## CREDIT CARD FRAUD DETECTION

***Project submitted in partial fulfillment of the requirements for the award of the degree of***

# BACHELOR OF TECHNOLOGY

### IN

**COMPUTER SCIENCE AND ENGINEERING**

**BY**

**G.RISHIKA REDDY (18C91A0524)**

**D.VEERASENA REDDY (18C91A0517)**

**BADSHAH ALAM (18C91A0505)**

**Under the Esteemed guidance of**

**MISS.MD.FOUZIYA M.Tech**

Assistant Professor



### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**HOLY MARY INSTITUTE OF TECHNOLOGY & SCIENCE**

**(COLLEGE OF ENGINEERING)**

***(Approved by AICTE New Delhi, Permanently Affiliated to JNTU Hyderabad, Accredited by NAAC with ‘A’ Grade)***

**Bogaram (V), Keesara (M), Medchal District -501 301.**

### 2021 - 2022

**HOLY MARY INSTITUTE OF TECHNOLOGY & SCIENCE**

#### (COLLEGE OF ENGINEERING)

***(Approved by AICTE New Delhi, Permanently Affiliated to JNTU Hyderabad, Accredited by NAAC with ‘A’ Grade)***

**Bogaram (V), Keesara (M), Medchal Dist-501301.**

### DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



**CERTIFICATE**

This is to certify that the mini project entitled “CREDIT CARD FRAUD DETECTION” is being submitted by G.RISHIKA REDDY (18C91A0524), D.VEERASENA REDDY (18C91A0517), BADSHAH ALAM(18C91A0505) in

Partial fulfillment of the academic requirements for the award of the degree of Bachelor of

Technology in “COMPUTER SCIENCE AND ENGINEERING” HOLY MARY INSTITUTE OF

TECHNOLOGY & SCIENCE, JNTU Hyderabad during the year 2021- 2022.

**INTERNAL GUIDE HEAD OF THE DEPARTMENT**

Miss.MD.FOUZIYA M.Tech DR .B.NARSIMHA M.Tech, Ph.D.

Assistant Professor Professor & HoD

Dept. of Computer Science & Engineering. Dept. of Computer Science & Engineering

### EXTERNAL EXAMINER

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, who’s constant guidance and encouragement crowns all effort with success.

I take this opportunity to express my profound gratitude and deep regards to My Guide **Miss.MD.Fouziya, Assistant Professor**, Dept. of Computer Science & Engineering, Holy Mary Institute of Technology & Science for his / her exemplary guidance, monitoring and constant encouragement throughout the project work.

My special thanks to **Dr. B. Narsimha, Head of the Department**, Dept. of Computer Science & Engineering, Holy Mary Institute of Technology & Science who has given an immense support throughout the course of the project.

I also thank to **Dr. P. Bhaskara Reddy,** the **honorable Director** of my college Holy Mary Institute of Technology & Science for providing me the opportunity to carry out this work.

At the outset, I express my deep sense of gratitude to the beloved **Chairman A. Siddartha Reddy** of **Holy Mary Institute of Technology & Science**, for giving me the opportunity to complete my course of work

I am obliged to **staff members** of Holy Mary Institute of Technology & Science for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

Last but not the least I thank **ALMIGHTY** and My **Parents**, and **Friends** for their constant encouragement without which this assignment would not be possible.

**G RISHIKA REDDY (18C91A0524)**

**D VEERASENA REDDY (18C91A0517)**

**BADSHAH ALAM (18C91A0505)**

**DECLARATION**

This is to certify that the work reported in the present project titled **“CREDIT CARD FRAUD DETECTION”** is a record of

work done by me in the Department of Computer Science & Engineering, Holy Mary Institute of Technology and Science.

No part of the thesis is copied from books/journals/internet and wherever the portion is taken, the same has been duly referred in the text the reported are based on the project work done entirely by me not copied from any other source.

**G RISHIKA REDDY (18C91A0524)**

**D VEERASENA REDDY (18C91A0517)**

**BADSHAH ALAM (18C91A0505)**

CONTENTS

ABSTRACT

Chapter Name of the Chapter Page No.

### INTRODUCTION 1

* 1. Problem Statement 2
  2. Objectives 3
  3. Motivation 3
  4. Existing System 3
  5. Proposed System 4

### LITERATURE SURVEY 5

1. **SOFTWARE REQUIREMENTS SPECIFICATIONS**
   1. Software Requirements 8

Hardware Requirements 8

### SYSTEM DESIGN

System Design 9

UML diagrams 11

### IMPLEMENTATION

* 1. Environmental Setup 17

Module Description 25

Sample Code 28

### SYSTEM TESTING

* 1. Tests 32

### RESULT SCREEN SHOTS 34

### CONCLUSION 37

### BIBLIOGRAPHY 38

LIST OF TABLES 39

LIST OF FIGURES 40

LIST OF SCREEN SHOTS 42

**ABSTRACT**

Now a day’s online transactions have become an important and necessary part of our lives. It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. As frequency of transactions is increasing, number of fraudulent transactions are also increasing rapidly. Such problems can be tackled with Machine Learning with its algorithms. This project intends to illustrate the modelling of a data set using machine learning with Credit Card Fraud Detection. Fraud detection has become an important tool and probably the best way to stop such frauds. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

### 1.INTRODUCTION

**CREDIT CARD:**

A credit card is a payment card issued to users (cardholders) to enable the cardholder to pay a merchant for goods and services based on the cardholder's promise to the card issuer to pay them for the amounts plus the other agreed charges.

The card issuer (usually a bank) creates a revolving account and grants a line of credit to the cardholder, from which the cardholder can borrow money for payment to a merchant or as a cash advance.

**FRAUD:**

Fraud is intentional deception to secure unfair or unlawful gain, or to deprive a

victim of a legal right. Fraud can violate civil law (i.e., a fraud victim may sue the fraud perpetrator to avoid the fraud or recover monetary compensation), a criminal law (i.e., a fraud perpetrator may be prosecuted and imprisoned by governmental authorities), or it may cause no loss of money, property or legal right but still be an element of another civil or criminal wrong. The purpose of fraud may be monetary gain or other benefits, for example by obtaining a passport, travel document, or driver's license, or mortgage fraud, where the perpetrator may attempt to qualify for a mortgage by way of false statements.

**CREDIT CARD FRAUD:**

Credit card fraud is an inclusive term for fraud committed using a payment card, such as a credit card or debit card.

The purpose may be to obtain goods or services, or to make payment to another account which is controlled by a criminal.

The Payment Card Industry Data Security Standard (PCI DSS) is the data security standard created to help businesses process card payments securely and reduce card fraud.

Credit card fraud can be authorised, where the genuine customer themselves processes a payment to another account which is controlled by a criminal, or unauthorised, where the account holder does not provide authorisation for the payment to proceed and the transaction is carried out by a third party.

1

**Machine learning**

Fraud Detection with Machine Learning becomes possible due to the ability of ML algorithms to learn from historical fraud patterns and recognize them in future transactions. Machine Learning algorithms appear more effective than humans when it comes to the speed of information processing. Also, ML algorithms are able to find sophisticated fraud traits that a human simply cannot detect. Machine Learning Fraud Detection Models

**Email Phishing Detection Models:**

Phishing emails represent spam letters that have fraudulent intentions. Phishers make fake websites and their URLs highly similar, both visually and semantically, to the originals. They are mostly threats to the Banking sector, multinational companies, and even medical establishments.

Logistic Regression is one of classic Machine Learning algorithms for phishing detection. Logistic Regression uses a linear model to predict a number in range from “0” or “1”, meaning fraud or not.

**Identity Theft Detection Models**

To prevent identity theft, a method such as patterns identification can significantly improve the accuracy of fraud detection. For example, if an individual’s behaviour patterns are stored to a database. That way, the previous behaviour patterns recorded for a certain user are constantly being compared to the activity in the account. In the event that this activity largely differs from the norm, fraud can be suspected. Each new transaction contributes to the behavioural analytics process done by the model, helping it to train better.

**1.1 Problem statement:**

Credit card frauds are increasing heavily because of fraud financial loss is increasing drastically. Every year due to fraud Billions of amounts lost. To analyze the fraud there is lack of research. Many machine learning algorithms are implemented to detect real world credit card fraud.

2

**1.2 Objectives:**

The main objective of this thesis is to perform predictive analysis on credit card transaction dataset using machine learning techniques and detect the fraudulent transactions from the given dataset. The focus is to identify if a transaction comes under normal class or fraudulent class using predictive models. Different sampling techniques will be implemented to tackle the class imbalance problem and series of machine learning algorithms like logistic regression, random forest will be implemented on the dataset, and the results will be reported.

**1.3 Motivation:**

Now a day the customers prefer the most accepted mode via credit card for the convenient way of paying bills online shopping is easiest way. At the same time the fraud transaction risks using credit card is a main problem which should be avoided.so there are many techniques available to avoid these risks effectively in existing research they model the sequence of operations in credit card transaction processing using a random forest and logistic regression and shown how it can be used for the detection of frauds.to avoid computational complexity and to provide better accuracy in fraud detection in proposed system.

**1.4 Existing system:**

In existing system, the K-means clustering model produced a low accuracy. Using K-means there were quite a few non-fraudulent activities, which wrongly got detected as frauds. Therefore, K-means would not be the preferred model, as it doesn’t correctly predict frauds and it also produced a lot of false positives.  The Traditional detection method mainly depends on database system and the education of customers, which usually are delayed, inaccurate and not in-time. After that methods based on discreate analysis and mining algorithms are widely used which can detect fraud by credit rate for cardholders and credit card transaction. For a large amount of data it is not efficient.

3

**DISADAVANTAGES OF EXISTING SYSTEM:**

• Frauds were misclassified as non-frauds resulting in increased False Positive rate.

• Increased False Positive rate resulted in low accuracy.

• False Negative rate was low but it affects the accuracy.

**1.5 Proposed system:**

The proposed system is a machine learning application to detect frauds in credit card transactions using Random Forest algorithm. These algorithms are used to classify the credit card data set and then regression is performed.

The Random Forest Algorithm has been found to produce a good estimate of the generalization error and to be resistant to over fitting. This algorithm has been found to produce a good accuracy and precision. The proposed system overcomes the above mentioned issue in an efficient way. Using random forest and logistic algorithms the fraud is detected and the false alert is minimized and it produces an optimized result. Here the random forest and logistic algorithms are made where a set of interval valued parameters are optimized.

**ADVANTAGES OF PROPOSED SYSTEM:**

1.One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks.

2.Random forest algorithm builds multiple decision trees and merges them together to get a more accurate and stable prediction.

3.The random forest algorithm also works well when data possess missing values.

4

**2.LITERATURE SURVEY**

The fraud detection is a complex task and there is no system that correctly predicts any transaction as fraudulent. The properties for a good fraud detection system are:

1. Should identify the frauds accurately.

2. Should detect the frauds quickly.

3. Should not classify a genuine transaction as fraud.

Outlier detection is a critical task as outliers indicate abnormal running conditions from which significant performance degradation may happen. Techniques used in fraud detection can be divided into two:

1) Supervised techniques where past known legitimate/fraud cases are used to build a model which will produce a suspicion score for the new transactions

2) Unsupervised are those where there are no prior sets in which the state of the transactions are known to be fraud or legitimate.

**Supervised outlier detection technique:**

supervised outlier detection techniques assume the availability of a data set which has been needed for the normal as well as the outlier class. Supervised method detects fraudulent transactions that can be used to differentiate between those accounts or transactions which are known to be fraudulent and those which are known to be legitimate. Classification techniques such as statistical discriminate analysis and neural networks can be used to discriminate between fraudulent and non-fraudulent transactions to give transactions a suspicion score. Supervised methods are only trained to differentiate between legitimate transactions and previously known fraud.

**Decision Tree:**

Decision trees are statistical data mining technique that uses independent attributes and a dependent attributes which are logically AND in a tree shaped structure. The classification rules extracted from decision trees are IF-THEN expressions and all the tests have to succeed if each rule is to be generated. Decision tree usually separates the complex problem into many simple ones and resolves the sub problems through repeatedly using [11]. Decision trees are predictive decision support tools which create mapping from observations. Decision tree methods are C5.0, C&RT and CHAID. The data mining techniques including decision trees and SVMs to the credit card fraud detection problem is

useful in reducing the bank’s risk.

5

**Logistic Regression:**

The two data mining approaches, are support vector machines and random forests, together with the well known logistic regression, as part of an attempt to detect the credit card fraud. It is well-understood, easy to use, and it is most commonly used for data-mining. Thus it provides a useful baseline for comparing performance of newer methods. Supervised learning methods for fraud detection face two challenges. They are:

1. The unbalanced class sizes of legitimate and fraudulent transactions, with legitimate transactions far outnumbering fraudulent ones.

2. The second is to develop supervised models for fraud that can arise from potentially undetected fraud transactions, leading to mislabeled cases in the data to be used for building the model.

For the purpose of the above problems, the fraudulent transactions are those specifically identified by the institutional auditors as those that caused an unlawful transfer of funds from the bank sponsoring the credit cards. These transactions were observed to be fraudulent expose.

The study is based on real-life data of transactions from an international credit card operation.

S P Maniraj [1] In this paper, they describe Random forest algorithm applicable on Find fraud detection. Random forest has two types. They describe in detail and their accuracy 91.96% and 96.77%respectively. This paper summaries second type is better than the first type.

Suman Arora [2] In this paper, many supervised machine learning algorithms apply on 70% training and 30% testing dataset. Random forest, stacking classifier, XGB classifier, SVM, Decision tree and KNN algorithms compare each other i.e. 94.59%, 95.27%, 94.59%, 93.24%, 90.87%, 90.54% and 94.25% respectively. Summaries of this paper, SVM has the highest ranking with 0.5360 FPR, and stacking classifier has the lowest ranking with 0.0335.

Kosemani Temitayo Hafiz [3] In this paper, they describe flow chart of fraud detection process. i.e. data Acquisition, data pre-processing, Exploratory data analysis and methods or algorithms are in detail. Algorithms are K- nearest neighbor (KNN), random tree and Logistic regression accuracy are 96.91%, 94.32%, 57.73% and 98.24% respectivel

6

**Applications:**

* Banking Industry
* Credit Card Fraud Detection
* Customer Segmentation
* Predicting Loan Defaults on LendingClub.com
* Healthcare and Medicine
* Cardiovascular Disease Prediction
* Diabetes Prediction
* Breast Cancer Prediction
* Stock Market
* Stock Market Prediction
* Stock Market Sentiment Analysis
* Bitcoin Price Detection
* E-Commerce
* Product Recommendation
* Price Optimization
* Search Ranking

7

**3.SOFTWARE REQUIREMENTS SPECIFICATIONS**

**3.1 Software Requirements:**

Operating system : Windows 10

Programming Language : Python

IDE : jupyter/syder, Anaconda Python

**3.2 Hardware Requirements:**

System : Intel i5 core

Hard Disk : 100GB

Monitor : 15” LED

Input devices : Keyboard,Mouse

RAM : 8 GB

8

**4.SYSTEM DESIGN**

**4.1 System Design:**

The fraud detection module will work in the following steps:

1.The Incoming set of transactions and amount are treated as credit card transactions.

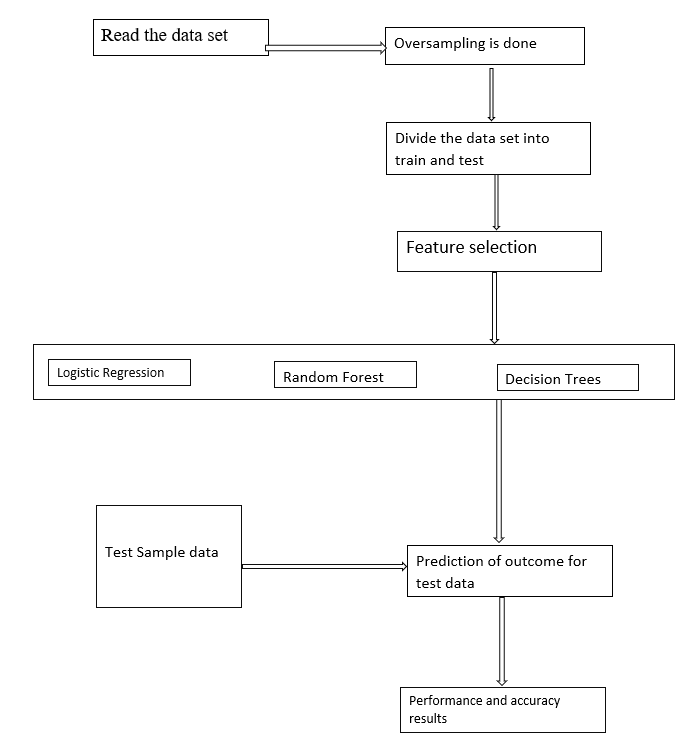
2.The credit card transactions are given to machine learning algorithms as an input.

3.The output will result in either fraud or valid transaction by analyzing the data and observing a pattern and using machine learning algorithms such as local outlier factor and isolation forest to do anomaly detection.

4.The fraud transactions are given to alarm which alerts the user that fraud transaction has occurred and the user can block the card to prevent further financial loss to him as well as the credit card company.

5.The valid transactions are treated as genuine transactions.

9



10

**4.2 UML diagrams:**

The activity of the uml diagram of credit card approval system which shows the flow between the activity of credit card, consumer, document, application, limits. The main activity involved in this uml activity diagram of credit card are as follows:

• Credit card activity

• Consumer activity

• Document activity

• Application activity

• Limits activity

Features of the activity UML diagram of credit card

• Admin user can search credit card, view description of a selected credit card, add credit card, update credit card and delete credit card.

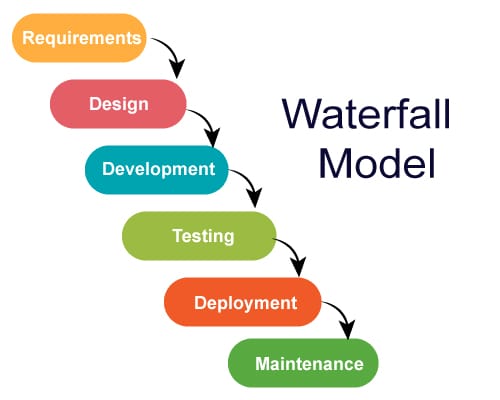
• It shows the activity flow of editing , adding and updating of consumer.

• User will be able to search and generate report of document, application, limits.

• All objects such as (credit card, consumer, limits) are interlinked.

• It shows the full description and flow of credit card, application, limits, document, consumer.

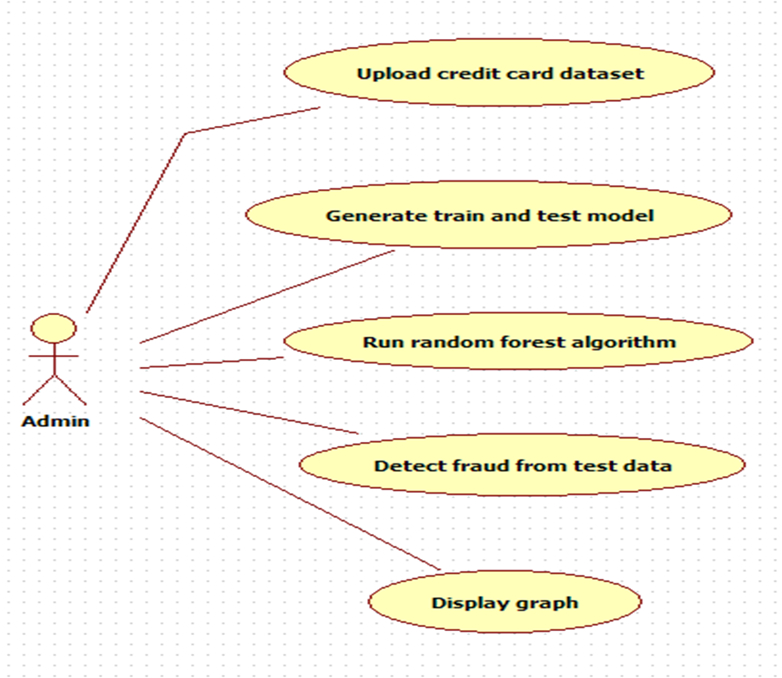
11



12

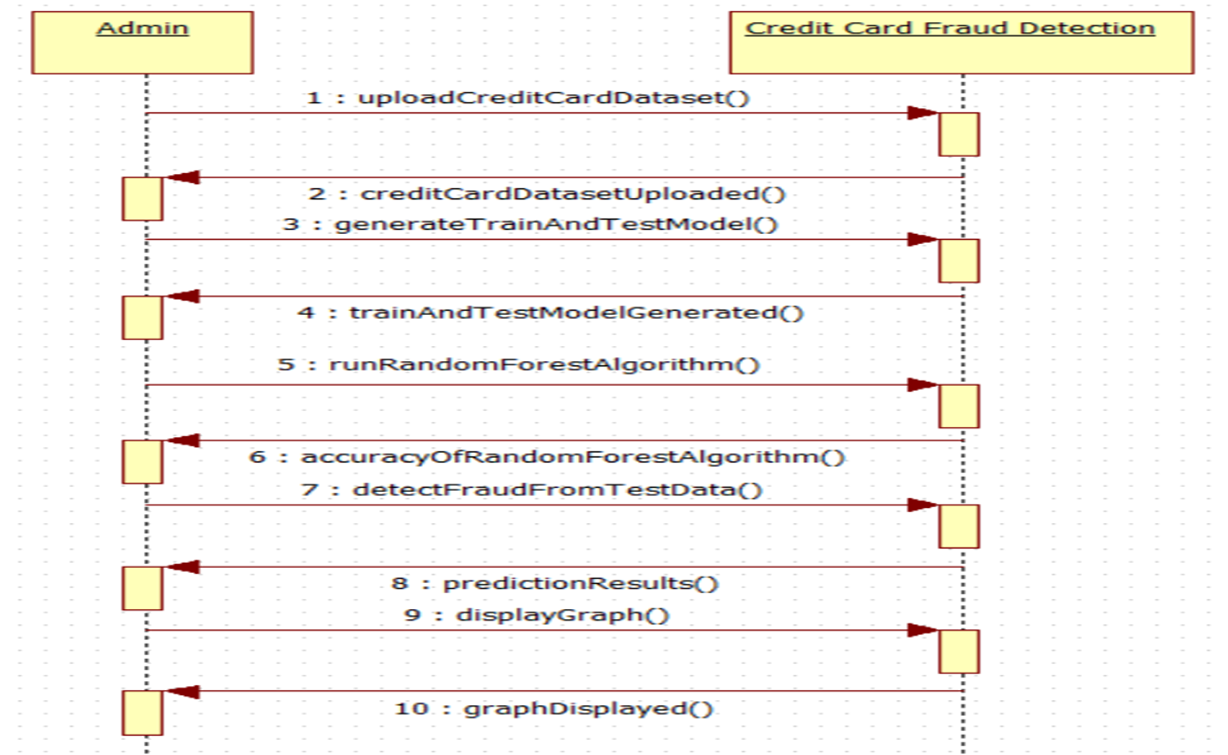
**Usecase diagram:**

A use case diagram is a type of behavioural diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals and any dependencies between those use cases. The below diagram figure shows the overall use case diagram for credit card fraud detection. A use case diagram is a type of behavioural diagram defined by the unified modeling language .The USE CASE diagram below describes the interaction between the customers and card issuers

****

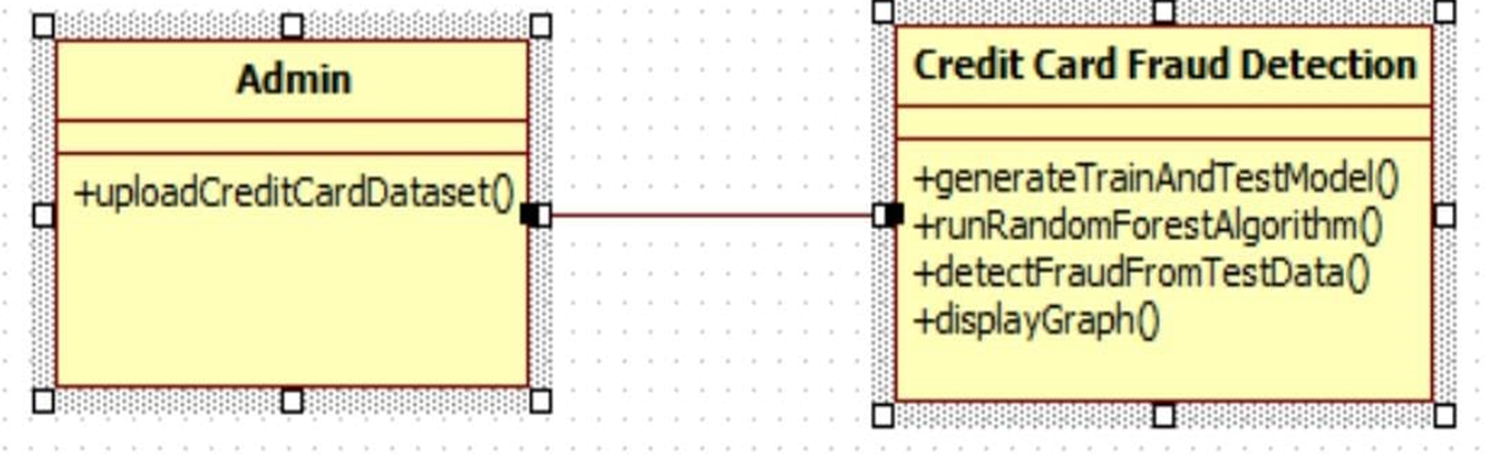
13

**Sequence Diagram:**

****

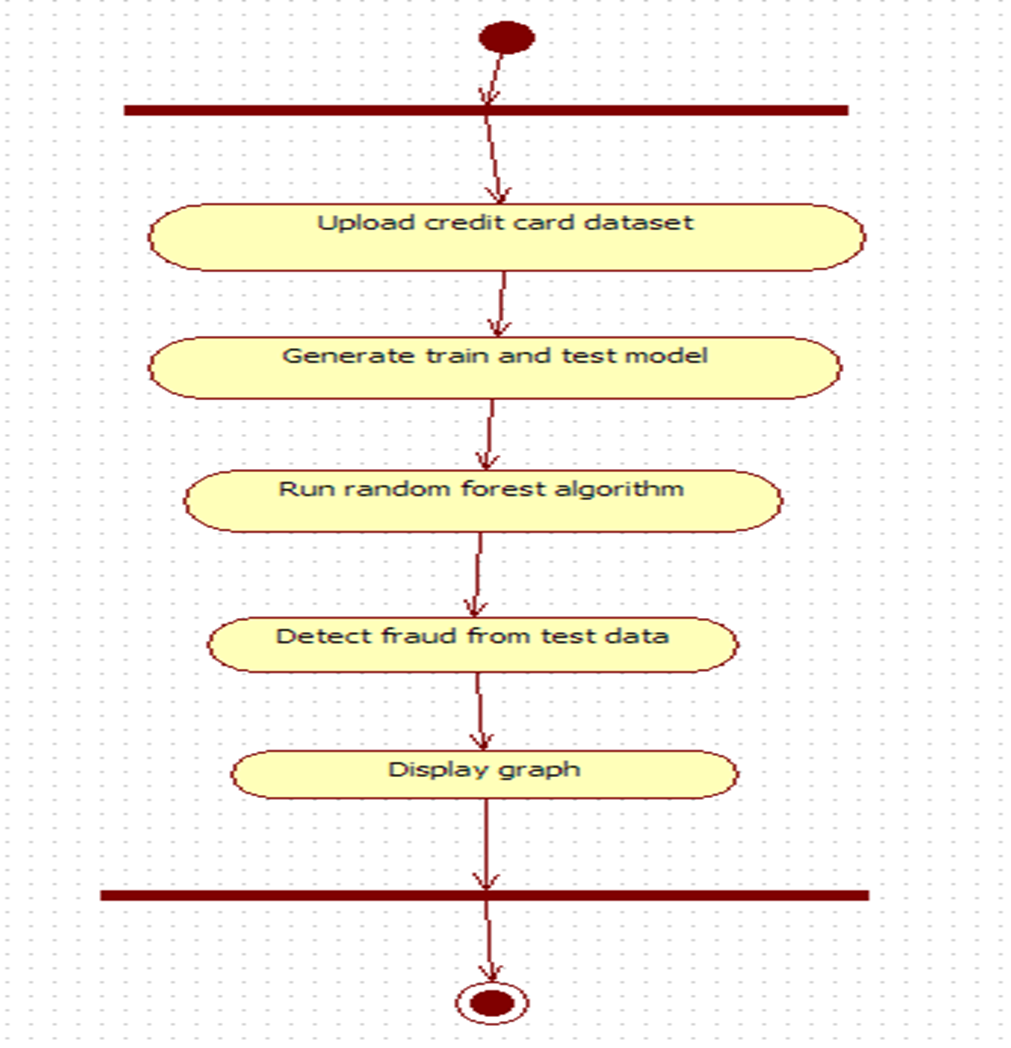
14

**Class diagram:**

****

15

**Activity diagram:**

****

16

**5.IMPLEMENTATION**

**5.1 Environmental Setup:**

• we need to install and setup the IDE

• after installing we need to set the path in environmental variables

• the process for installing is as below

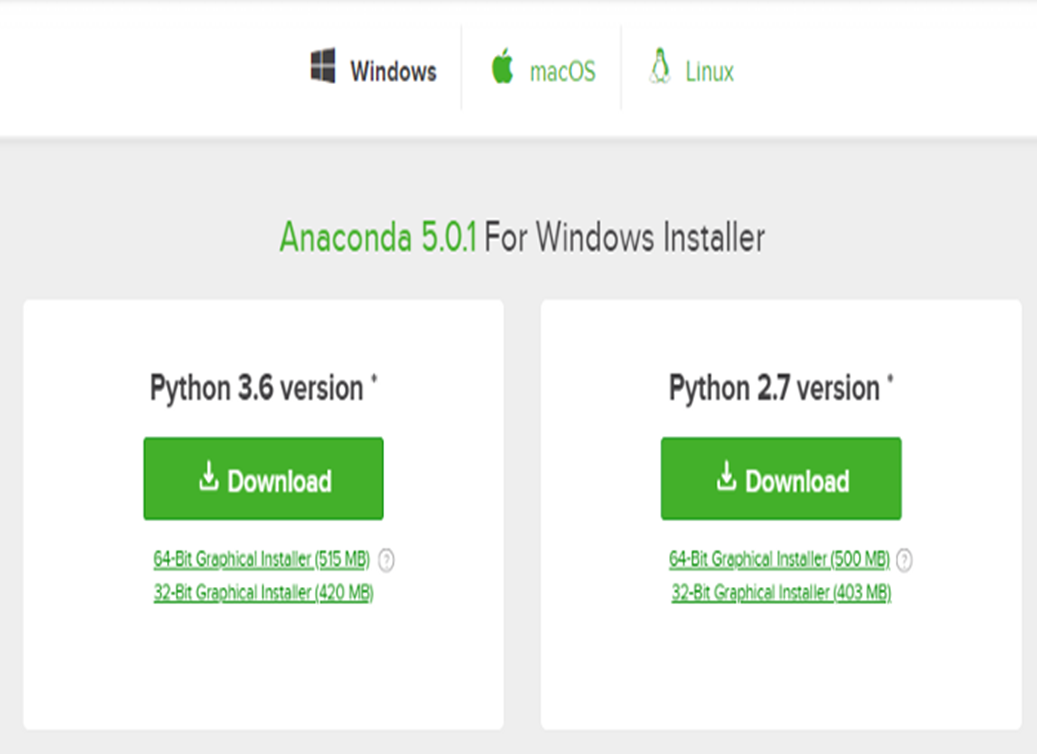
Step 1) Installing Anaconda

1. Downloads and install Anaconda from https://repo.anaconda.com/archive/Anaconda3-2021.05-Windows-x86\_64.exe.

2. After opening link u can see this download option

3. click on the download option in above image.

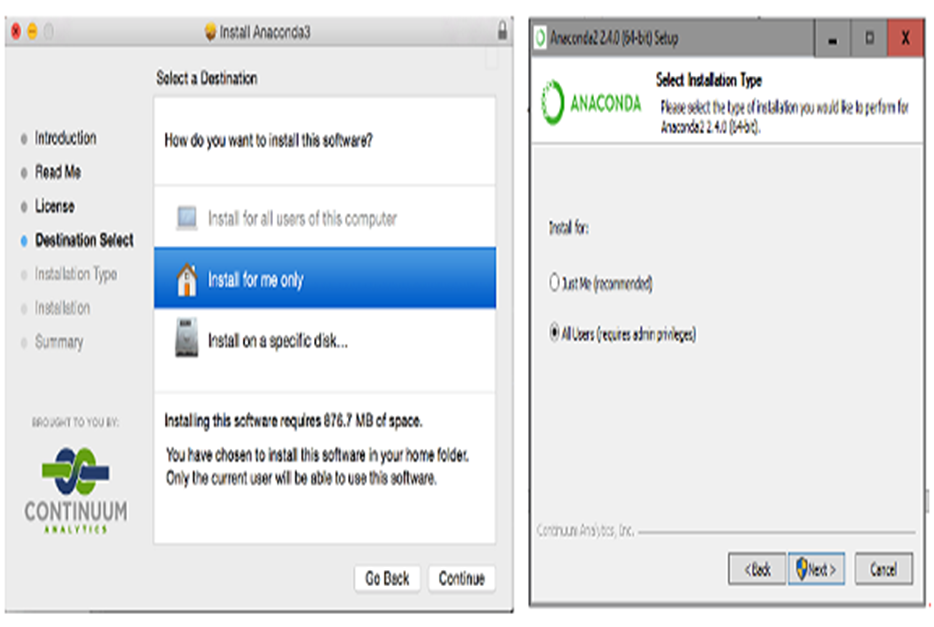
4. after downloading start installation.



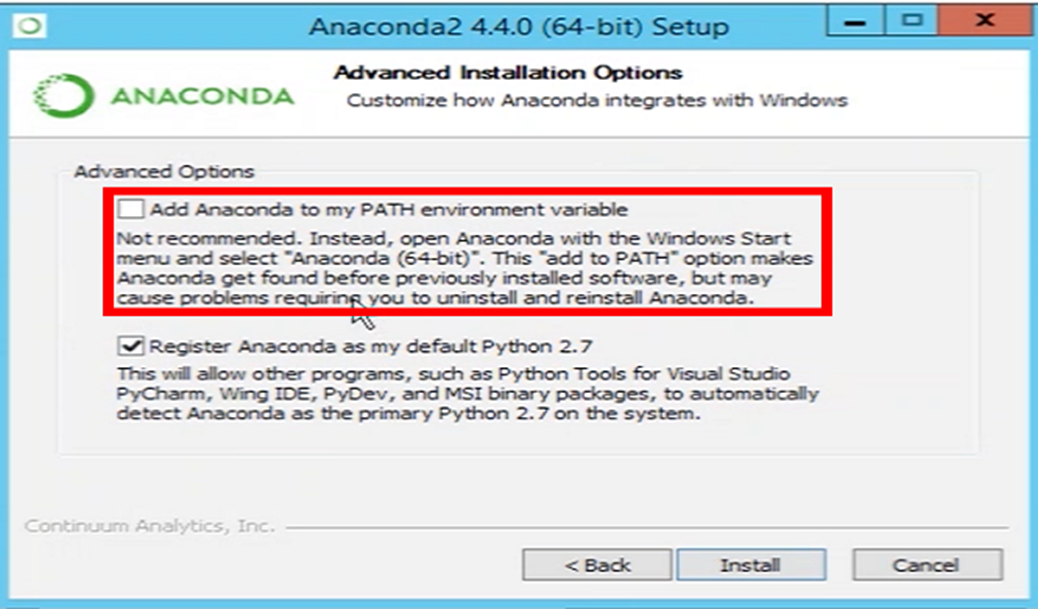
17

3. Select the default options when prompted during the installation of Anaconda as shown above.

4. ensure that the path to the folder where Anaconda is installed is added to your computer/system.



18



5. Open “Anaconda Prompt” by finding it in the Windows (Start) Menu.

6. Type the command in red to verified Anaconda was installed.

> python --version

Python 3.7.3

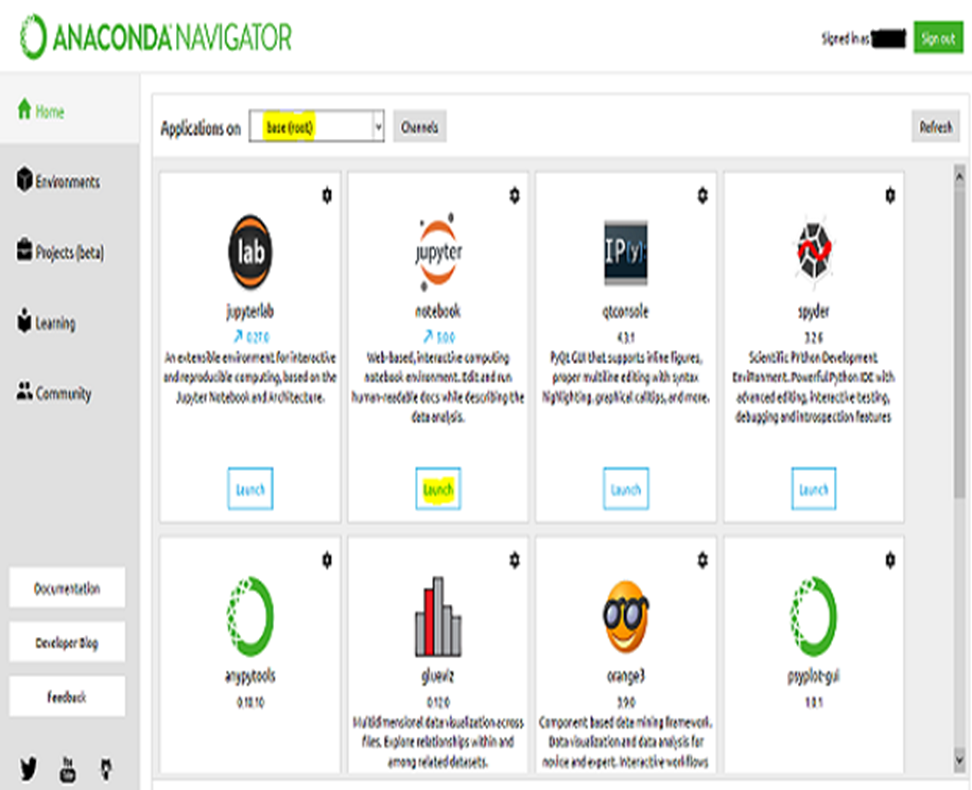
7. Type the command in red to update Anaconda.

> conda update --all –yes

• Start Jupyter Notebook

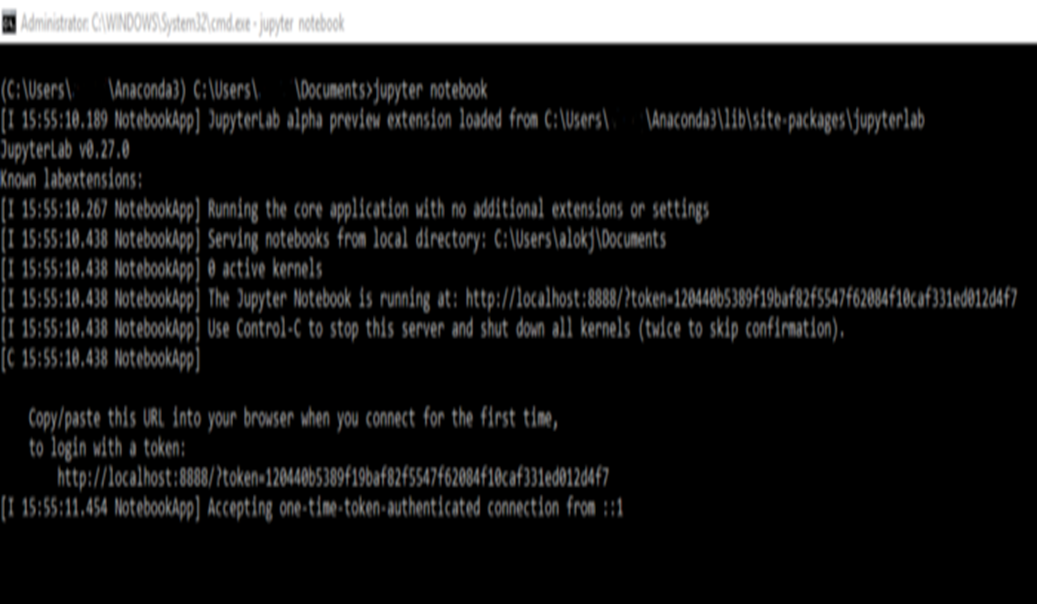
1. open anaconda navigator and the screen which is similar to below appears.

19

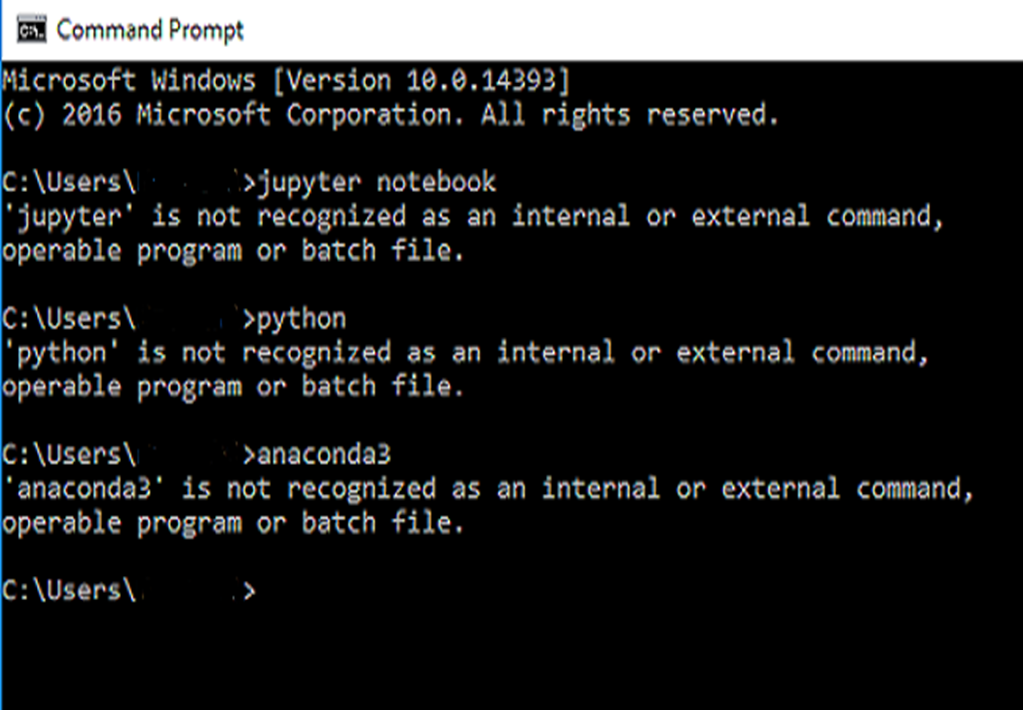


1. open anaconda prompt to open jupyter note book.

20



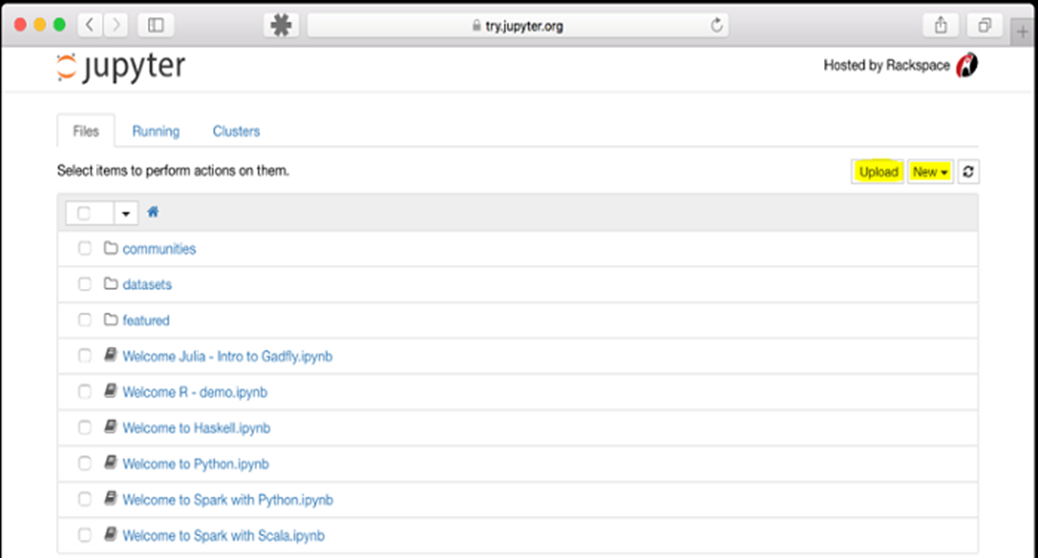
21



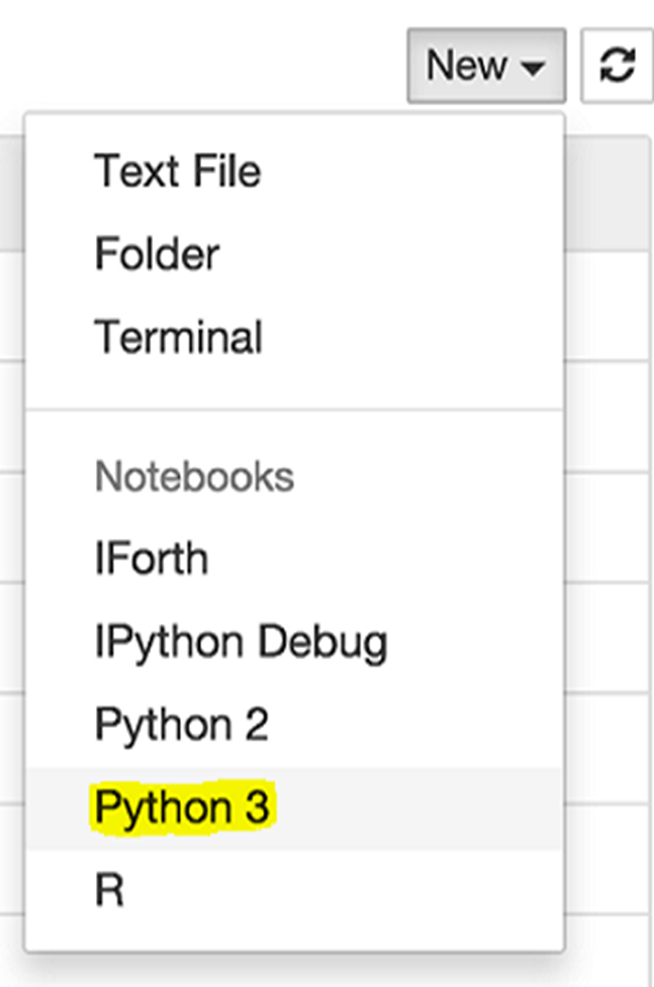


22

4. now open jupyter new kernel.



23



•Installing required packages.

24

**5.2 Module description:**

* **Data Collection:** Dataset used in this model are transactions made by credit cardholders.
* **Data Pre-Processing:** Pre-processing is the process of three important and common steps as follows:

**Formatting:** It is the process of putting the data in a legitimate way that it would be suitable to work with. Most recommended format is .csv files.

**Cleaning:** Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on.

**Sampling:** This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding the behaviour and pattern of data in an integrated way.

* **Feature Extraction:** Feature extraction is the process of studying the behaviour and pattern of the analyzed data and draw the features for further testing and training. Finally, our models are trained using the Classifier algorithm. Random forest algorithm was used to classify pre-processed data.
* **Display Graph:** The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data.
* It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions.

25

**NUMPY:**

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

• A powerful N-dimensional array object

• Sophisticated (broadcasting) functions

• Tools for integrating C/C++ and Fortran code

• Useful linear algebra, Fourier transform, and random number capabilities

NumPy can also be used as an efficient multi-dimensional container of generic data.

Arbitrary data-types can be defined using Numpy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

**PANDAS**

Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or PYTHON

**SCIKIT LEARN**

Scikit-learn is the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

sklearn is used to build machine learning models.

It should not be used for reading the data, manipulating and summarizing it. There are better libraries for that (e.g. NumPy, Pandas etc.)

26

**MATPLOTLIB**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

27

**5.3 Sample code:**

**#importing the necessary packages**

import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

from matplotlib import gridspec

**#loading the data**

data = pd.read\_csv('credit card fraud.csv')

**#describing the data**

data.head()

**#describing the data**

import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

from matplotlib import gridspec

**#checking whether any null values are present**

data.info()

**#determining the fraudulent and not fraud transactions**

Fraud = data[data['Class'] == 1]

notfraud = data[data['Class'] == 0]

outlier\_fraction = len(Fraud)/float(len(notfraud))

print(outlier\_fraction)

print('Fraud Cases: {}'.format(len(data[data['Class'] == 1])))

print('notfraud Transactions: {}'.format(len(data[data['Class'] == 0])))

**#Correlation matrix**

corrmat = data.corr()

fig = plt.figure(figsize = (12, 9))

28

sns.heatmap(corrmat, vmax = .8, square = True)

plt.show()

**#dividing the X and the Y from the dataset**

X=data.drop(['Class'], axis=1)

Y=data['Class']

print(X.shape)

print(Y.shape)

**#getting just the values for the sake of processing (its a numpy array with no columns)**

X\_data=X.values

Y\_data=Y.values

**#training and testing the data**

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_data, Y\_data, train\_size=0.70, test\_size=0.30, random\_state=1)

Building logistic regression model using skicit learn

**#Instantiate the model to an empty object**

from sklearn.linear\_model import LogisticRegression

model = LogisticRegression()

**#Train the model using 'fit' method**

model.fit(X\_train, y\_train)

**#Test the model using 'predict' method**

y\_pred = model.predict(X\_test)

29

**#Print the classification report**

From sklearn.metrics import classification\_report,confusion\_matrix,accuracy\_score

print(classification\_report(y\_test, y\_pred))

print(classification\_report(y\_test, y\_pred))

print(confusion\_matrix(y\_test, y\_pred))

print(accuracy\_score(y\_test, y\_pred))

**#printing the confusion matrix**

LABELS = ['Notfraud', 'fraud']

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

plt.figure(figsize=(12, 12))

sns.heatmap(conf\_matrix, xticklabels=LABELS, yticklabels=LABELS, annot=True, fmt="d");

plt.title("Confusion matrix")

plt.ylabel('True class')

plt.xlabel('Predicted class')

plt.show()

**#Building the Random Forest Classifier (RANDOM FOREST)**

from sklearn.ensemble import RandomForestClassifier

# random forest model creation

rfc = RandomForestClassifier()

rfc.fit(X\_train,y\_train)

**# predictions**

y\_pred = rfc.predict(X\_test)

**#Print the classification report**

From sklearn.metrics import classification\_report,confusion\_matrix,accuracy\_score

print(classification\_report(y\_test, y\_pred))

print(confusion\_matrix(y\_test, y\_pred))

print(accuracy\_score(y\_test, y\_pred))

30

**#printing the confusion matrix**

LABELS = ['Notfraud', 'fraud']

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

plt.figure(figsize=(12, 12))

sns.heatmap(conf\_matrix, xticklabels=LABELS, yticklabels=LABELS, annot=True, fmt="d");

plt.title("Confusion matrix")

plt.ylabel('True class')

plt.xlabel('Predicted class')

plt.show()

31

**6.SYSTEM TESTING**

**Tests:**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

The testing phase involves the testing of development system using various data. Preparation of the test data plays a vital role in system testing. After preparing the test data, the system under study was tested using those data. While testing the system, by using the test data, errors were found and corrected by using the following testing steps and corrections were also noted for future use. Thus, a series of testing is performed on the proposed system before the system is ready for implementation.

**Unit testing:**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration.

This is a structural testing, that relies on knowledge of its construction and is invasive.

Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

32

**Integration testing:**

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent.

Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

**Validation Testing:**

At the culmination of the integration testing, the software is completely assembled as a package, interfacing errors have been uncovered and corrected, and a final series of software validation testing began. Here we test if the system functions in a manner that can be reasonably expected by the customer. The system is tested against the system requirement specification.

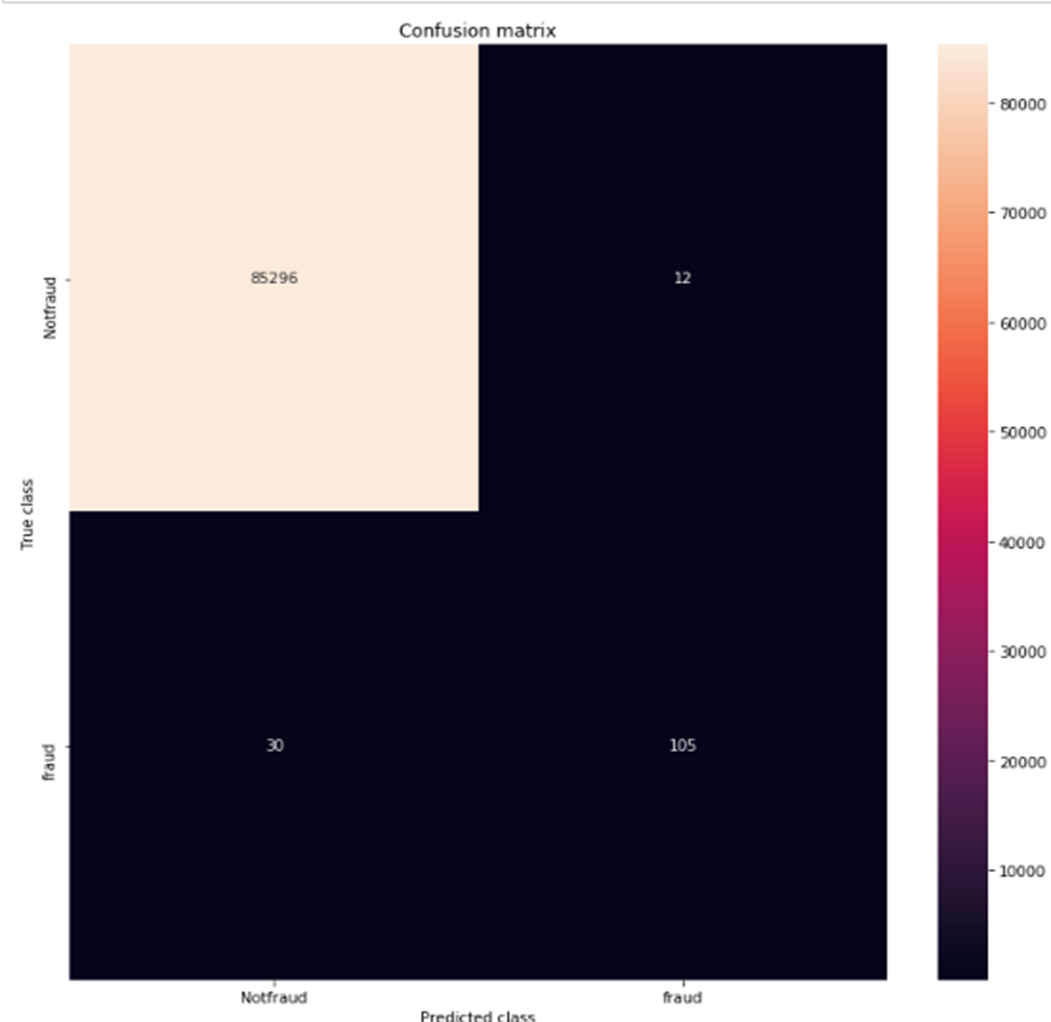
**Output Testing:**

After performing validation testing, the next phase is output testing of the proposed system, since no system can be useful if it does not produce the desired output in the specified format. The output generated or displayed by the system under consideration is tested by asking the user about the format required by them, here, the output format is considered in two ways: One is on the screen and the other is on the printed form. Beta testing is carried output by the client, and minor errors that have been discovered by the client are rectified to improve the user friendliness of the system.

33

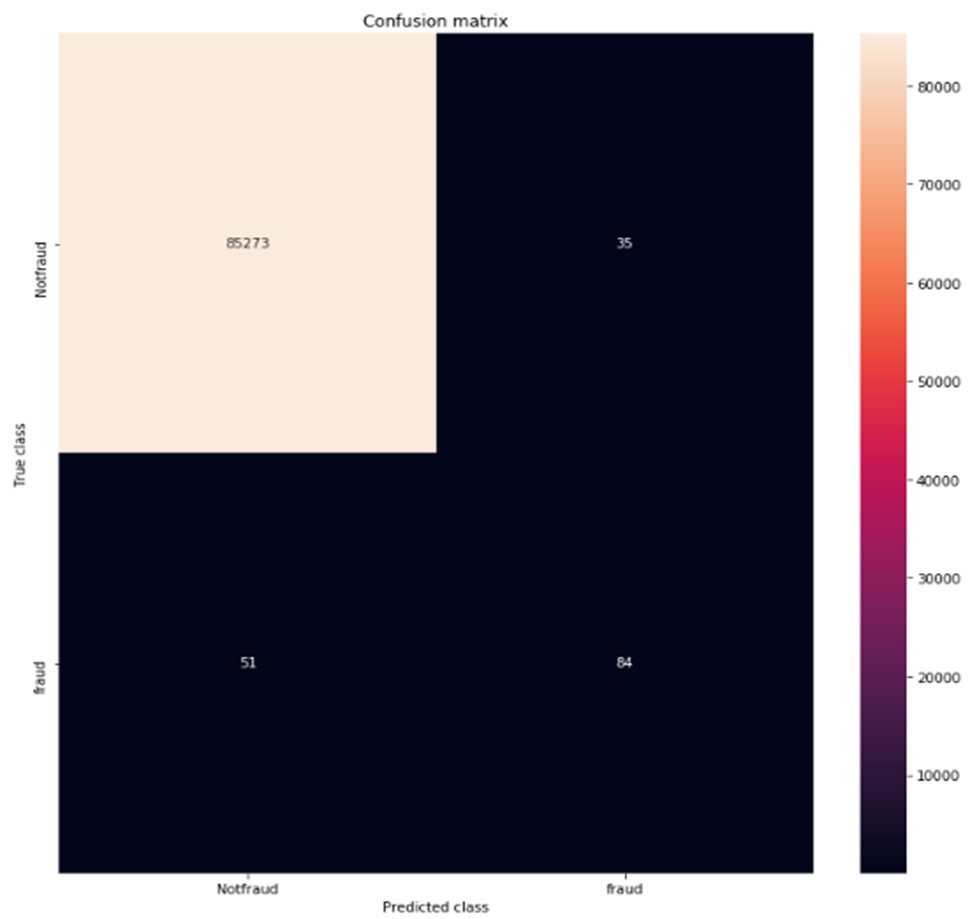
**7.RESULT SCREENSHOTS**

**Random forest:**



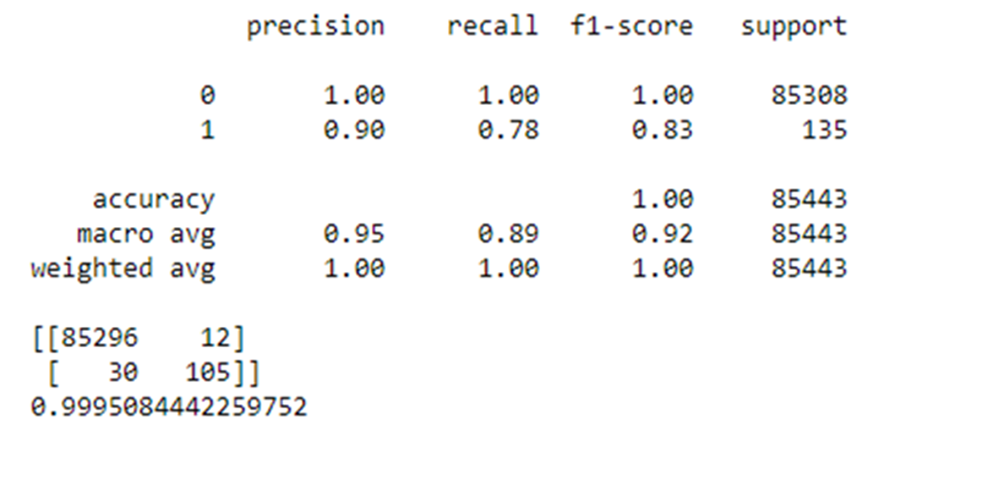
34

**Logistic regression:**

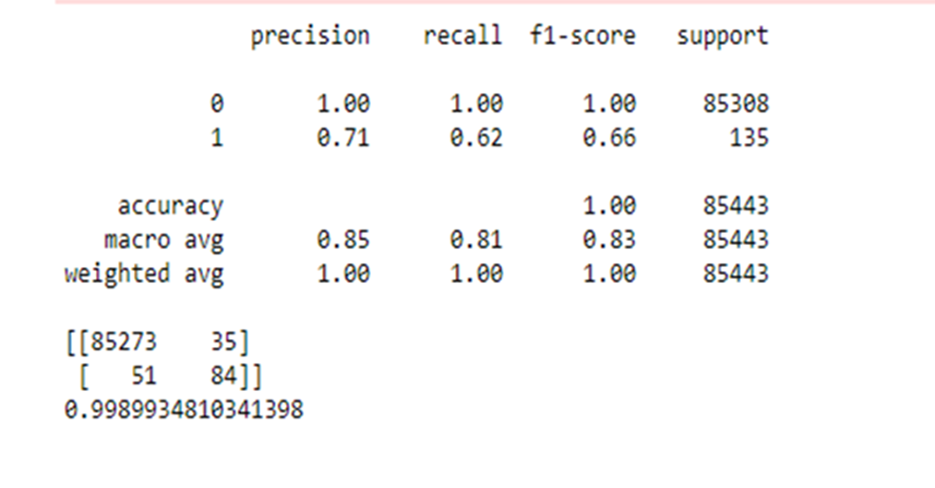
****

35

**Random forest**

****

**Logistic Regression**

****

36

**8.CONCLUSION**

* This system is capable of providing most of the essential features required to detect fraudulent and legitimate transactions.
* Although there are several fraud detection techniques available today but none is able to detect all frauds completely when they are actually happening, they usually detect it after the fraud has been committed. So we need a technology that can detect the fraudulent transaction when it is taking place so that it can be stopped then and there and that too in a minimum cost.
* The major drawback of all the techniques is that they are not guaranteed to give the same results in all environments. They give better results with a particular type of dataset and poor or unsatisfactory results with other.

• The Random forest algorithm will perform better with a larger number of

training data.

37

**9.BIBLIOGRAPHY**

1. Nathalie Japkowicz, Learning from Imbalanced Data Sets: A Comparison of Various Strategies (2000), AAAI Technical Report WS-00–05

2.https://www.researchgate.net/publication/336552027\_A\_Review\_On\_Credit\_Card\_Fraud\_Detection\_Using\_Machine\_Learning

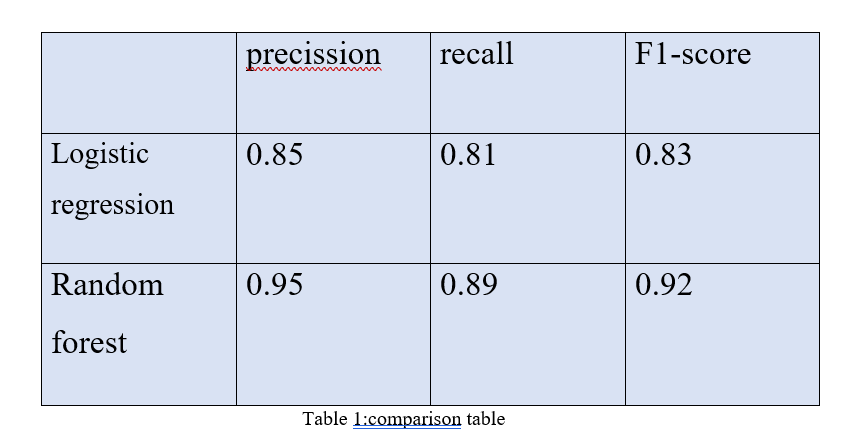
3.https://www.researchgate.net/publication/337591236\_Credit\_Card\_Fraud\_Detection\_Using\_Machine\_Learning\_With\_Python

4. Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319, 2011.

5. Y. Kou, C-T. Lu, S. Sinvongwattana, Y-P. Huang, "Survey of Fraud Detection Techniques", Proceedings of the 2004 IEEE International Conference on Networking Sensing & Control, 2004.

38

**LIST OF TABLES**

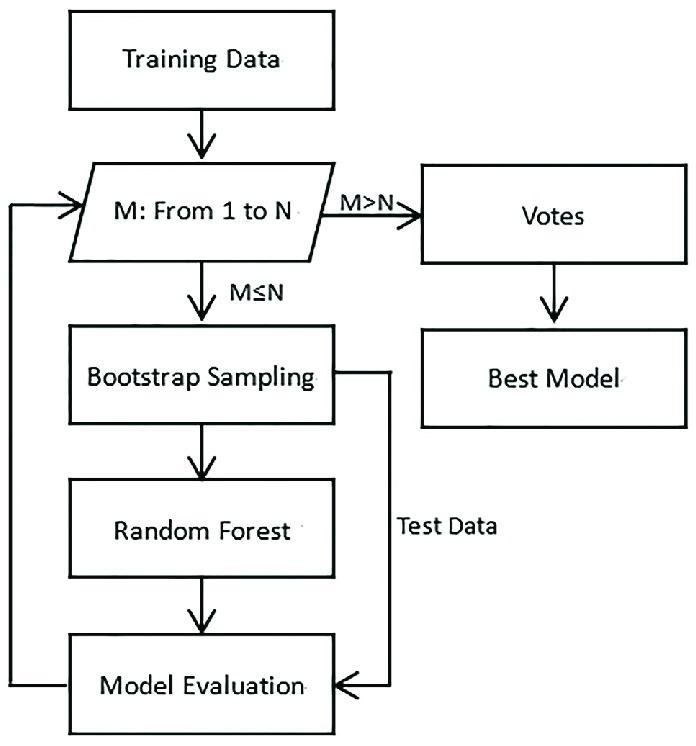


39

**LIST OF FIGURES**

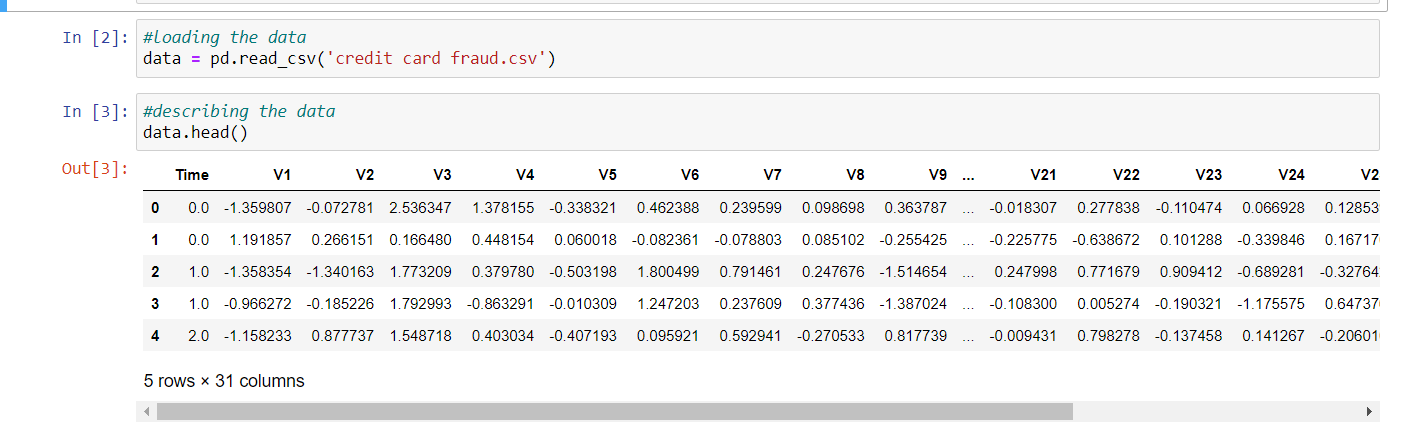
****

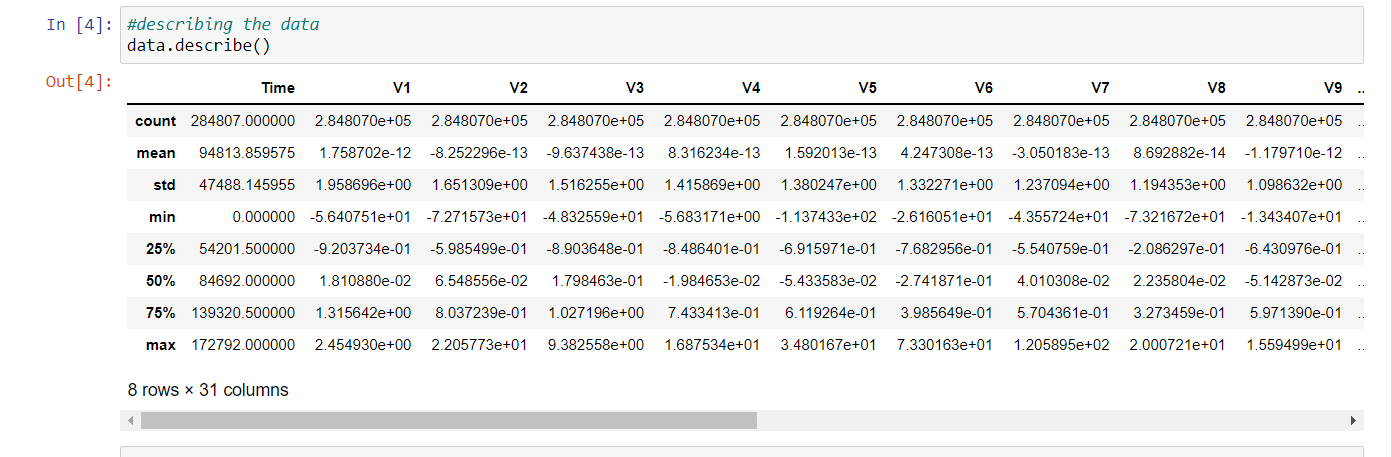
40

****

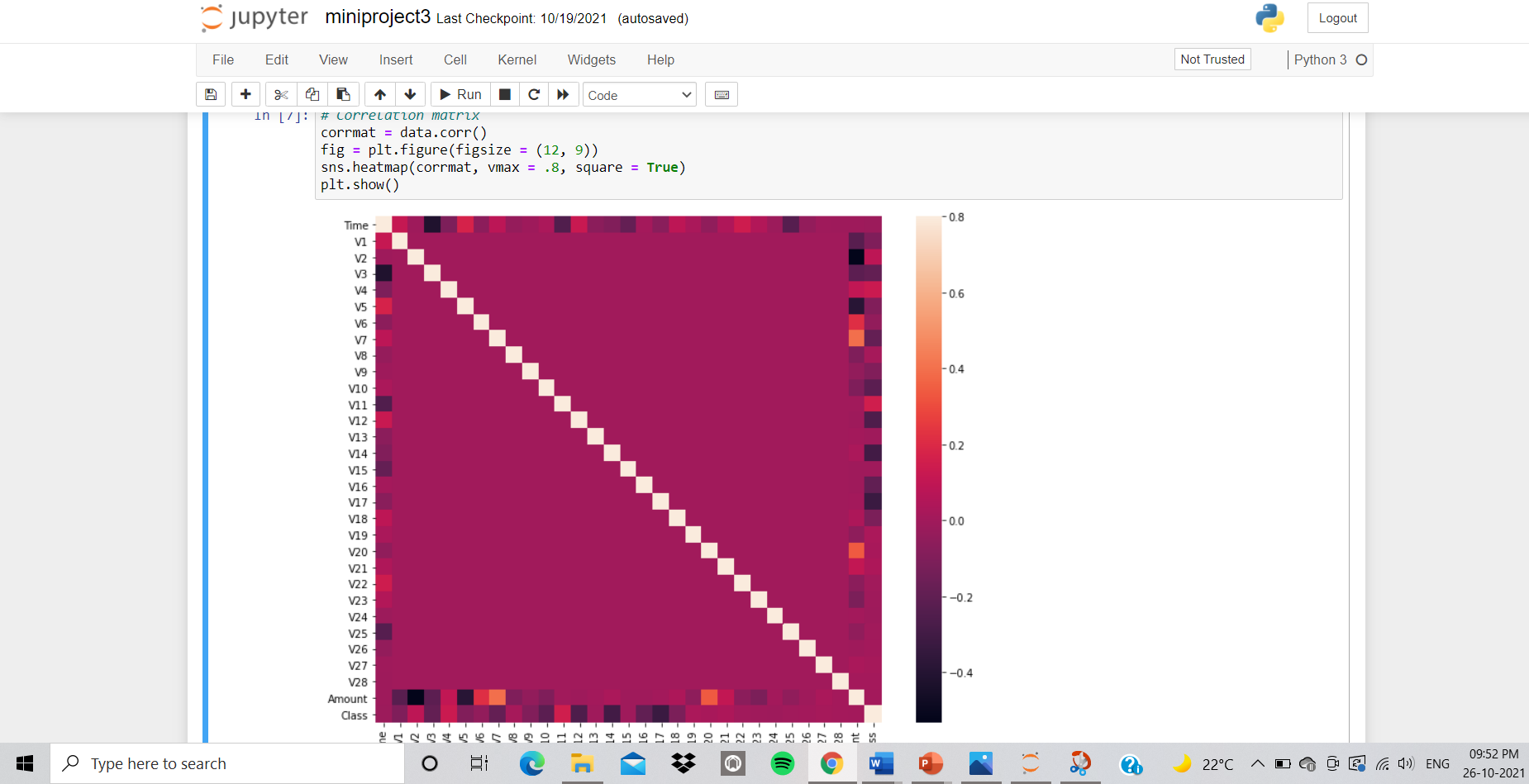
41

**LIST OF SCREENSHOTS**





42



43