IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING PREDICTIVE MODEL TO DETECT THE FRAUDULENT USE OF CREDIT CARD

[1]Utsab Ray, [2]Barish Mondal, [3]Karabi Ganguly

[1]B.Tech, Department of Biomedical Engineering, [2] B.Tech, Department of Electrical Engineering,
[3]Associate Professor and HoD, Department of Biomedical Engineering

[1]Department of Biomedical Engineering,
[1]JIS College of Engineering, Kalyani, West Bengal, India

Abstract: In recent years, credit cards have become the most popular payment method. As technology advances, so does the quantity of fraud instances, necessitating the development of a fraud algorithm for accurately detect and eliminate fraudulent actions. People have been looking for creative ways to illegally access someone's finances since the invention of payment systems. This threat has escalated in recent years, as the vast majority of transactions are now conducted totally online with credit card information. Credit Card Fraud is a broad term that applies to any sort of fraud employing a payment card, particularly a credit card. The sole goal of such infractions is usually to obtain products as well as services, or to pay a significant transfer to another account even without owner's permission. This project intends to concentrate mostly on algorithms for machine learning. Here a comparative study between several algorithms namely SVM, Decision Tree, Naïve Bayes, KNN and Logistic Regression has been done. The Logistic regression algorithm has been employed due to its better accuracy. The accuracy is being used to calculate the results. Thus, a predictive study has been developed which is used to compute the status of any credit card with respect to the transaction.

Keywords **- Credit Card, Fraud, Algorithms, Logistic Regression, Accuracy, Predictive system.**

I. INTRODUCTION

Fraud involving credit cards is an increasing concern in today's society, as is fraud in public offices, corporate industries, financial industries, and a variety of other organizations. In today's society, the heavy reliance on the internet is the cause of a spike in credit card fraud transactions, although fraud has increased not only internet but also offline. Although information retrieval techniques [1] are applied, the results are not very reliable in detecting credit card fraud. The only way to lower these costs is really to defraud using efficient algorithms, which is a plausible method of reducing credit card scams. The only way to reduce these costs is to detect fraud using efficient algorithms, which is a promising technique to reduce credit card fraud. The major goal is to create a fraud detection system that detects fraudulent transactions in less time and with more accuracy by utilizing machine learning-based categorization techniques. As technology advances, cash payments are decreasing and online payments are increasing, allowing fraudsters to conduct anonymous transactions.

In some online payment methods, only the cardholder name, maturity date, and CVV are necessary, and that data may be lost without our knowledge; in some circumstances, we are unaware that our data is being taken.We are unaware that our data has been leaked due to purchases made over the internet when criminals utilize phishing techniques to obtain the details. He only requires card data for some purchases to commit fraud, and the user may

not be aware that his/her credit card information has been compromised. The card information should be kept secret. Information may be disclosed as a result of phishing sites, and the device could be lost or stolen. The best technique to determine if a transaction is fraudulent or not is to examine the customer's purchasing habits using available data and utilize machine learning to determine whether such a transaction is legitimate or not.

Fraud in any form is a criminal offence, and credit card fraud means stealing money. Many research have been conducted to determine if a transaction is fraudulent or not. Having several hurdles and attempting to overcome them [2]. To begin, many employed Data Mining Methods to detect fraudulent transactions by employing some unconventional methodology, and these days criminals are so skilled that they may commit fraud without breaking any regulations [3]. As a result, using computer vision is common. So, in this case, a significantly unbalanced data set is analyzed in order to provide us with the optimal algorithm to apply along with its problems. Because it is heavily imbalanced, regardless of whether the proposed method is efficient or not, it provides an accuracy of around 99.9%. As in [4], under sampling is thought to produce beneficial outcomes in this case. As shown in [5], outlier detection and eradication techniques are utilized to accurately identify fraudulent transactions in a credit card payment dataset.

People in today's generation are more interested in receiving things online than going out and buying them, and as a result, the rise of ecommerce platforms is increasing, as is the risk of credit card theft. To minimize such financial crimes, we must first identify the optimum algorithm for decreasing credit card frauds.

II. RELATED WORK

New ways for detecting credit card fraud using a variety of research methods or fraud detection tools, with a focus on neural network models, data analysis, and distributed storage mining. A variety of other techniques are employed to detect credit card fraud. After conducting a literature review on multiple methodologies of credit scoring, we can infer that there are many additional ways in Machine Learning on its own to recognize credit card fraud.

Fraud detection using credit cards research employs combination Machine Learning[6][7] and Computational Intelligence algorithms[8]. In this segment, we improve the job completed in two areas: (i) the widely available tool for detecting fraud, and (ii) the approaches for dealing with imbalanced data. A [9] provides some ways for dealing with unbalanced data. They are (a) modular, (b) sampling methods, and (c) procedures that are similar.

In 2019, Yashvi Jain, Namrata Tiwari, Shripriya Dubey, and Sarika Jain studied various techniques [10] for credit card fraud detection, including support vector machines (SVM), artificial neural networks (ANN), Gradient boosting, Markov Prototype, K-Nearest Neighbours (KNN) Proposed Fuzzy system, and Decision Trees. In their paper, they discovered that the k-nearest neighbor, decision trees, and SVM algorithms provide medium-level accuracy. A high detention rate is provided by neural networks, naïve bayes, fuzzy systems, and KNN. Algorithm pertaining to Logistic Regression & fuzzy logic imposed lowest accuracy. At the middle level, the LogisticRegression, SVM, decision trees provide a high detection rate. There are two methods, ANN and Nave Bayesian Networks, that perform better across all parameters. These are quite expensive to train. All algorithms have a significant flaw. The disadvantage is that these techniques do not produce the same results in various contexts. They produce better outcomes with one type or dataset and worse results with another.

Heta Naik and Prashasti Kanikar conducted research on numerous algorithms [11] in 2019. Among the categorization algorithms, Nave Bayes stands out. Bayes' theorem determines the likelihood of an event occurring. Logistic regression algorithms are comparable to linear regression algorithms.

In 2019, Sahayasakila V, D.Kavya Monisha, Aishwarya, Sikhakolli Venkatavisalakshiswshai Yasaswi revealed the Whale Optimization Techniques (WOA) and SMOTE, two essential algorithmic techniques [12]. They were primarily concerned with increasing convergence speed and resolving the data imbalance issue. The SMOTE and WOA techniques are used to solve the class imbalance problem. The SMOTE technique discriminates all synthetic transactions, which are then re-sampled to ensure data accuracy and optimised using the WOA technique. The method also enhances the system's convergence speed, accuracy, and efficiency.

Navanushu Khare and Saad Yunus Sait [13] presented their work in 2018. They started with a very skewed dataset and moved their way up. Accuracy, sensitivity, precision, are used to assess project. The accuracy for Logistic Regression is 97.7%, Decision Trees is 95.5%, Random Forest is 98.6%, and SVM classifier is 97.5%, according to the results. They discovered that perhaps the Random Forests really does have the maximum accuracy of all the methods and is the best algorithm for detecting fraud. They also discovered that now the SVM algorithm suffers from data imbalance and does not perform much better in detecting credit card fraud.

III. PROPOSED WORK

The work has been conceptualized in figure(i)..

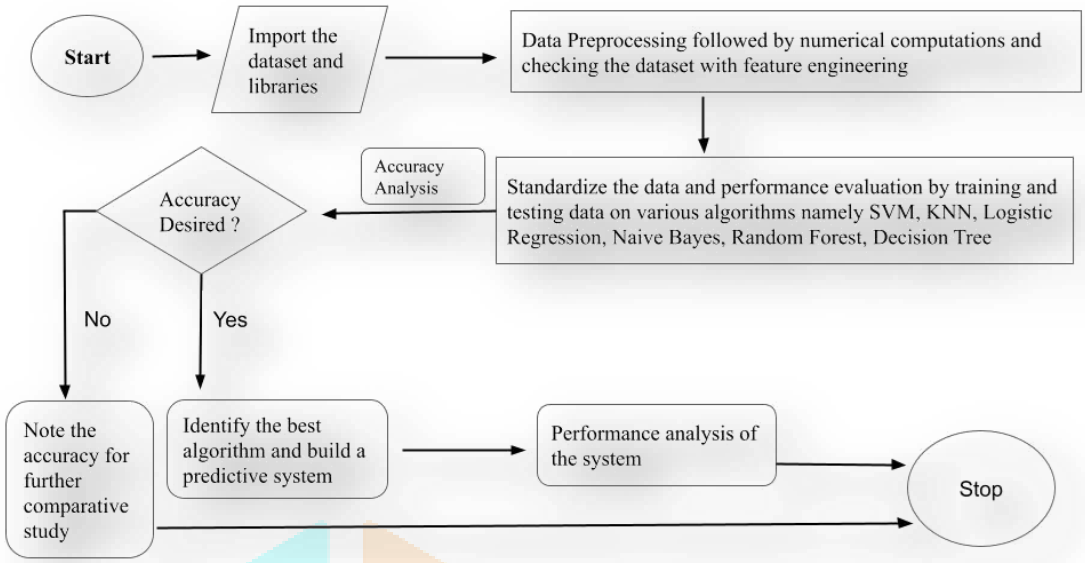

Figure (i) - Proposed Methodology

Initially dataset has been gathered and machine learning pipeline has been initiated. Machine Learning follows mainly the principle of training data proceeding testing the model and checking for predictions and further deployment. The pipeline of machine learning workflow has been identified in figure (ii).

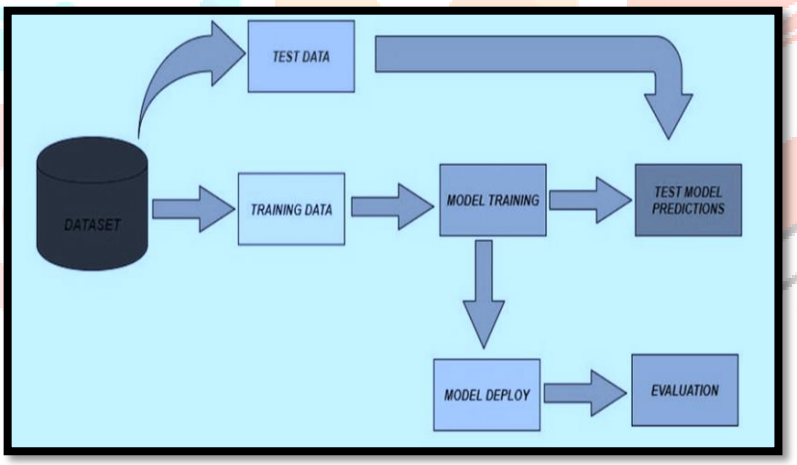

Figure (ii): Machine Learning workflow

Following this pipeline the algorithms namely Random Forest,KNN, Decision Tree, Naïve Bayes and Logistic Regression are deployed on this particular dataset to compute credit card frauds and it is visualized that Logistic Regression performs best on testing data and it is further required to develop a predictive system which can easily identify the specific fraudulent transaction pertaining to any particular card.

IV. RESULTS AND DISCUSSIONS

V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
.617801	-0.991390	-0.311169	1.468177	-0.470401	0.207971	0.025791	0.403993	0.251412	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
.065235	0.489095	-0.143772	0.635558	0.463917	-0.114805	-0.183361	-0.145783	-0.069083	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
.066084	0.717293	-0.165946	2.345865	-2.890083	1.109969	-0.121359	-2.261857	0.524980	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
.178228	0.507757	-0.287924	-0.631418	-1.059647	-0.684093	1.965775	-1.232622	-0.208038	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
.538196	1.345852	-1.119670	0.175121	-0.451449	-0.237033	-0.038195	0.803487	0.408542	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

Figure (iii): Classify the dataset

In figure (iii) dataset classification which is a significant measure to perform any computational check.

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	
284802	172786.0	-11.881118	10.071785	-9.834783	-2.066656	-5.364473	-2.606837	-4.918215	7.305334	1.914428	4.356170	-1.593105	2.711941	-0.689256	4.626942	-0.924459	1.107641	1.9
284803	172787.0	-0.732789	-0.055080	2.035030	-0.738589	0.868229	1.058415	0.024330	0.294869	0.584800	-0.975926	-0.150189	0.915802	1.214756	-0.675143	1.164931	-0.711757	-0.(
284804	172788.0	1.919565	-0.301254	-3.249640	-0.557828	2.630515	3.031260	-0.296827	0.708417	0.432454	-0.484782	0.411614	0.063119	-0.183699	-0.510602	1.329284	0.140716	0.1
284805	172788.0	-0.240440	0.530483	0.702510	0.689799	-0.377961	0.623708	-0.686180	0.679145	0.392087	-0.399126	-1.933849	-0.962886	-1.042082	0.449624	1.962563	-0.608577	0.5
284806	172792.0	-0.533413	-0.189733	0.703337	-0.506271	-0.012546	-0.649617	1.577006	-0.414650	0.486180	-0.915427	-1.040458	-0.031513	-0.188093	-0.084316	0.041333	-0.302620	-0.(

Figure (iv): Identification of dataset corresponding values with respect to credit cards as per dataset

In figure (iv) the stratification of the dataset has been done to visualize the card information which is an important step ahead computation of fraud identification.

```
# dataset informations
credit_card_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column  Non-Null Count   Dtype  
---  ------  --------------   -----  
 0   Time    284807 non-null  float64
 1   V1      284807 non-null  float64
 2   V2      284807 non-null  float64
 3   V3      284807 non-null  float64
 4   V4      284807 non-null  float64
 5   V5      284807 non-null  float64
 6   V6      284807 non-null  float64
 7   V7      284807 non-null  float64
 8   V8      284807 non-null  float64
 9   V9      284807 non-null  float64
 10  V10     284807 non-null  float64
 11  V11     284807 non-null  float64
 12  V12     284807 non-null  float64
 13  V13     284807 non-null  float64
 14  V14     284807 non-null  float64
 15  V15     284807 non-null  float64
 16  V16     284807 non-null  float64
 17  V17     284807 non-null  float64
 18  V18     284807 non-null  float64
 19  V19     284807 non-null  float64
 20  V20     284807 non-null  float64
 21  V21     284807 non-null  float64
 22  V22     284807 non-null  float64
```

Figure (v): Understanding of Data

Figure (v) shows that analysis of the data employs algorithms to continuously improve itself, but adequate data is essential for these models to function successfully, which entails obtaining and analysing data using tables and graphics.

```
# distribution of legit transactions & fraudulent transactions
credit_card_data['Class'].value_counts()
```

```
0    284315
1       492
Name: Class, dtype: int64
```

This Dataset is highly unblanced

Figure (vi): Distribution of Transactions

In figure(vi) the distribution of transactions has been diversely visualized and it is depicted that the dataset is highly imbalanced as the distribution of fraud transactions {1} and true transactions {0} varies to a larger extent.

```
[ ] # statistical measures of the data
    legit.Amount.describe()

count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64

[ ] fraud.Amount.describe()

count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

Figure (vii): Calculation of statistical measures

Considering the dataset in figure (vii) statistical measures are computed. Scientific study requires statistical methodologies. In fact, statistical approaches dominate scientific research since they involve planning, designing, data collection, analysis, meaningful interpretation, and publishing of study findings.

```
new_dataset.tail()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	
279863	169142.0	-1.927883	1.125653	-4.518331	1.749293	-1.566487	-2.010494	-0.882850	0.697211	-2.064945	-5.587794	2.115795	-5.417424	-1.235123	-6.665177	0.401701	-2.897825	-4.57(
280143	169347.0	1.378559	1.289381	-5.004247	1.411850	0.442581	-1.326536	-1.413170	0.248525	-1.127396	-3.232153	2.858466	-3.096915	-0.792532	-5.210141	-0.613803	-2.155297	-3.26
280149	169351.0	-0.676143	1.126366	-2.213700	0.468308	-1.120541	-0.003346	-2.234739	1.210158	-0.652250	-3.463891	1.794969	-2.775022	-0.418950	-4.057162	-0.712616	-1.603015	-5.03
281144	169966.0	-3.113832	0.585864	-5.399730	1.817092	-0.840618	-2.943548	-2.208002	1.058733	-1.632333	-5.245984	1.933520	-5.030465	-1.127455	-6.416628	0.141237	-2.549498	-4.61
281674	170348.0	1.991976	0.158476	-2.583441	0.408670	1.151147	-0.096695	0.223050	-0.068384	0.577829	-0.888722	0.491140	0.728903	0.380428	-1.948883	-0.832498	0.519436	0.90

```
[ ] new_dataset['Class'].value_counts()

1    492
0    492
Name: Class, dtype: int64
```

Figure (viii):- Building a new dataset

In figure (viii) a new dataset has been has been developed to diminish the imbalance.

Splitting the data into Features & Targets

```
X = new_dataset.drop(columns='Class', axis=1)
Y = new_dataset['Class']

[ ] print(X)

            Time        V1        V2  ...       V27       V28  Amount
203131  134666.0 -1.220220 -1.729458  ...  0.173995 -0.023852  155.00
95383    65279.0 -1.295124  0.157326  ...  0.317321  0.105345   70.00
99706    67246.0 -1.481168  1.226490  ... -0.546577  0.076538   40.14
153895  100541.0 -0.181013  1.395877  ... -0.229857 -0.329608  137.04
249976  154664.0  0.475977 -0.573662  ...  0.058961  0.012816   19.60
...          ...       ...       ...  ...       ...       ...     ...
279863  169142.0 -1.927883  1.125653  ...  0.292680  0.147968  390.00
280143  169347.0  1.378559  1.289381  ...  0.389152  0.186637    0.76
280149  169351.0 -0.676143  1.126366  ...  0.385107  0.194361   77.89
281144  169966.0 -3.113832  0.585864  ...  0.884876 -0.253700  245.00
281674  170348.0  1.991976  0.158476  ...  0.002988 -0.015309   42.53

[984 rows x 30 columns]
```

Figure (ix):- Splitting into features and targets

The dataset has been split into features and targets to avoid data over fitting and further imbalances as per figure (ix).

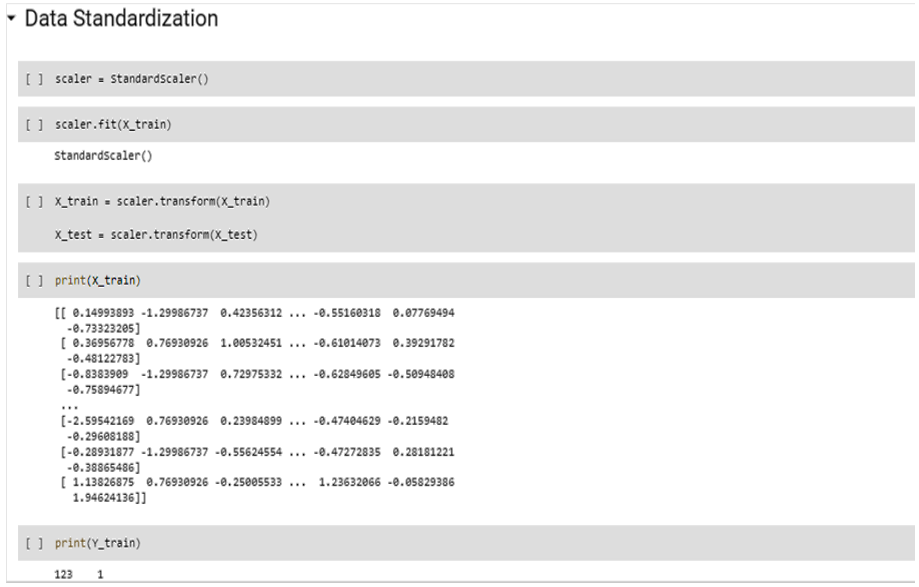

Figure (x):- Data Standardization

Figure (xi):- Accuracy analysis of KNN

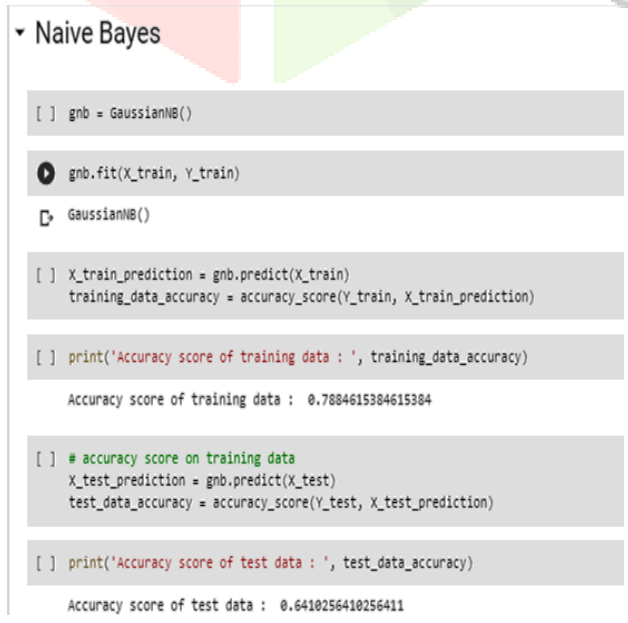

Figure (xii):- Accuracy of Naïve Bayes

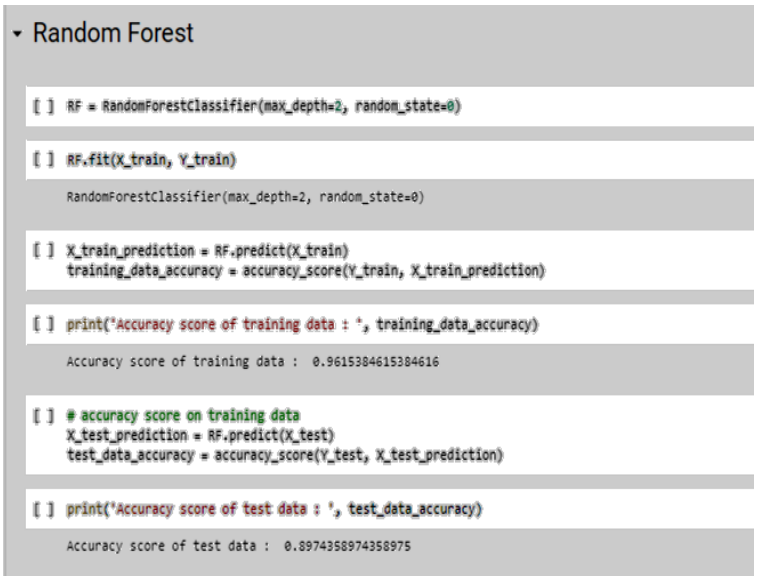

Figure (xiii):- Accuracy of Random Forest

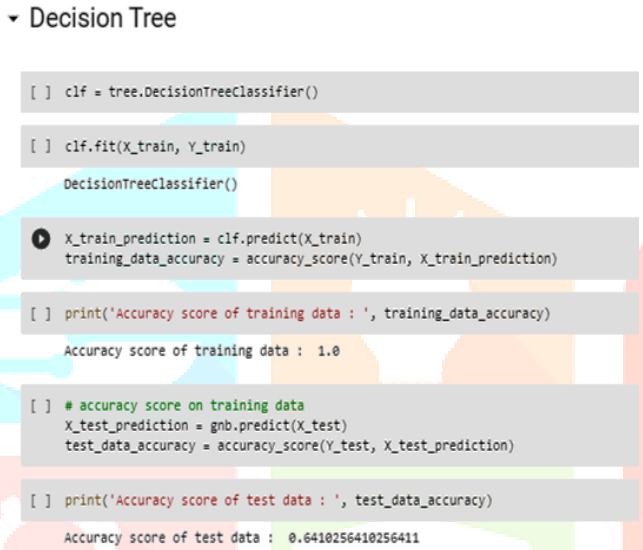

Figure (xiv):- Accuracy of Decision Tree

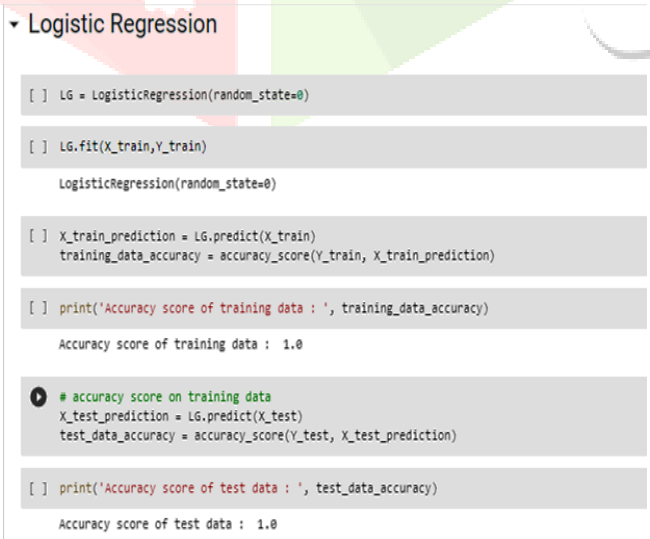

Figure (xv):- Accuracy of Logistic Regression

Thus, the comparative study between various algorithms with respect to their algorithms can be summarized as below in table 1:

Sl. No	Algorithm Name	Training Accuracy	Testing Accuracy	Rank Based on Accuracy
1	Decision Tree	100%	64.10%	4
2	Random Forest	96.15%	89.74%	3
3	Logistic Regression	100%	100%	1
4	Naïve Bayes	78.84%	64.10%	5
5	KNN	97.43%	97.43%	2

Table 1: Comparative study of algorithms

So, from Table 1, it is clear that Logistic Regression performs best and Naïve Bayes has lease accuracy on performance. Therefore, predictive system is developed with the help of Logistic Regression algorithm.

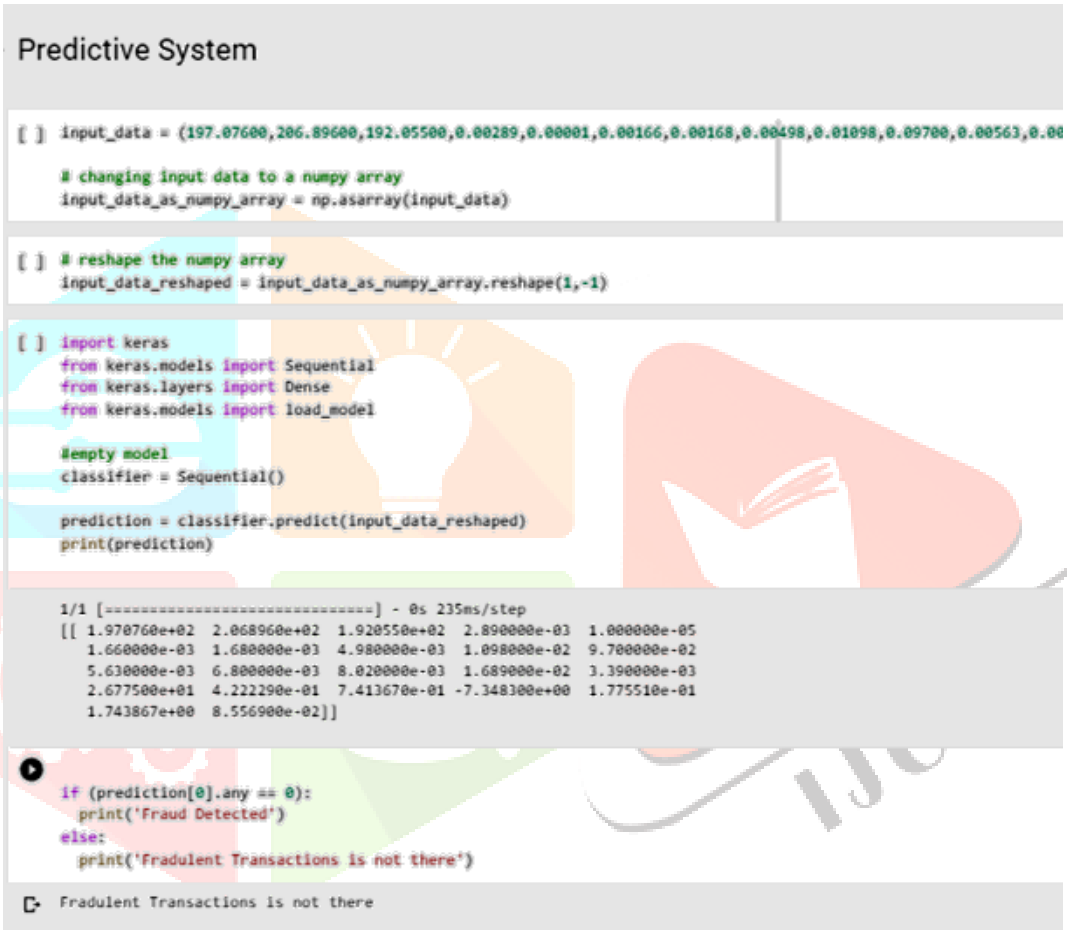

Figure (xvi):- Predictive System

V. CONCLUSION

From this work, we can conclude that a predictive system is developed with the help of Logistic Regression algorithm that performed best on this respective dataset to compute the fraud for a particular credit card. This system will help to check and identify any sort of fraudulent transaction on any credit card

VI. ACKNOWLEDGMENT

We would like to thank all the faculty and staff members from the Department of Biomedical Engineering and Department od Electrical Engineering for their support.

REFERENCES

[1] RandulaKoralage, , Faculty of Information Technology, University of Moratuwa,Data Mining Techniques for Credit Card Fraud Detection..

[2] Yashvi Jain, NamrataTiwari, ShripriyaDubey, Sarika Jain, "A Comparative Analysis of Various Credit Card Fraud Detection Techniques, Blue Eyes Intelligence Engineering and Sciences Publications 2019"

[3] John Richard D. Kho, Larry A. Vea "Credit Card Fraud Detection Based on Transaction Behaviour" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[4] Learning Robert A. Sowah, Moses A. Agebure, Godfrey A. Mills, Koudjo M. Kaumudi, "New Cluster Undersampling Technique for Class Imbalance "of 2016 IJMLC

[5] Mohamed Jaward Bah, Mohamed Hammad "Progress in Outlier Detection Techniques: A Survey" Hongzhi Wang, of the 2019 IEEE

[6] Adi Saputra1, Suharjito2L: Fraud Detection using Machine Learning in e-Commerce, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 9, 2019.

[7]Dart Consulting,Growth Of Internet Users In India And Impact On Country's Economy: https://www.dartconsulting.co.in/marketnews/growth-of-internet-users-in-india-and-impact-on-countryseconomy/

[8] Roy, Abhimanyu, et al:Deep learning detecting fraud in credit card transactions, 2018 Systems and Information Engineering Design Symposium (SIEDS), IEEE, 2018

[9] Yong Fang1, Yunyun Zhang2 and Cheng Huang1, Credit Card Fraud Detection Based on Machine Learning, Computers, Materials & Continua CMC, vol.61, no.1, pp.185-195, 2019.

[10] Yashvi Jain, NamrataTiwari, ShripriyaDubey,SarikaJain:A Comparative Analysis of Various Credit Card Fraud Detection Techniques, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-5S2, January 2019.

[11] Heta Naik , Prashasti Kanikar: Credit card Fraud Detection based on Machine Learning Algorithms,International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 44, March 2019.

[12] Sahayasakila.V, D. KavyaMonisha, Aishwarya, SikhakolliVenkatavisalakshiseshsaiYasaswi: Credit Card Fraud Detection System using Smote Technique and Whale Optimization Algorithm,International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958, Volume-8 Issue-5, June 2019.

[13] NavanshuKhare ,SaadYunus Sait: Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models, International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 825-838 ISSN: 1314-3395.