Fraud Detection System Using Machine Learning

**Hanzla Jaffar**

**Department Of Software Engineering, University Of Management and Technology**

**CS453: Dig Data Programming**

**Dr. Mazher Jawed Awan**

**July 18, 2022**

Big Data Programming

July 2022

Fraud detection System Machine Learning

**Abstract**: Now days, when there is huge amount of database and this data is handle by machine learning. To make secure people information and their privacy we are making fraud detection system, it will assure savage of information.

In order to visualize scalability and performance of model we make training data that for model to teach and will compare with original data for better accuracy. Here 4 models we are use Logistic Regression, Random Forest Tree, Decision Tree, GaussianNB, through these models and with data training data set we will find accuracy by using machine learning.

Keywords: Spark ML, Big data, Python

1. **Introduction**

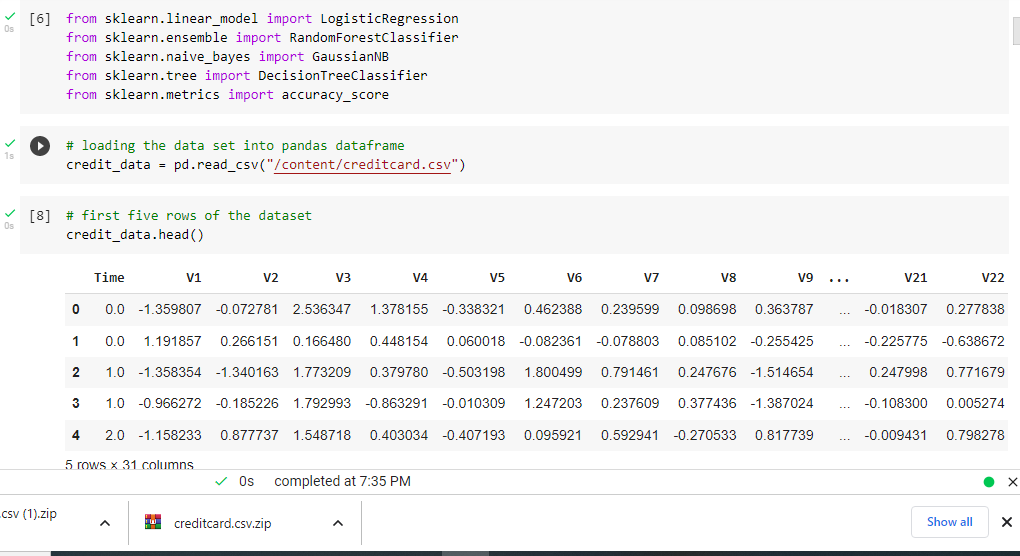
In this era, when everybody use social media and like to work in an easy way through online, people information store on database and this data manage through dig data programming, using multiple methods machine learning spark streaming etc. With manage of data there is need to secure data and to avoid any kind of online fraud. To avoid fraud we code by using different models with proper data set, including training set and testing test that we will make in code to make training data with actual data for good result.

1. **Import Dependencies**

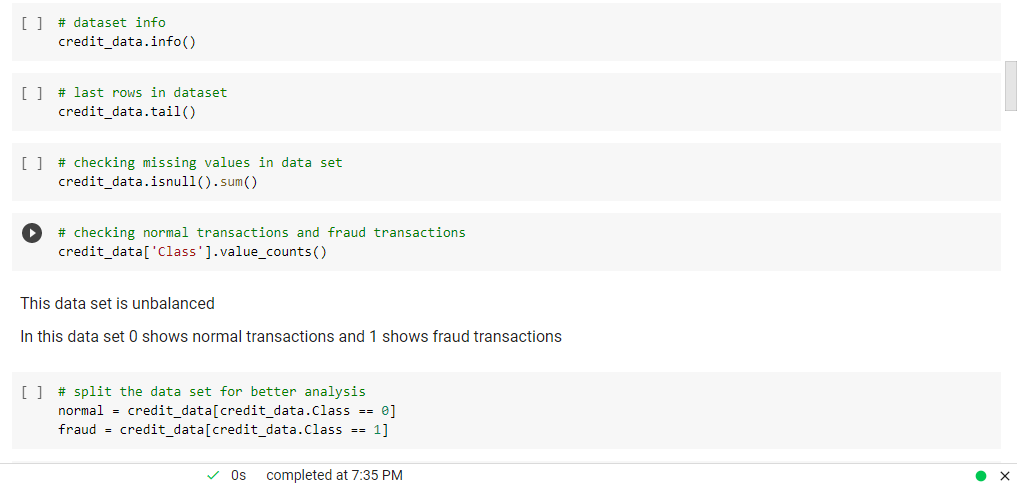
For better understanding, we add screenshots and their explanation



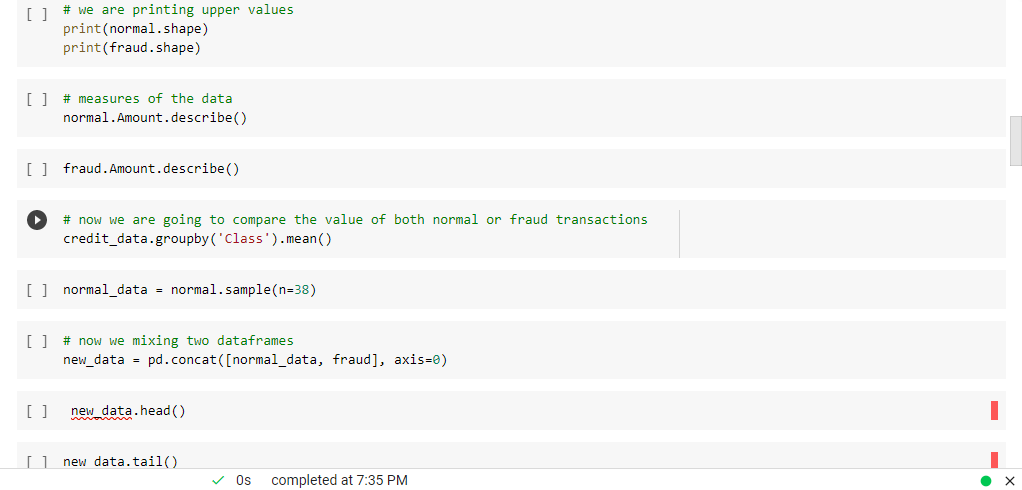
In this image import dependencies that we will use in code.



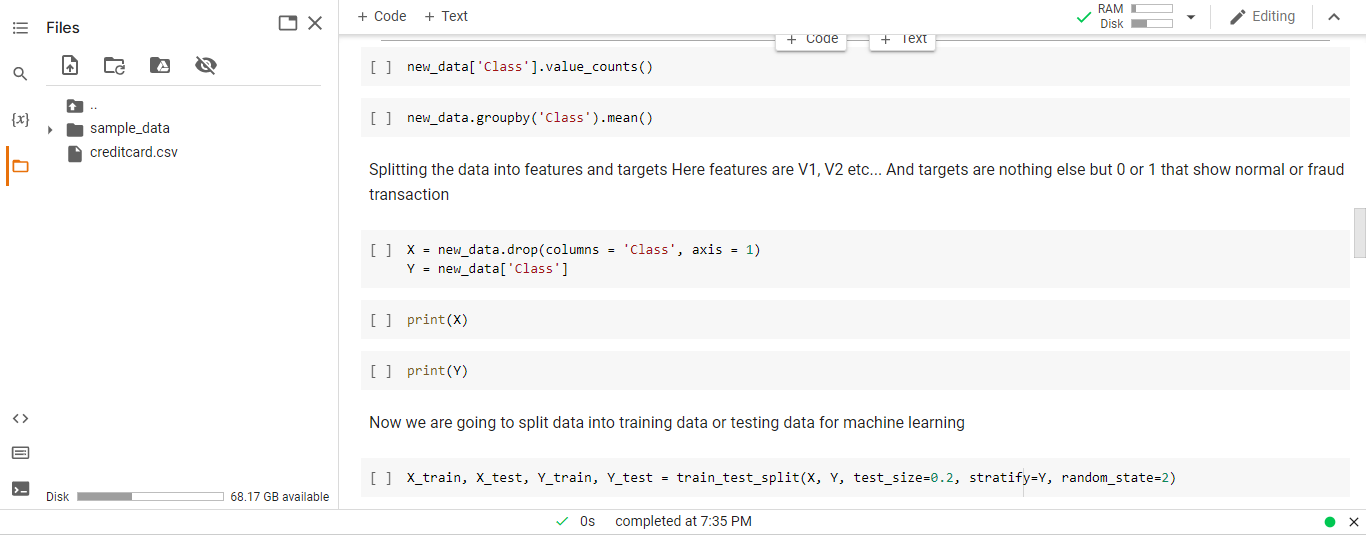
After import dependencies we make a variable name credit card and import data set by panda query after read dataset first five column has been printing using command.



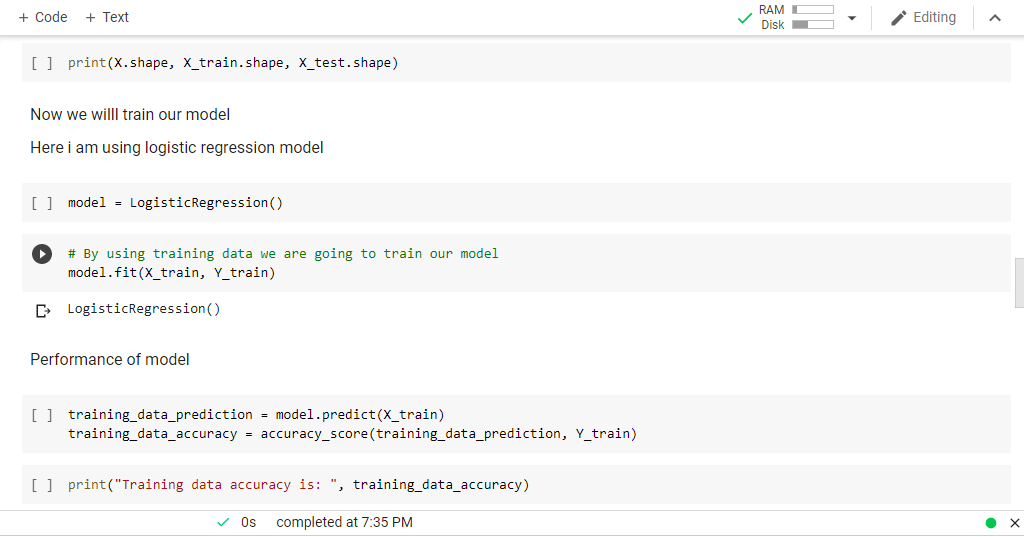
* In first line of screenshot I am taking info of data set
* IN second line showing last five line of data set
* In third line I am finding is there any null value present in dataset or not
* In 4th line taking class column of dataset and count values
* In 5th line making new Rdd name normal in which we store data of column name class which have zero value zero value means those transactions which are normal there is no fraud and in in fraud variable store that data in which fraud transactions are detect mention by 1



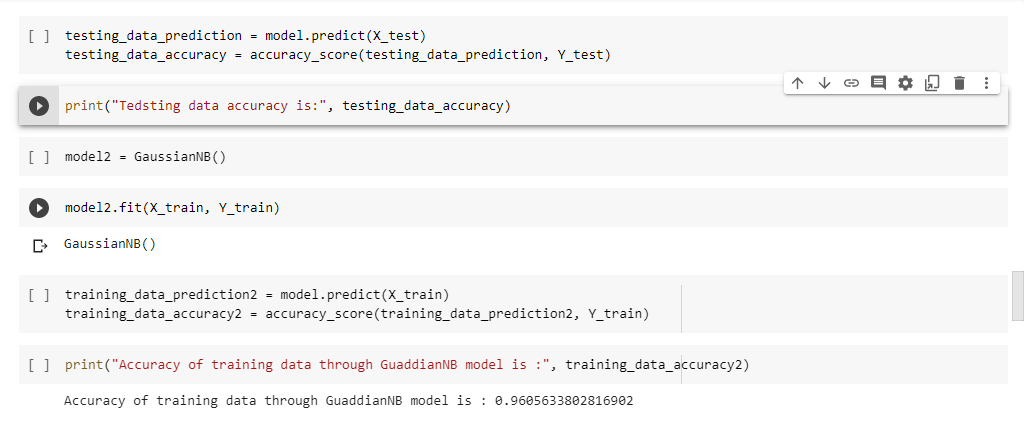
In this screenshot printing data store in rdd name normal and fraud that we make in last screenshot. Elaborating normal and fraud rdd amount describe in this system will show count, mean, maximum value, minimum value, percentage their datatypes. After describe both rdd’s, taking mean of column class from main dataset. Making new rdd name normal data and random 38 values from normal rdd. Now concatenate both rdd’s normal data and fraud to make balanced training dataset for ML both have same number of rows. Taking first and last five rows of new data rdd.



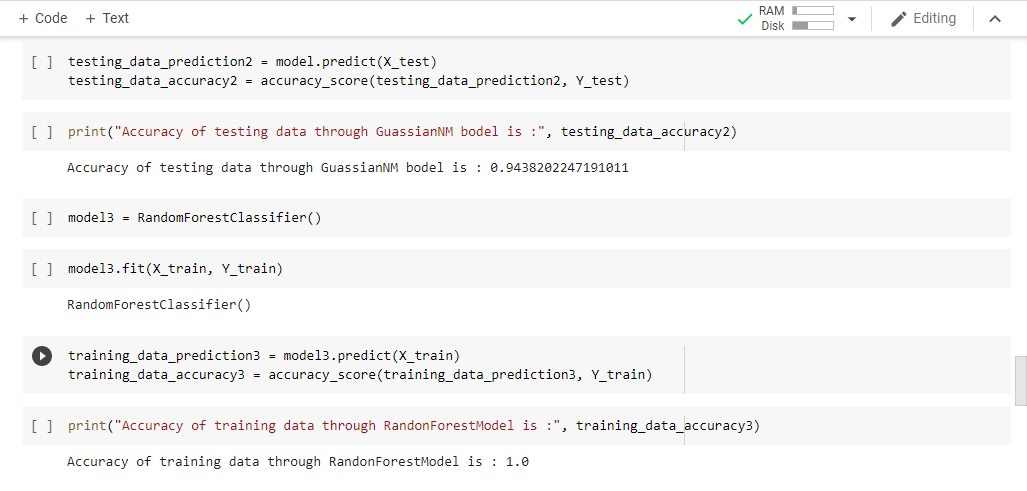
In upper screenshot make rdd name new data. Take class column of rdd new data and count values. After count taking mean of class column. Now make new rdd name X and Y in X rdd drop column class in column sequence and in Y rdd jut add column name class from new data rdd. Now print both rdd X and Y. Now make four variable name X\_test, X\_test, X\_train, Y\_train then apply train\_test\_split function in which give parameters X and Y as training data in test\_size variable there will be 20 percent data in test data and 80 percent in training data ,stratify variable mean this variable is independent, random\_state variacle show how data split random \_state = 2 it means it will arrange data like I arrange.



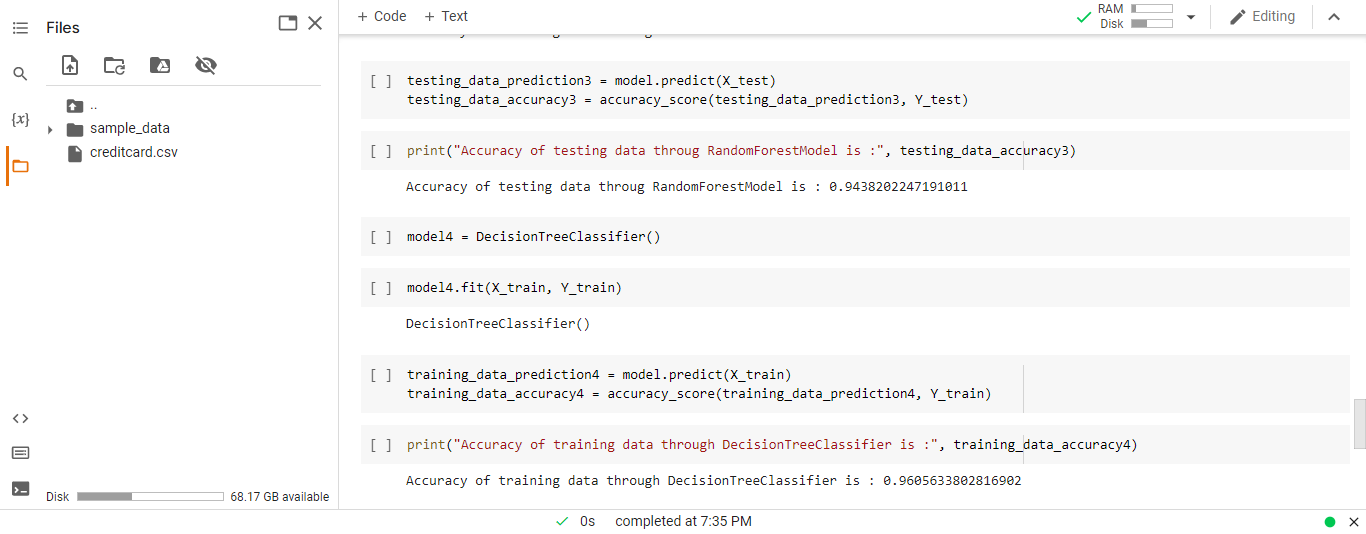
Print X.shape mean it will show whole dataset also X\_train that include 80 percent dataset and in X\_test contain 20 percent dataset. Now create object of model name Logistic Regression. Now check performance of model how accurate is this. By giving training data in new make variable first we check accuracy or training data and then testing data.



Using second model Guassian model and repeat all processes done in upper model.



To check accuracy of training data and testing data use Random Forest Classifier model. IN thi also repeat all processes done in before models. And print accuracy of training data and testing data.



Here we use another model Decision Tree by implement all old process and in last print.