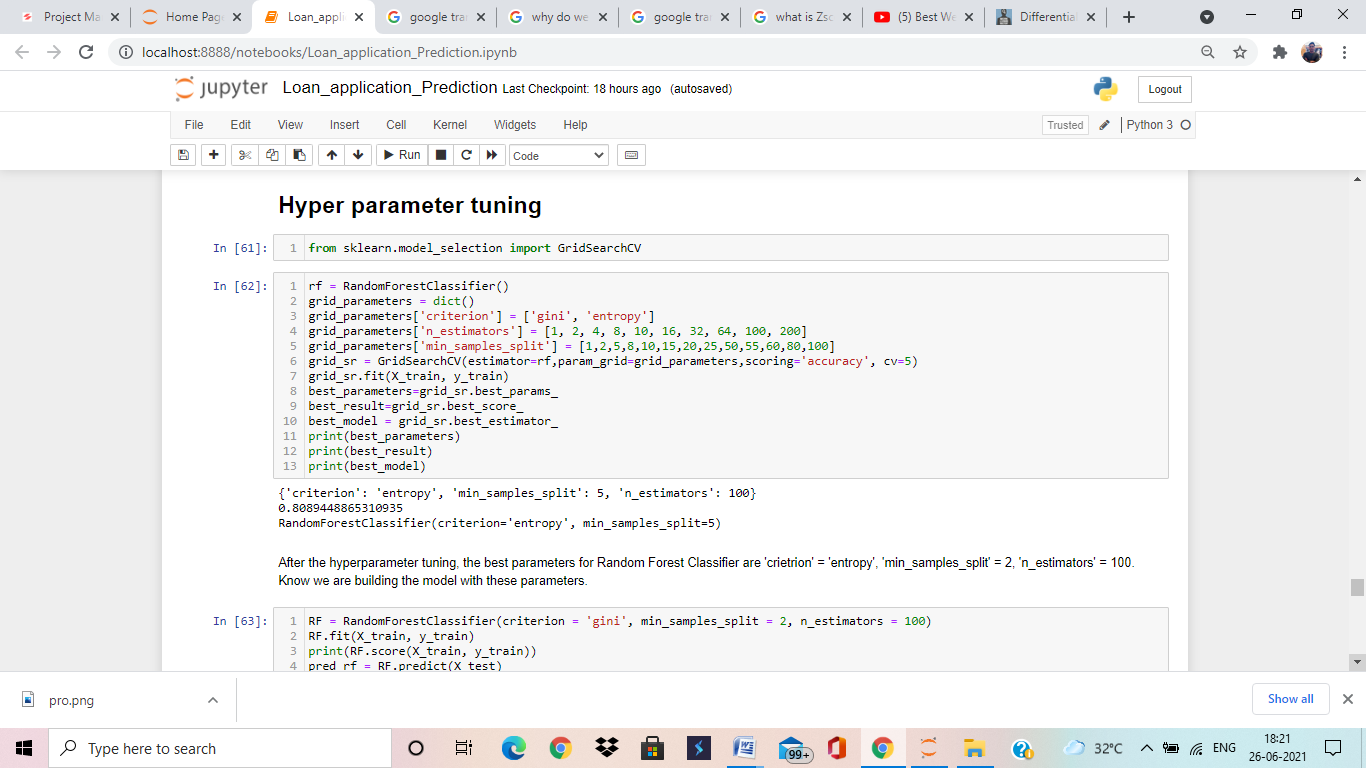


Here, we are applying Support Vector Classifier algorithms and Support Vector Classifier is giving us 80% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.

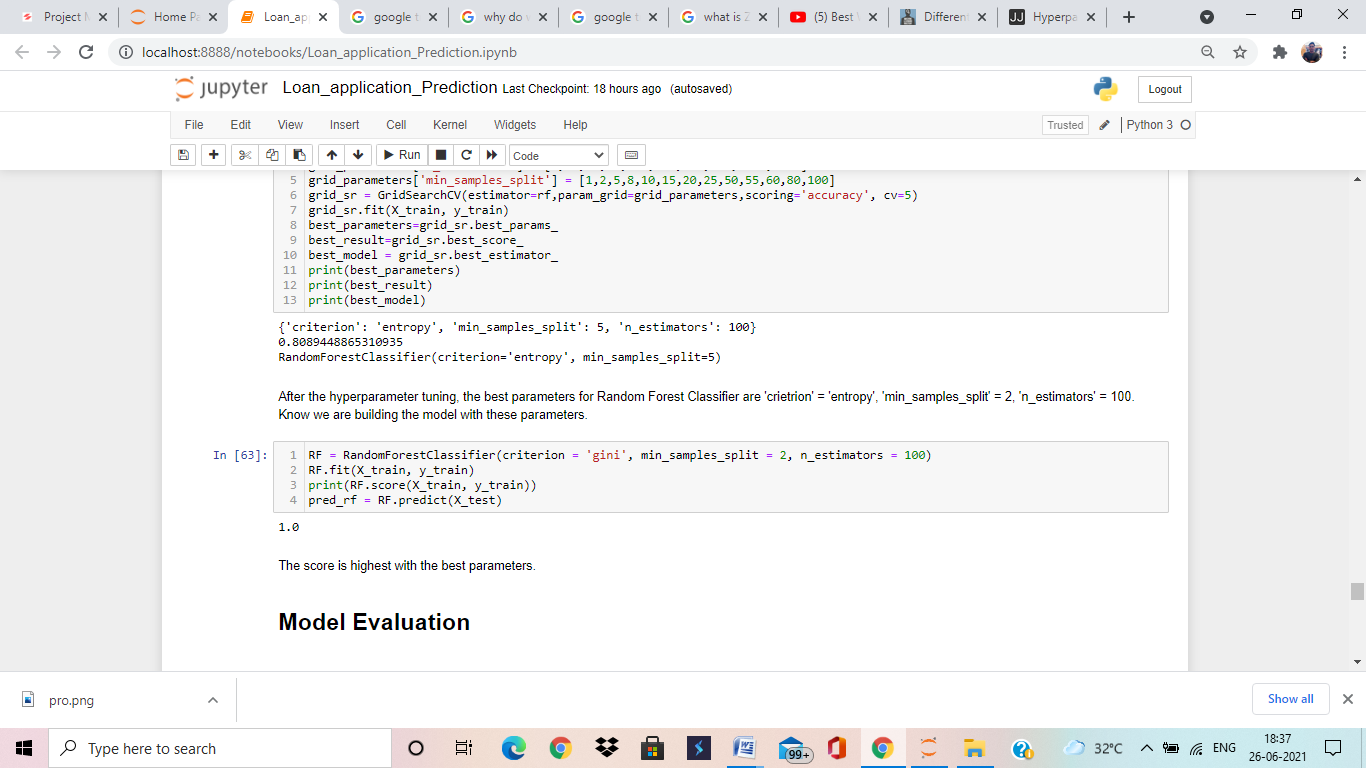
**Cross Validation**



After applying Cross Validation on all algorithms know we have accuracy without over-fitting/under-fitting. Cross validation basically use to remove High Bias/Variance from data to train model in proper manner. All algorithms performing well except MultinomialNB. It is giving low score.

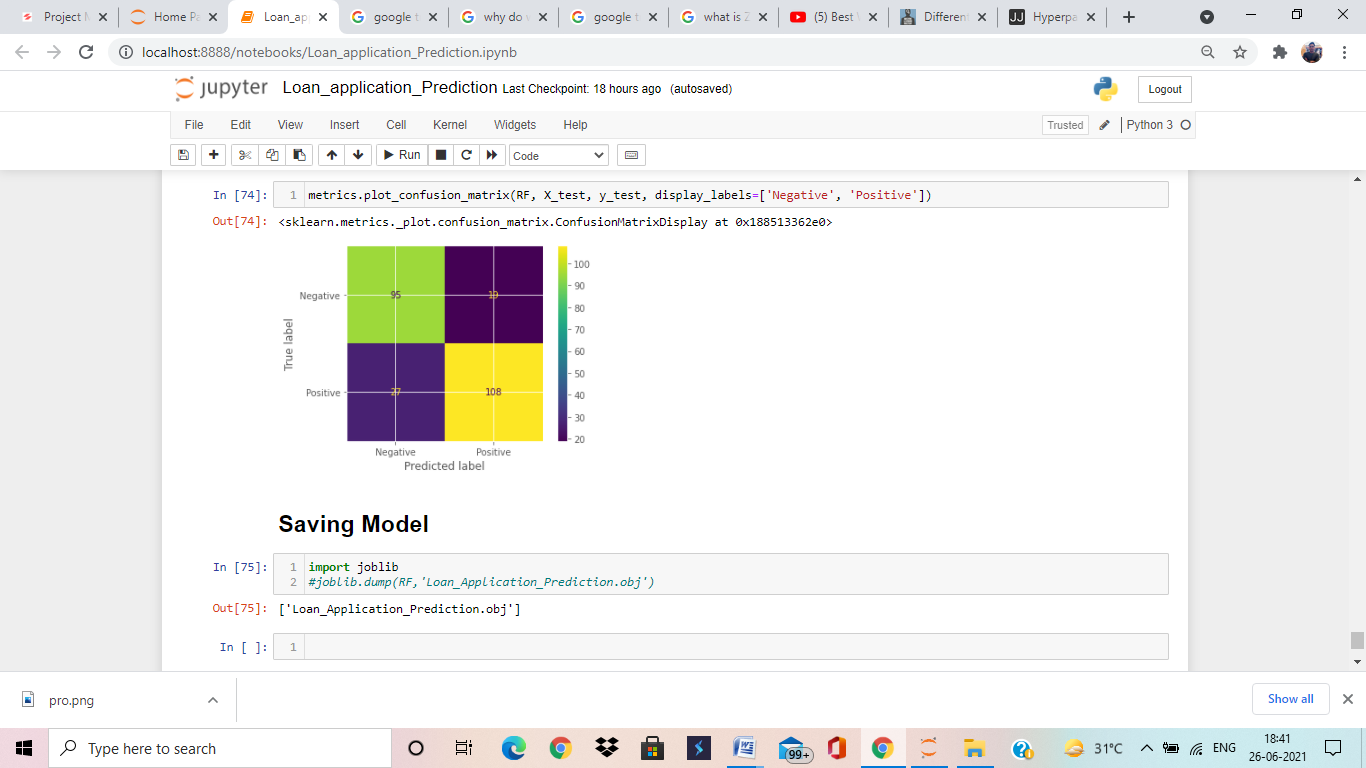
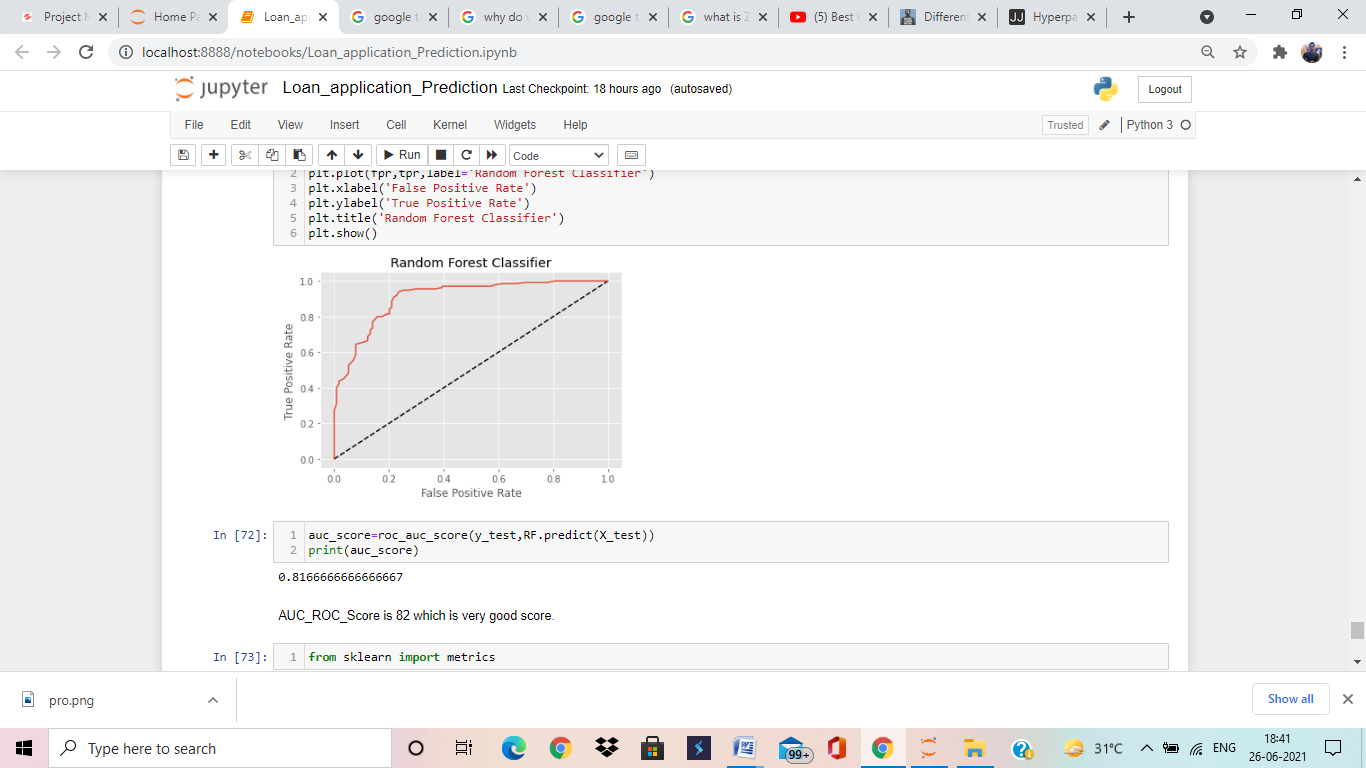


We are applying Hyper Parameter Tuning to find best parameters to get good accuracy and train model with these parameters. According to Hyper Parameter Tuning we find Random Forest Classifier parameter(Criterion: ‘entropy’, min\_samples\_split:5,n\_estimators:100) giving us best accuracy. Know we are building model with these parameters.



Here we build model with best suitable parameters and finally we are getting 100% accuracy by Random Forest Classifier algorithms.

* **Model Evaluation**



Here, we are try to evaluating the model with metrics and as we are getting very good accuracy apporx.82%. My model performing well knows we can save this model.

* **Conclusion**

Random Forest Classifier is the best model. It is performing very well, we applied all techniques on this model. We trained this model with best parameters and at lost we check accuracy by metrics.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*