

**Insurance Claims-Fraud Detection**

Submitted By

Vicky

* **Problem Definition:**

Insurance fraud is constantly raising concerns among insurance companies and regulators, due to huge losses caused to business every year. For every Rs. 100 premium earned by the insurance companies, they would be paying Rs. 213 as claims. Motor insurance which contributes almost one third to the non-life insurance in India is highly affected by fraud. Many insurance companies are showing loss due to high claim ratio in motor sector. The motor insurance market in India is growing at respectable rate, and this provides greater opportunity for insurance companies to tap this growing market. Nonetheless, the insurance companies face greater challenges in terms of reducing claim processing cost, and detection and control of fraud. Hence, the purpose of this paper is to explore the magnitude of the problem and methods to prevent and control fraud in motor insurance. And also study highlights ways to reduce claim processing cost. The paper reveals the scarcity of research of such a topical issue in developing economies such as India and suggests the urgent need to discuss among insurance companies about the effects of insurance fraud on the industry, with a view to tackle the problem.

It is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

We will work with some auto insurance data to demonstrate how we can create a predictive model that predicts if an insurance claim is fraudulent or not.

* **Data Analysis:**

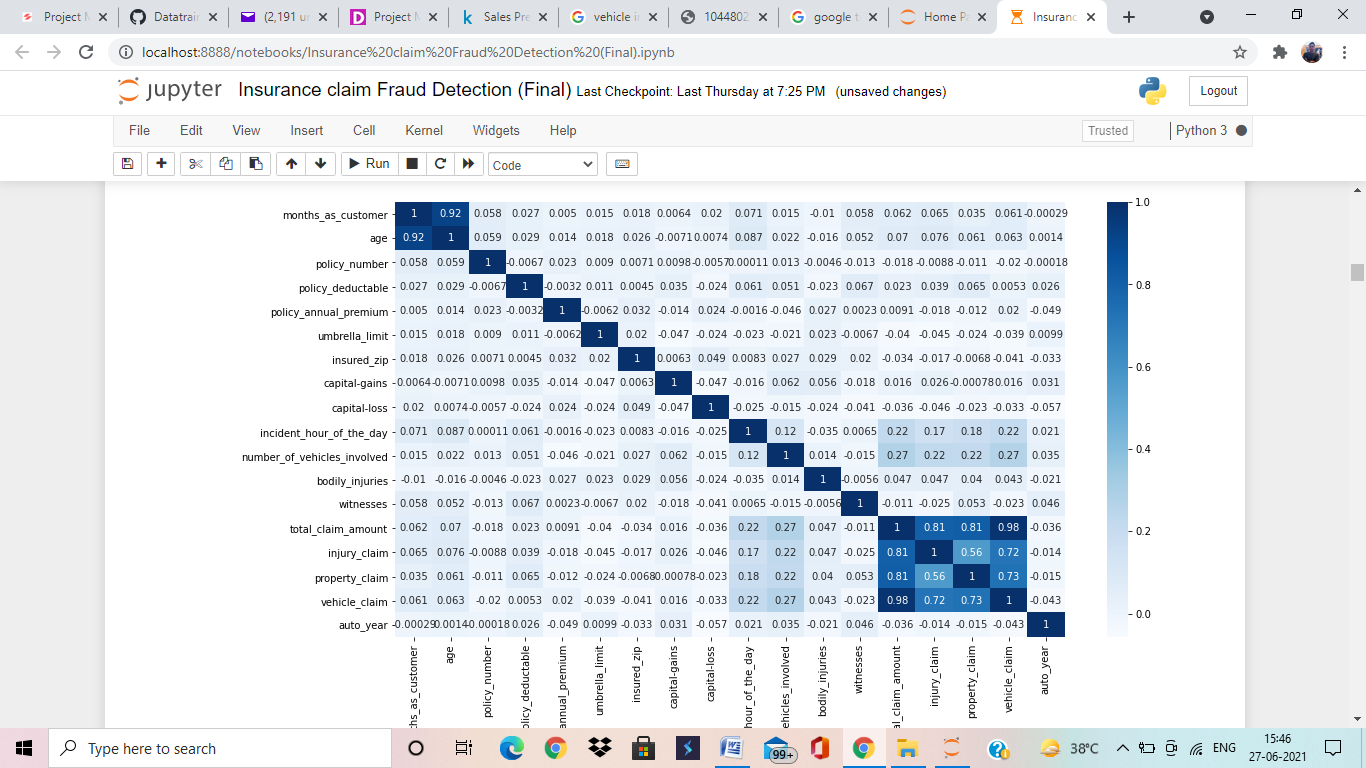
According to given dataset, we have following features:

* months\_as\_customer
* age
* policy\_number
* policy\_bind\_date
* policy\_state
* policy\_csl
* policy\_deductable
* policy\_annual\_premium
* umbrella\_limit
* insured\_zip
* insured\_sex
* insured\_education\_level
* insured\_occupation
* insured\_hobbies
* insured\_relationship
* capital-gains
* capital-loss
* incident\_date
* incident\_type
* collision\_type
* incident\_severity
* authorities\_contacted
* incident\_state
* incident\_city
* incident\_location
* incident\_hour\_of\_the\_day
* number\_of\_vehicles\_involved
* property\_damage
* bodily\_injuries
* witnesses
* police\_report\_available
* total\_claim\_amount
* injury\_claim
* property\_claim
* vehicle\_claim
* auto\_make
* auto\_model
* auto\_year
* fraud\_reported (Target Variable)

According to given dataset, we will try to analysis the data with different-different approches.

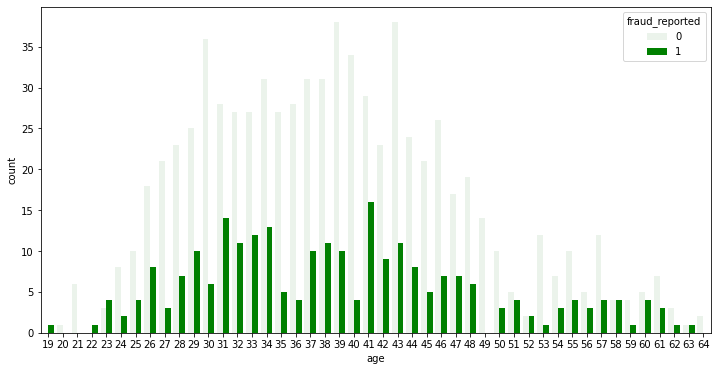
1. If we check shape of our data we have 1000 rows and 40 features.
2. All features are not numerical in nature, we have also some categorical features as well. We will deal with categorical features during Features Engineering.
3. The Null values are not present in any features.

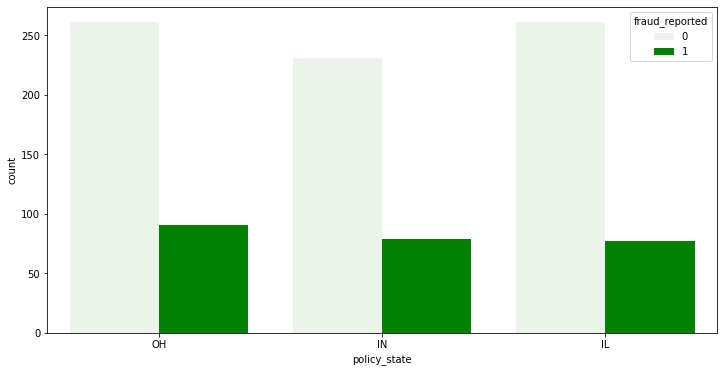
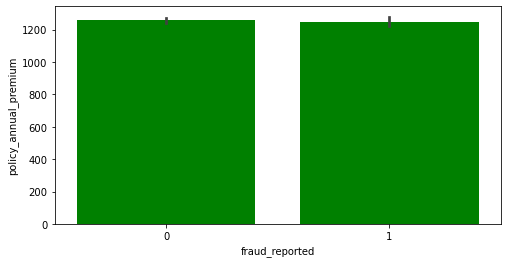
**Correlation:**

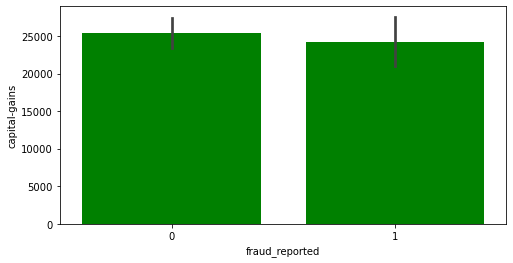
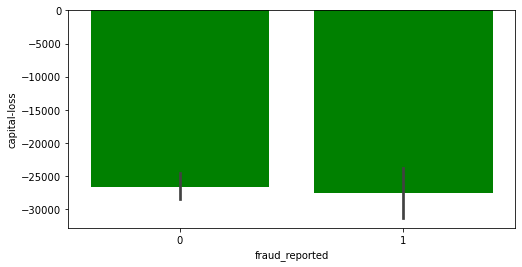
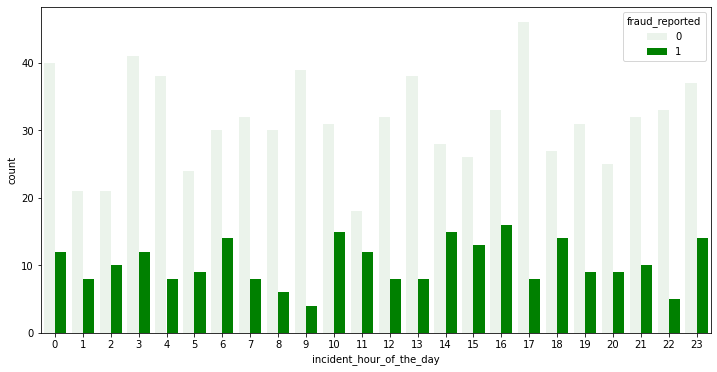


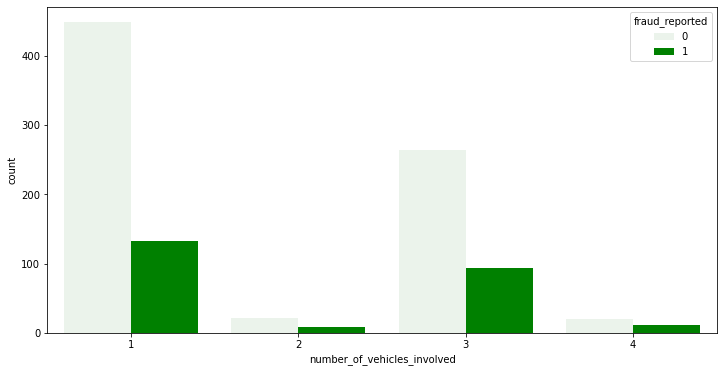
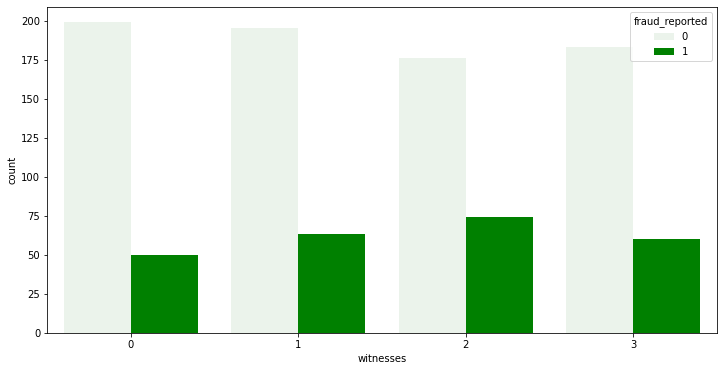
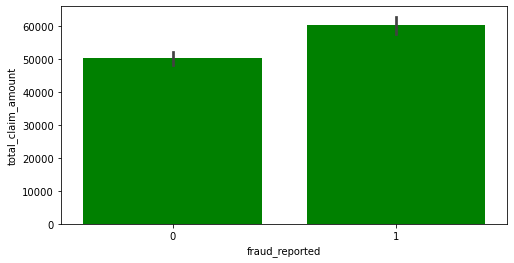
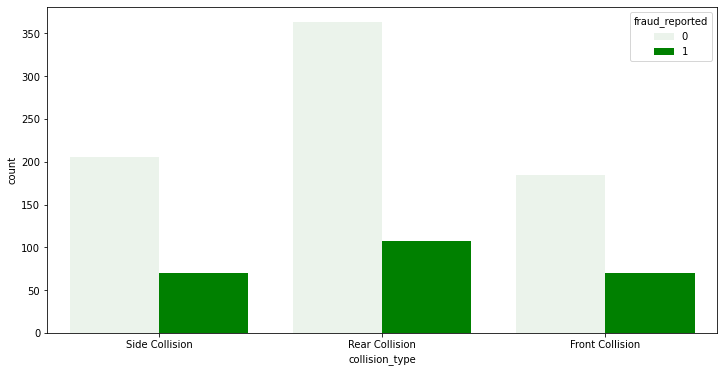
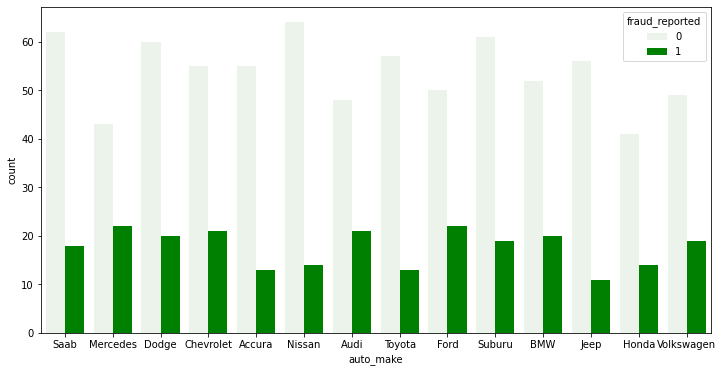
In correlation we are trying to observe which features is highly correlated with other one. According to above Total\_Claim\_Amount is highly correlated with Injury\_Claim, Property\_Claim, Vehicle\_Claim and Months\_as\_customer is highly correlated with Age.

Total Claim Amount is totally depends on different-different claims as we are getting from correlation. As we are getting from others features they also have relationship with each other but it is very less.









* On the bases of above graphically analyze the age between 30 to 43 have highest Fraud Reported. It means, mostly person are major and probably they have much experience of driving so, we can’t classified on the behalf of age whether they are Fraud or not.
* On the bases of above graphically analyze the Policy Annual Premium is equal in both condition.
* If we talk about Capital Gain and Capital Loss it is quite simple, if fraud is less defiantly capital gain will be increase and fraud reported is more that time capital loss well increase.
* If we talk about Incident hours of the day. The maximum cases happed between 10 to 18 hours.
* On the bases of above graphically evaluation the number of vehicles involved during incident is 1.
* In case of Property Damage, we have three observations Yes, No and third one is unknown. Unknown is the highest fraud reported.
* Bodily injury (and bodily injury claims) is a term describing physical harm to others caused by you, or anyone else covered on your insurance policy. On the graphical evaluation in this project here we are classifying into 3 (0,1,2) section and as we are getting 2 is the highest one.
* [Total Claim Amount](https://www.lawinsider.com/dictionary/total-claim-amount) means all amounts expended by Licensor and/or Licensee in connection with any third party claim of infringement or misappropriation in a Defense Action, including, but not limited to, attorney fees and legal costs, and/or a royalty or other amount that must be paid to a third party as a result of a final claim or judgment or settlement. In this case maximum amount claimed by fraud.
* According to dataset we have injury claim, property claim, vehicle claim and as we are getting in all three fraud is the maximum.
* According to insured occupation we have found highest fraud\_reported belongs to Executive Managerial.
* As we are getting maximum the case are Single Vehicle Collision and maximum collision is Rear collision. Accordingly Major Damage is higher than other damages.
* If we talk about auto\_make the maximum incident happed with Mercedes and lowest one is Jeep.

**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.

No null values are present in our dataset but we have Policy\_bind\_date features and as we know for better prediction we have to use Ordinal Encoding to extract information from its.

As we already said in the starting that we have many categorical features and we are using Label Encoder here to encoding them.

As we are getting the outliers are present in our data that’s why we are using Z-Score function to remove them. If we want to achieve good accuracy so 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 Fraud\_reported is our target variable.

Using MinMaxScaler to scaling the data into -1 to +1. Some time our features values is not equally scaled (same scale) so that time we have to scale all features (except target variable) on the same scale. According to our dataset some features are high in scale.

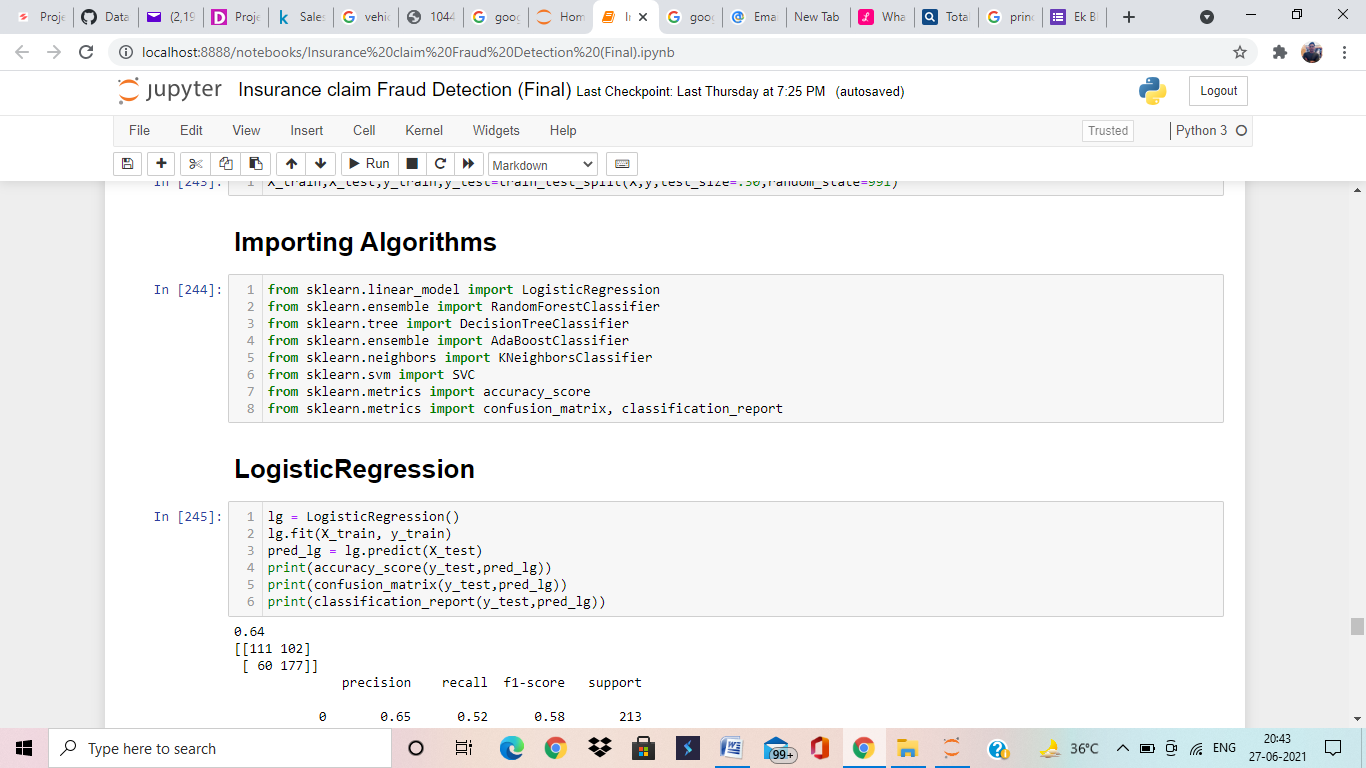
PCA (Principal Component Analysis) is a dimensionality reduction method. We have 40 features and we are trying to reduce dimensionality of these features. After applying PCA know we have 30 features but we are using only approx.4% of data.

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.

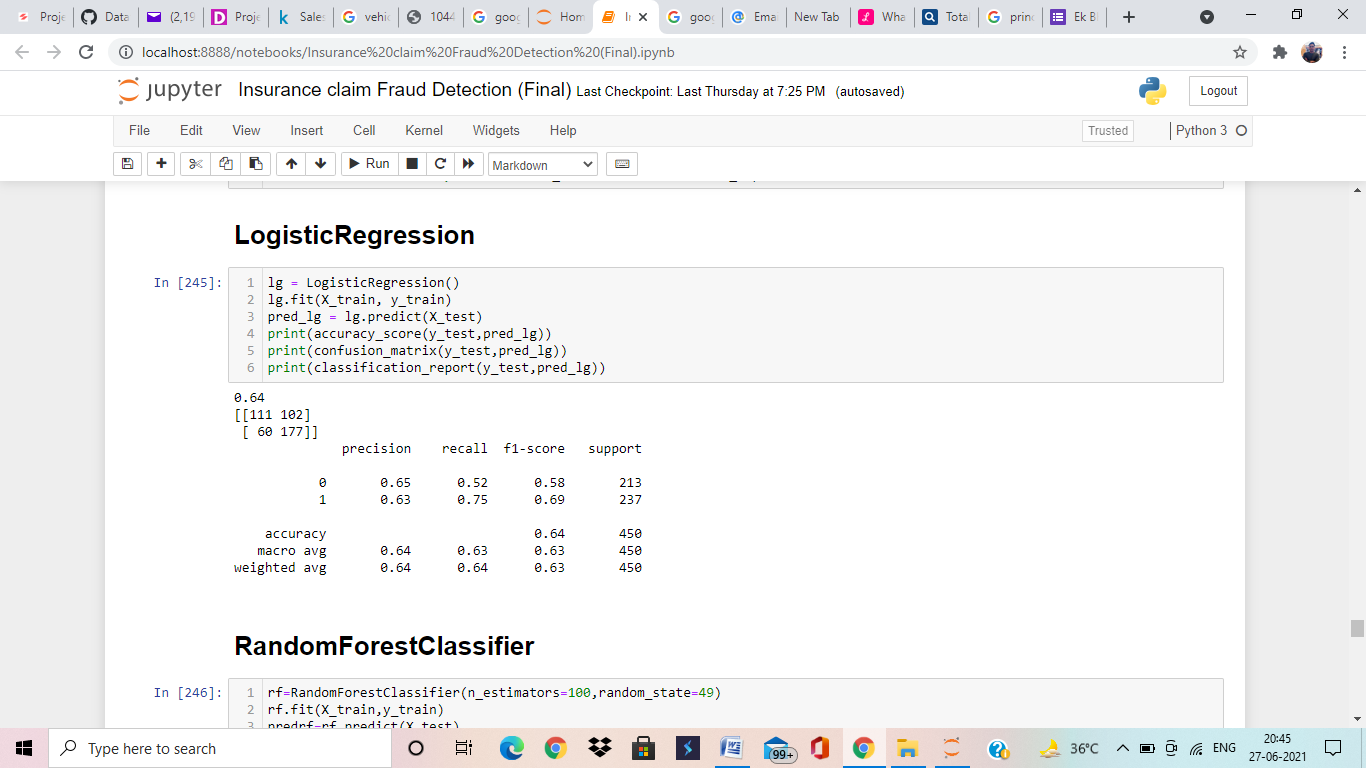
After data cleaning, features engineering, encoding and data scaling, know we are entering in Model building phase. We have already done different-different techniques on data.

This is Classification problem, so we are applying here multiple Logistic Regression algorithms and Classification algorithms.

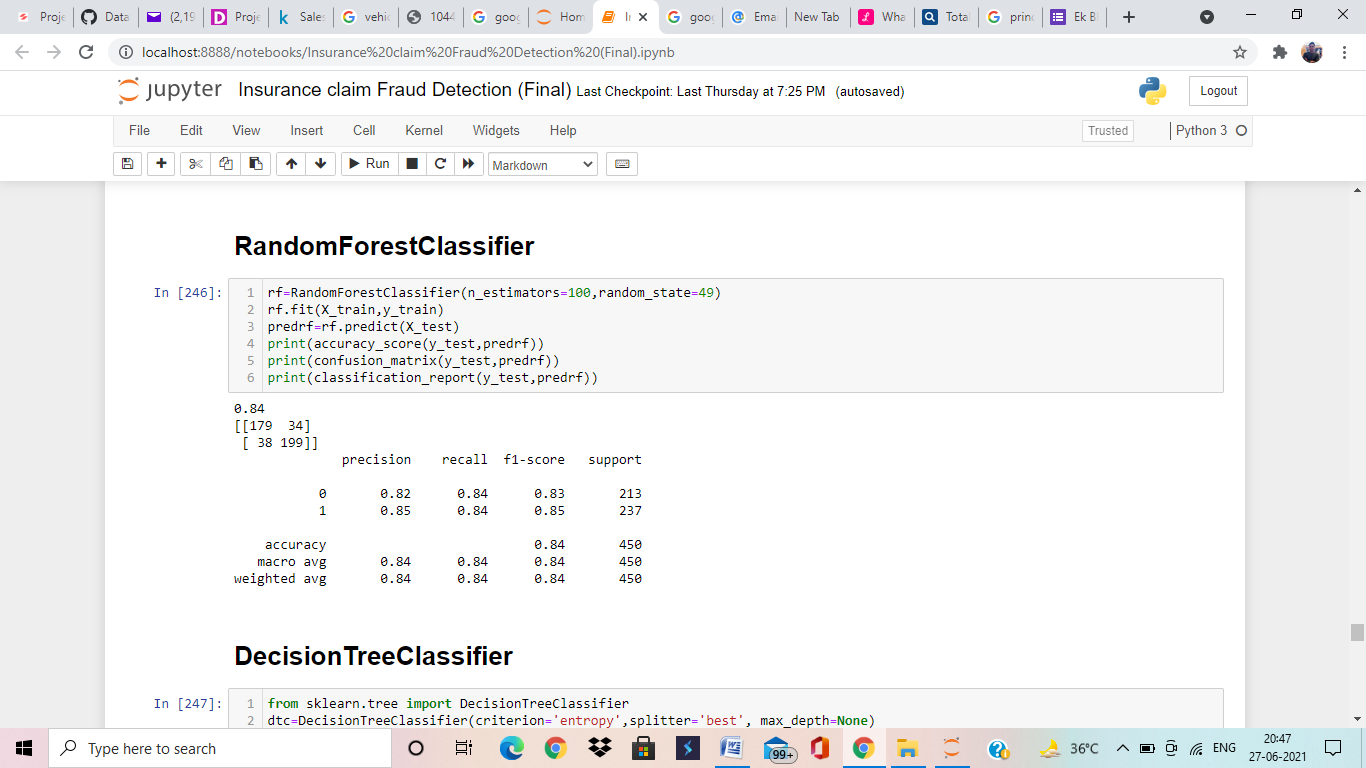
Take a look what algorithms we are applying.



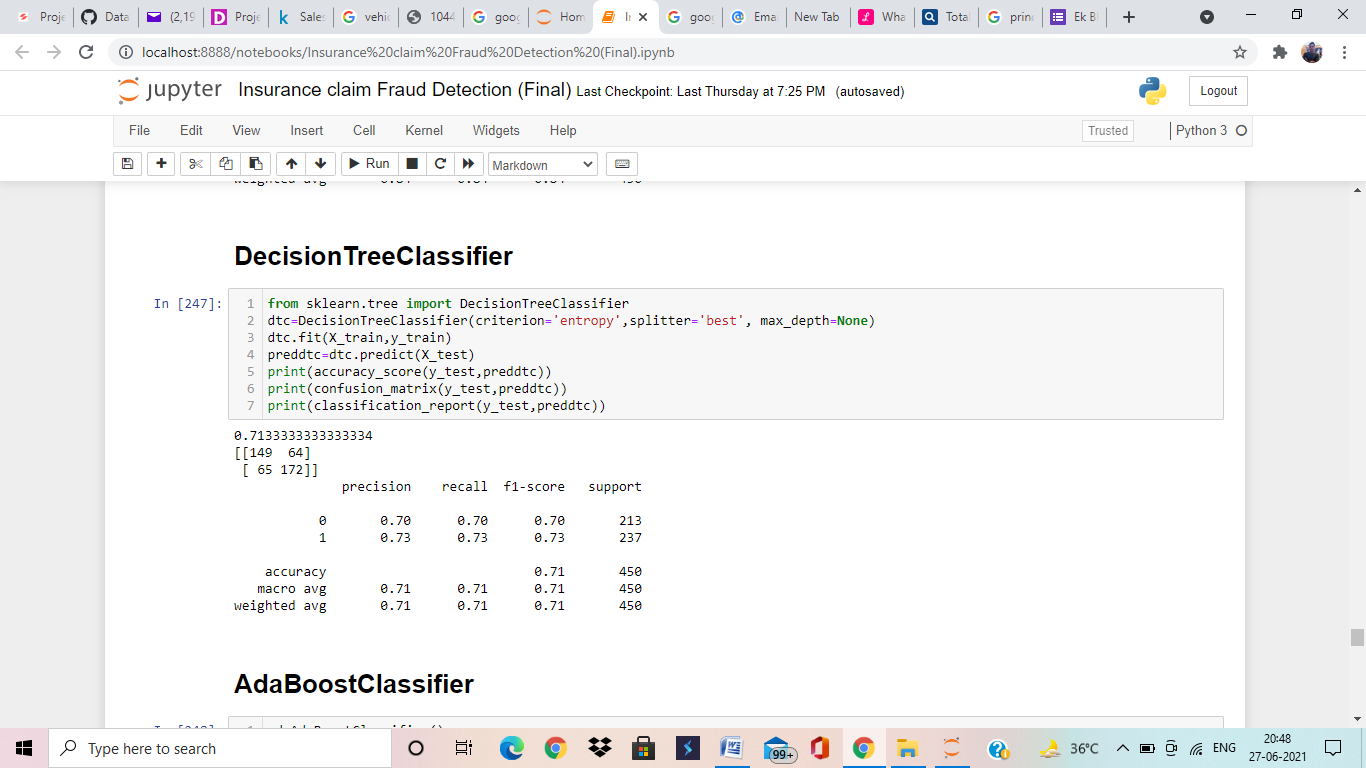
Know we will try to train our model with different-different algorithms for better accuracy.



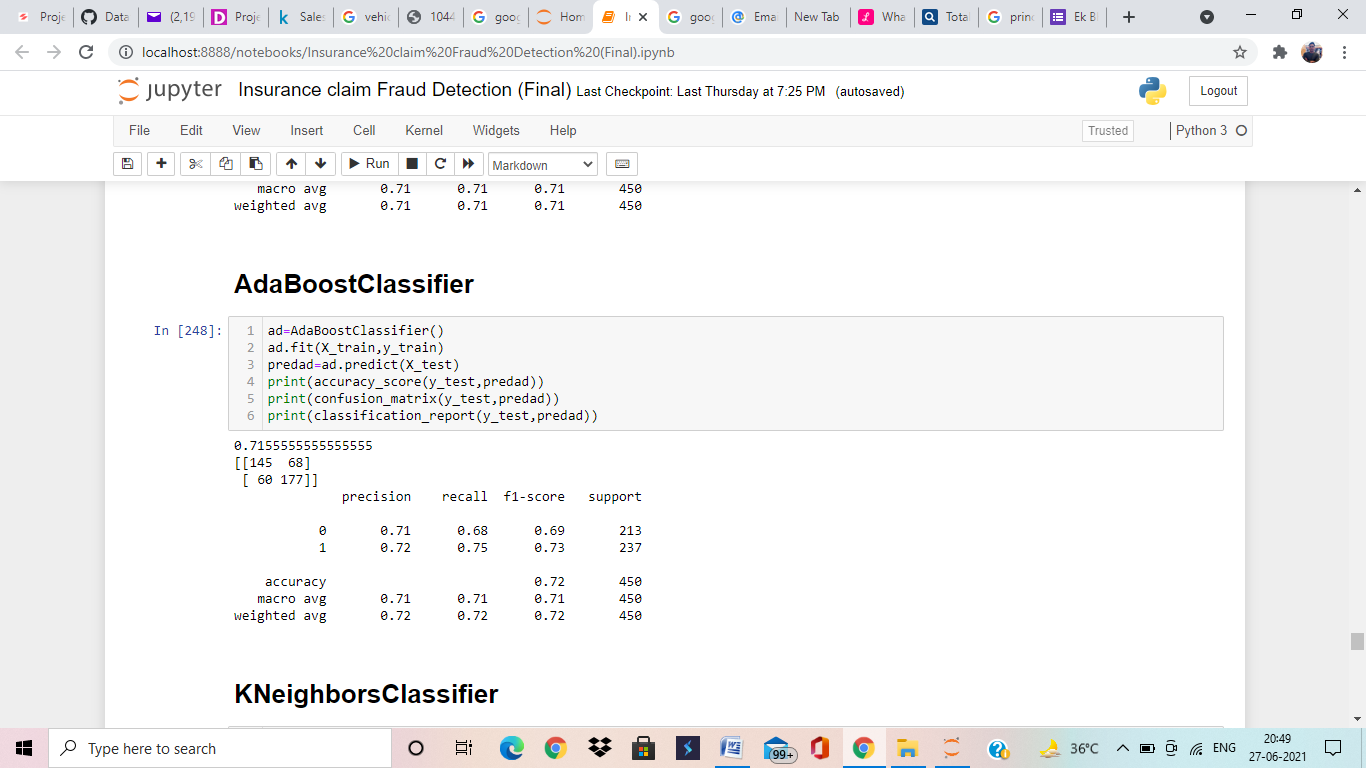
Here, we are applying Logistic Regression algorithms and Logistic Regression is giving us 64% accuracy, which is lower but good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



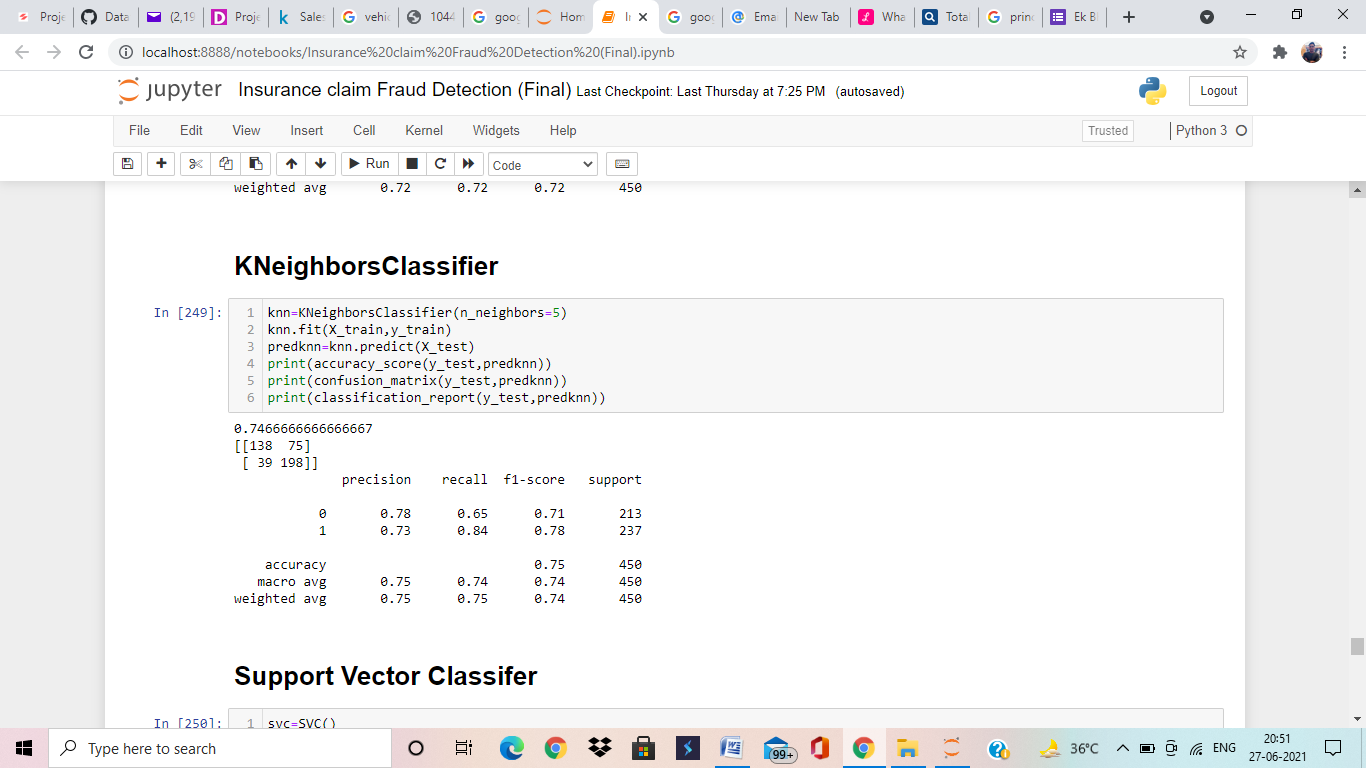
Here, we are applying Random Forest Classifier algorithms and Random Forest Classifier is giving us 84% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



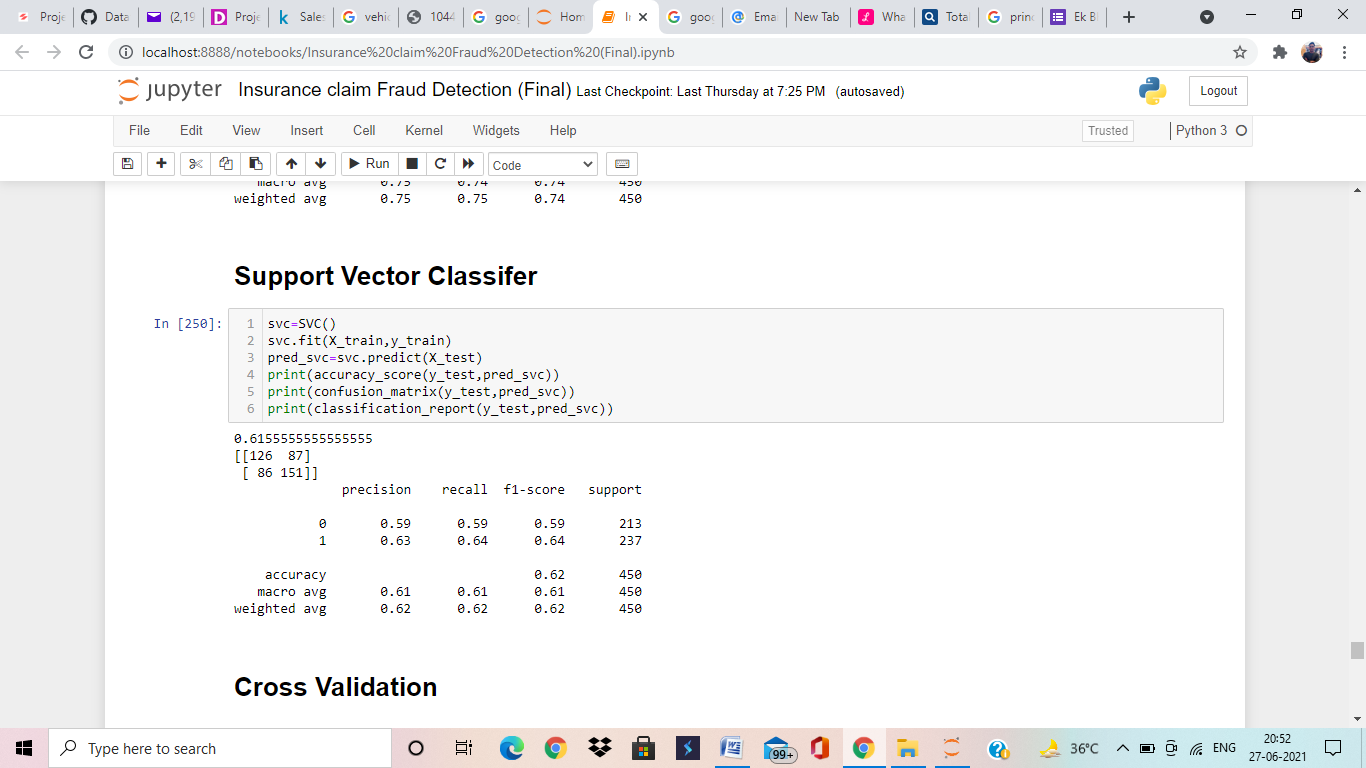
Here, we are applying Decision Tree Classifier algorithms and Decision Tree Classifier is giving us 71% accuracy, which is good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.



Here, we are applying AdaBoost Classifier algorithms and AdaBoost Classifier is giving us 71% 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.

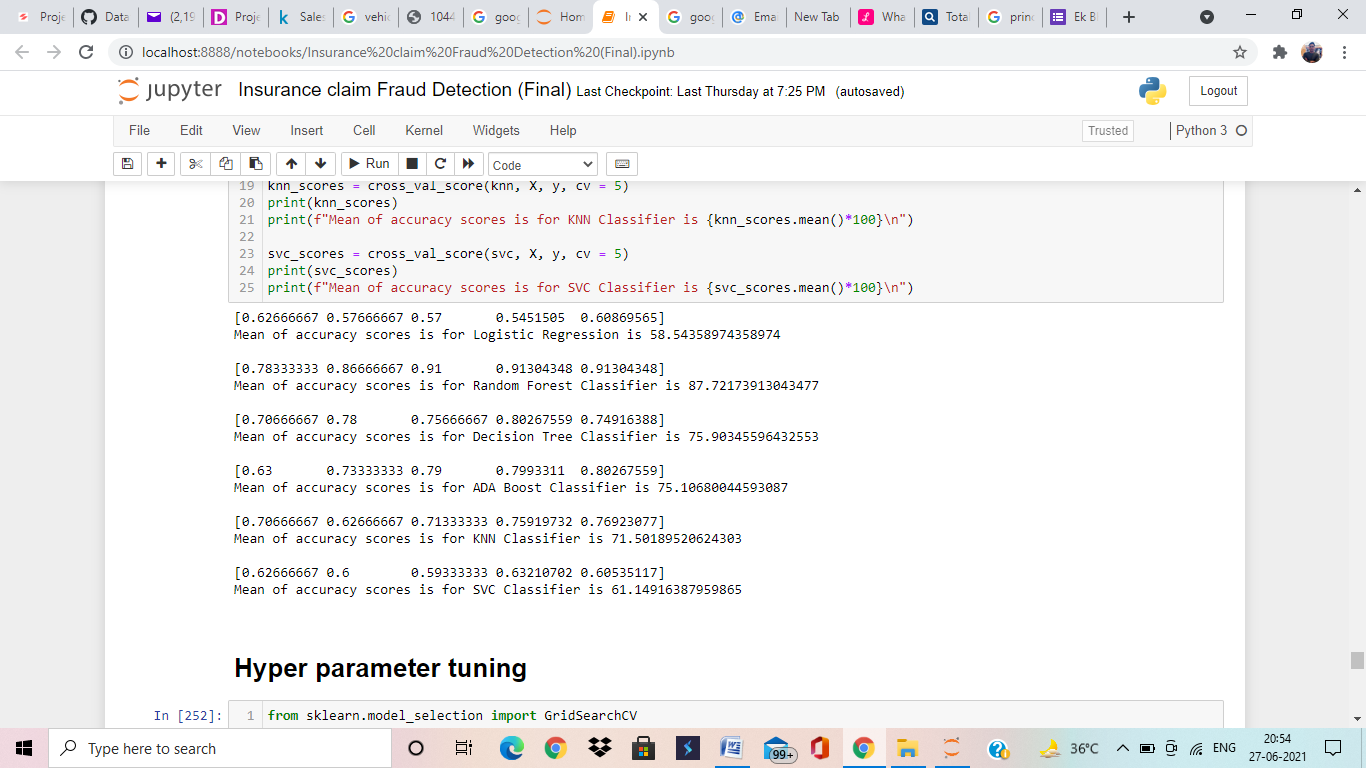


Here, we are applying KNeighbors Classifier algorithms and KNeighbors Classifier is giving us 74% accuracy, which is very good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.

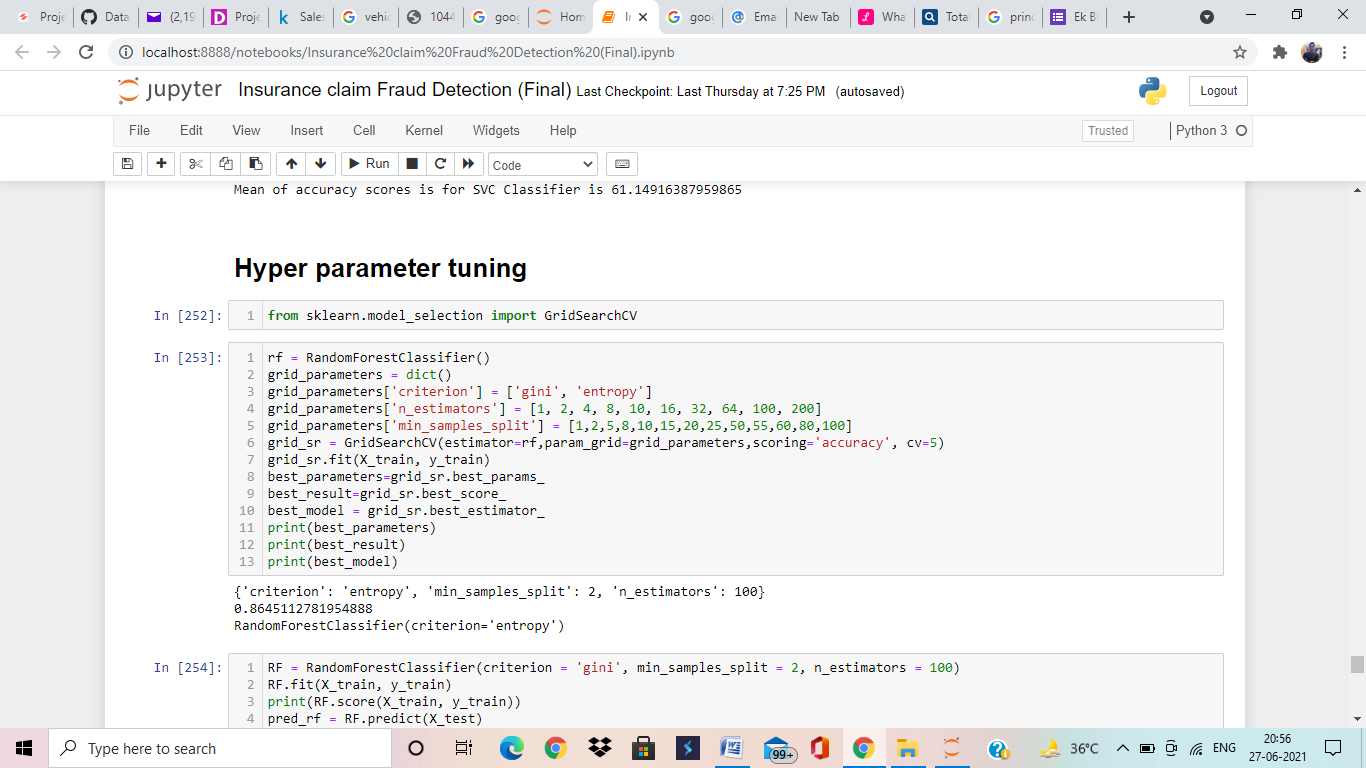


Here, we are applying Support Vector Classifier algorithms and Support Vector Classifier is giving us 61% accuracy, which is lower but good accuracy but there is may be over fitting/under fitting. We will deal with over fitting/under fitting during cross validation.

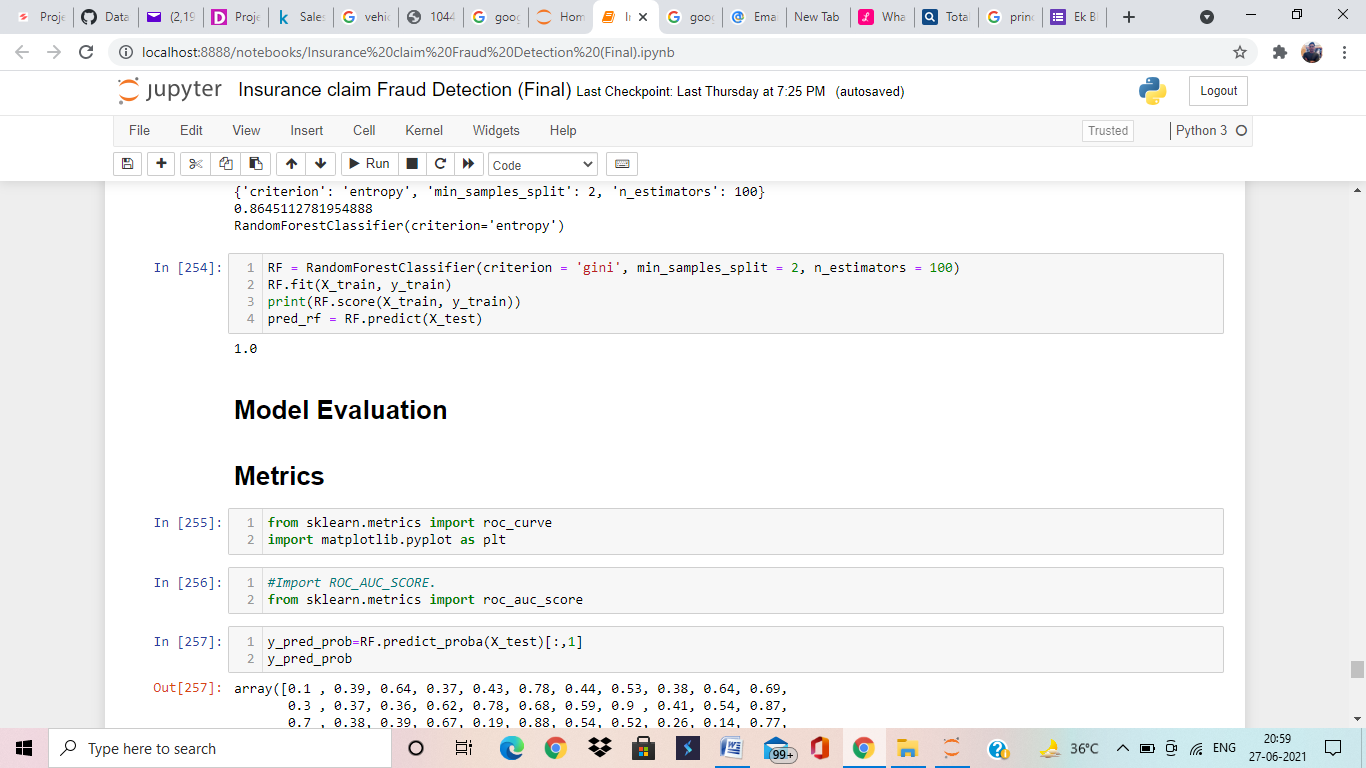
**Cross Validation**



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 a proper manner. All algorithms performing well except Logistic Regression. It is giving low score.

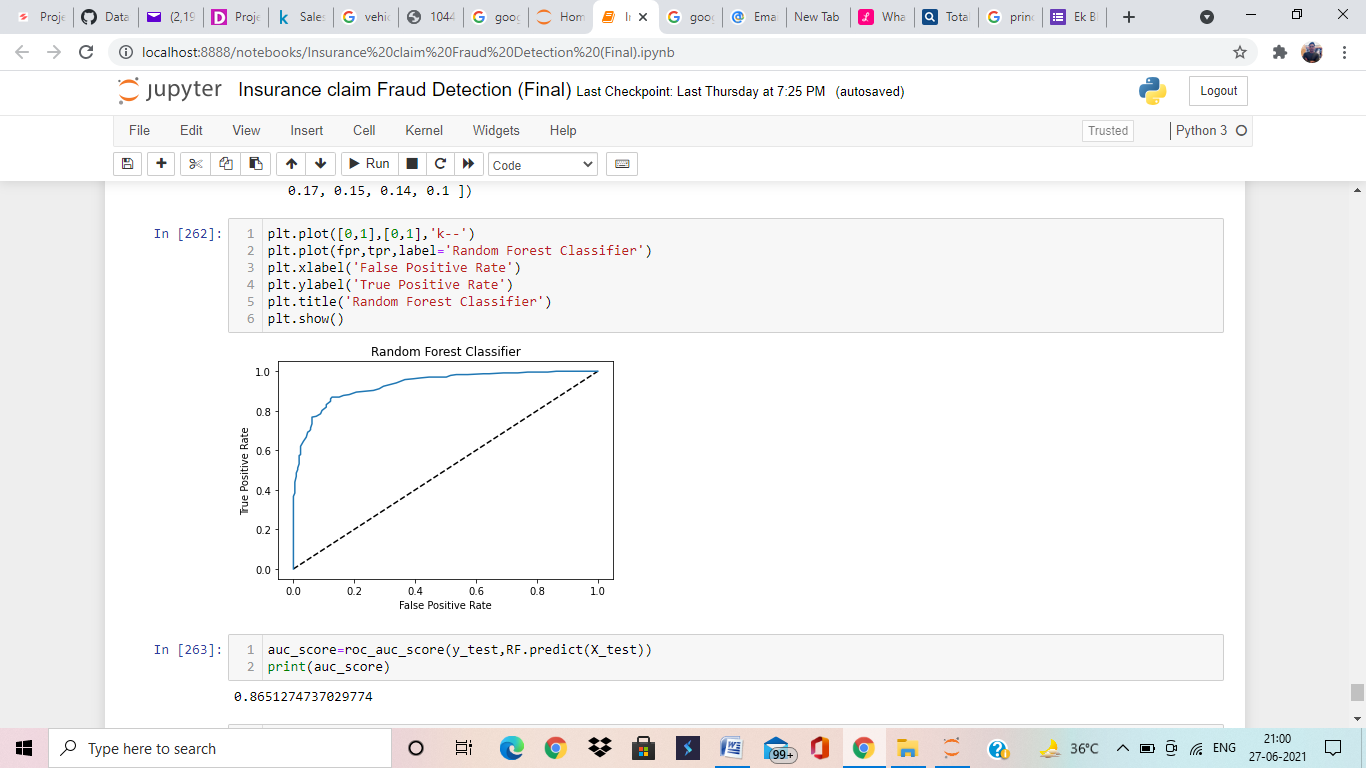
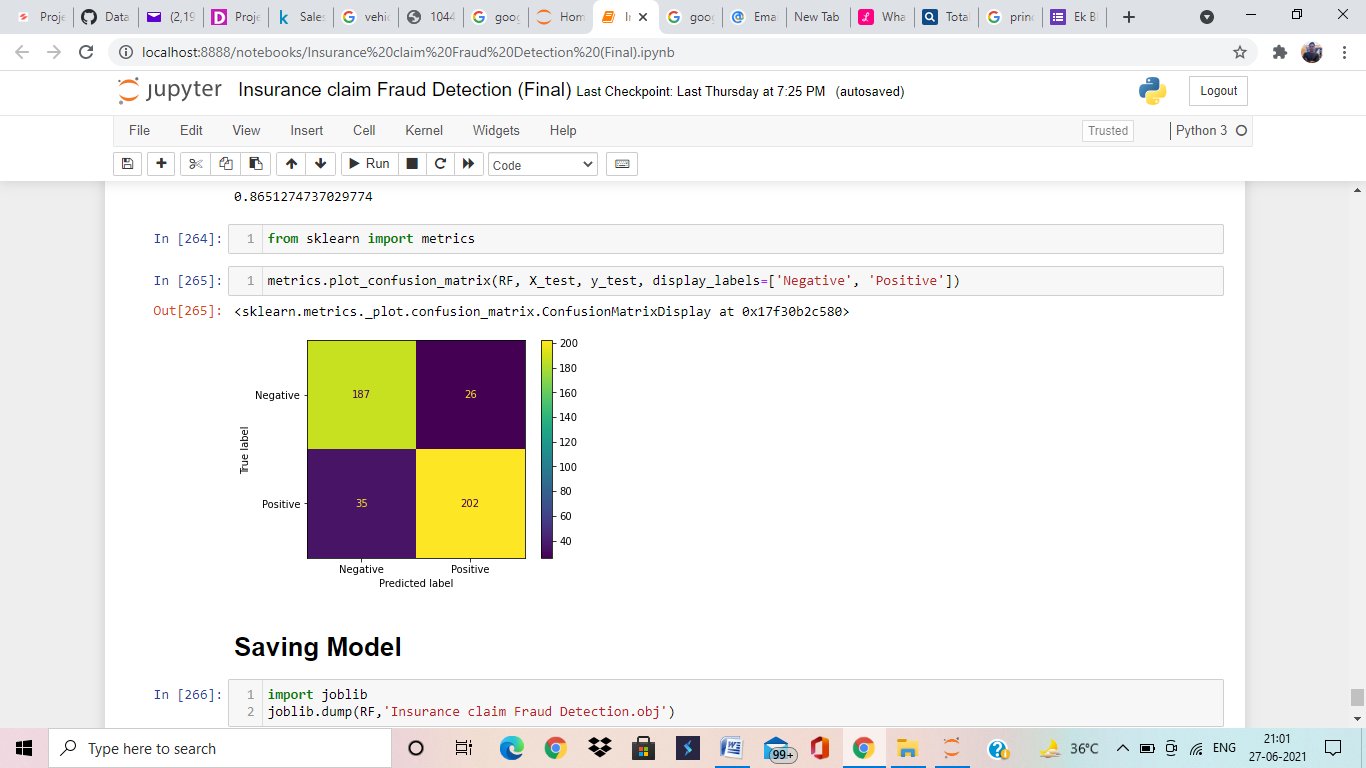


We are applying Hyper Parameter Tuning to find best parameters to get good accuracy and train model with these parameters. According to Hyper Parameter Tuning we find Random Forest Classifier parameter(Criterion: ‘entropy’, min\_samples\_split:2,n\_estimators:100) giving us best accuracy. Know we are building model with these parameters.



Here, we are building 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.86%. 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.

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