DETECTION OF CREDIT CARD FRAUDULENT TRANSACTION USING DECISION TREE AND LOGISTIC REGRESSION ALGORITHMS

[1]Ms.Benedict Tephila M
Assistant Professor,
Sri Krishna College of Engineering and Technology, Coimbatore.
benedicttephilam@skcet.ac.in

[2]Srivarshini S
UG Research Scholar,
Sri Krishna College of Engineering and Technology, Coimbatore.
19euec152@skcet.ac.in

[3]Swetha K
UG Research Scholar,
Sri Krishna College of Engineering and Technology, Coimbatore.
19euec161@skcet.ac.in

[4]Tanushree C R
UG Research Scholar,
Sri Krishna College of Engineering and Technology, Coimbatore.
19euec162@skcet.ac.in

Abstract

Credit card fraudulence is currently the major happening issue in the ongoing scene. This is presently grateful to the increase in web-based banking and disconnected exchanges. The con artist involves the card for his /her own utilizations where the card fraud happens on a regular basis. This happens for some unlawful purposes. In todays era, we face lots of credit card issues. To neutralize these fake exercises the discovery of charge card deceitfulness exchange framework is introduced. This undertaking mostly focuses to AI calculations. The algorithms applied are decision tree and logistic regression algorithm. The outcomes of these 2 algorithms assist accuracy, recall, precision, and F1-score. The decision tree and therefore the logistic regression algorithm are matched up and the best algorithm is considered based up on the factors like f1 score, exactness and recall are considered since its the technique mostly the best for recognizing extortion.

Keywords: *F1 Score, Decision tree, Logistic regression, accuracy, precision, recall.*

1.INTRODUCTION

Credit Card Fraud is characterized as a situation where a singular purpose somebody's extra card for personal cause and hence the authorities who provide the card are very much concern about the reality that the card should be utilized properly.

The consumption of credit cards for web based buying has been extraordinary due to the need and enhancement of the web based business, bringing about a high volume of Mastercard extortion. The need to detect master card thefts is critical in this era of digitalization. Checking and measuring the geste of different junkies to appraise descry or forestall horrendous geste is important for misrepresentation discovery. We might want to study ace card misrepresentation location so we can recognize it all the more successfully. The varied Detecting credit card fraud involves a variety of methods, algorithms, and kinds. The algorithm can determine whether a function is fraud or not. They must pass a dataset and have knowledge of fraudulent sales in order to detect fraud. They analyse the data and categorise all transactions. The recognition of extortion involves covering the molding of populaces of individuals addicts to avoid indefensible geste, that contains misreport, intreference, and levanting.

The usage of machine proficiency algorithms are done to break down all the accepted arrangements and report the uncertain cases. The card holders spoke with the experts to authenticate assuming that the deal has certified. The scrutinizers give criticism to the most robotized framework that is used to the guide and modernize the calculation to enhance the misrepresentation disclosure execution over the long run to coach eventually.

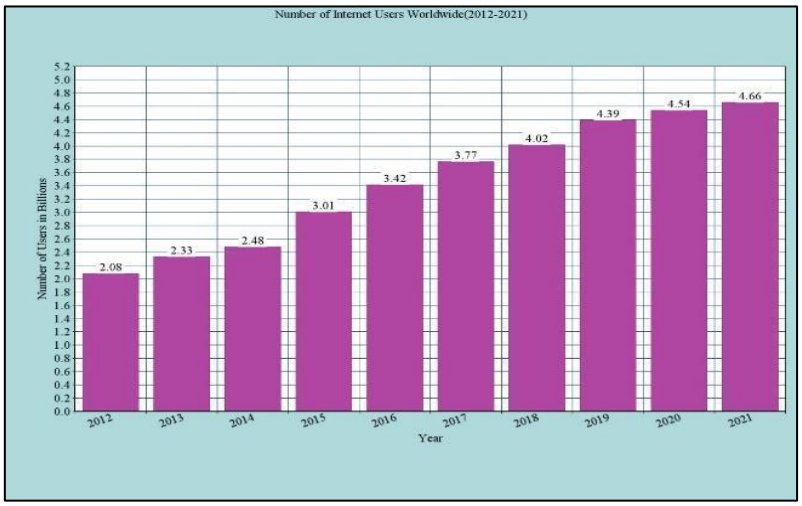

Fig. 1 Number of Internet Users Worldwide

2.LITERATURE SURVEY

The Uncertainty of Credit Card Recognition in some constant exchanges led by occasions, vulnerability is fundamental. Credit card fraud detection is an example of an uncertain area where fraud cases should be identified in real time and labelled before a sale is allowed or rejected. We provide IBM Proactive Technology Online (PROTON) open source query management application. The inclusion of inquiry components has an influence on all

areas of an event processing machine's armature and sensibility. PROTON's extensions incorporate the consideration of extra implicit qualities and capabilities, as well as help for new kinds of operands and occasion handling examples to deal with them all. Bayesian Network, KNN, ANN, SVM ,Fuzzy Sense Grounded System, are a portion of the vivid methods open for an fraud recognition framework. A comprehensive evaluation of existing and suggested credit card extortion identification models has been led, as well as a near investigation of various techniques utilizing quantitative factors, for example, delicacy, disclosure rate, and deception rate. Our study's end outlines the disadvantages of becoming a model and offers a better solution for overcoming them. A study of Credit Card Fraud Detection Systems Credit cards have turned into a well known instalment strategy for both on the web and disconnected buys, and the frequency of charge card misrepresentation is on the ascent. Credit card scams are becoming more common by the day, owing to the diverse that have been conceived for their identification. Fraudsters are gifted to the point that they plan new procedures to commit deceitful exchanges consistently, requiring steady innovation for their revelation strategies. Most of the techniques in Neural Networks have advanced in distinguishing vivid credit card false arrangements. This paper analyzes the different shaded strategies used in credit.

3.PROPOSED TECHNIQUE:

The rearmost machine proficiency calculations are made utilized in this proposed paper to descry ace card falseness. In the first place, the data required for this project is taken from the Kaggle organization. There are 31 sections in the data set that are named from v1 to v28 to safeguard delicate information. The opposite sections address Time, Quantum, and confusion. Time uncover the time space between the underlying exchange and the following one in a moment or two. Quantum is that the quantum of money transcated. The Class addresses the sort of sale. (i.e) 0 indicates a ligit deal and 1 signifies a deceitful bone. We compass different attract up to test for flimsiness inside the dataset and to acknowledge it outwardly. We have proposed this paper to arrange the arrangements which have both the extortion and the non-misrepresentation bargains inside the Kaggle dataset using the machine learning algorithms like Decision Tree Classifier and furthermore like the Logistic Regression Algorithm. They figure out which calculation best distinguishes ace card extortion. The credit extortion location issue's framework convergence (Fig. 2.) includes the word's splitting, model preparation, and model organization, as well as the evaluation measures. Subsequent to checking this dataset, we order a histogram for every segment. This is much of the time continuous to prompt a graphical likeliness of the dataset which might be utilized to certify that are not there in no key card that is certainly not a false identification worth and AI algorithms can undoubtedly re-utilize the data set. After this investigation, we incorporate a temperature guide to start a brilliant show of the word and become familiar with the connection between expectation factors as well as class fluctuation. The dataset is presently arranged and reused. The time and quantum column are formalized and in this way the Class column is eliminated to affirm the good form of the assessment. The data which is taken from the modules are gathered by calculating to be reused. The back modules which has been delineated makes sense of how these calculations cooperate. This open source library which is raised by usage of NumPy, SK learn and other modules of Matplotlib which gives a powerful devices which is used for scrutinizing the

information and machine education. This integrates bright sections, grouping and review computation and there has been an expectation to work with logical libraries. The jupyter journal stage was used to make a code in python to show the methodology that this paper proposed. Google collab stage is used to execute the program this executes the python code. The underneath figure 2 addresses the framework stream plan.

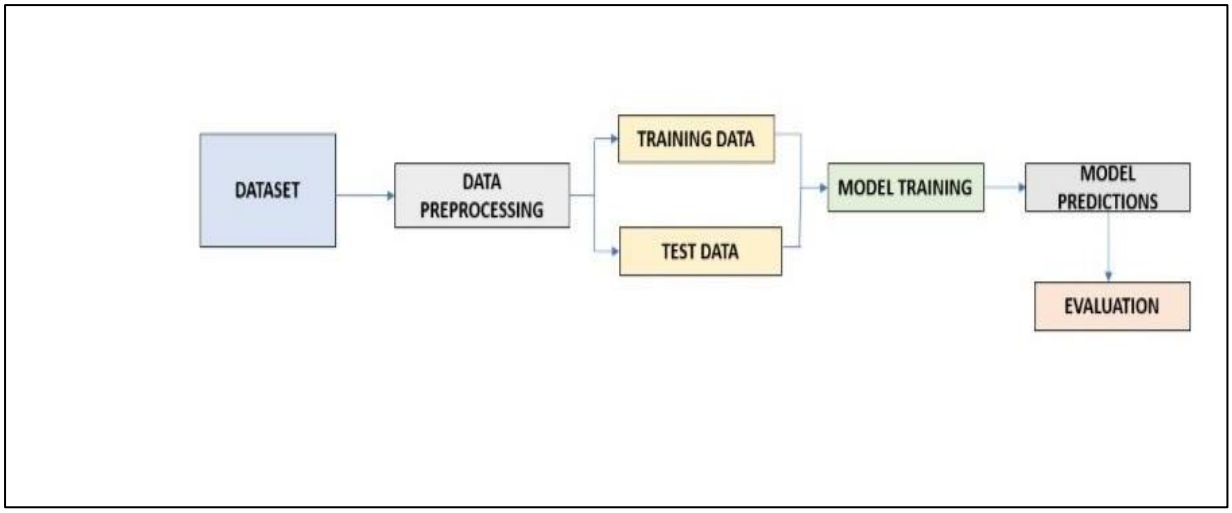

Fig. 2 System Flow Design

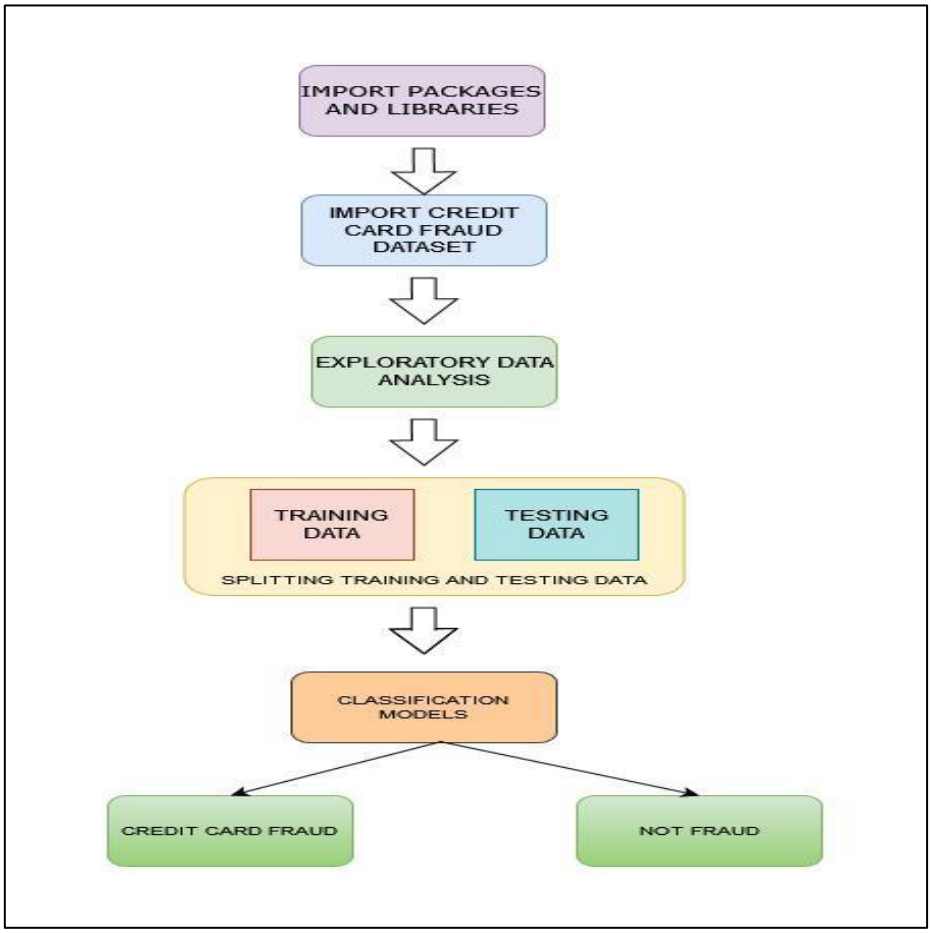

Fig. 3 Architecture Diagram

a. Importing the packages:

Pandas is used for information, NumPy is used for arrays, sklearn is used for information split, structure, and breaking down section models, lastly the matplotlib package for information perception and visual understanding will be our significant apparatuses for this plan. We should populate our Python territory with our essential packages in general.

b. Importing Dataset:

We will utilize the Kaggle credit card Misrepresentation Discovery dataset. It contains highlights V1 through V28, which are the zenith of PCA's accomplishments. The time point will be overlooked on the grounds that it makes little difference to the models. The leftover viewpoints are the 'Quantity' point, which contains the general quantum of magnate being executed, and the 'Class' point, which contains regardless of whether the deal is an extortion case.In our dataset , directing the degree of connections in the midst of the factors is a valuable data. This data is useful in picking with which highlights to prize or which machine proficiency model to choose . The correlation matrix gives a visual outline of the connection values in the midst of the elements and the outgrowth.

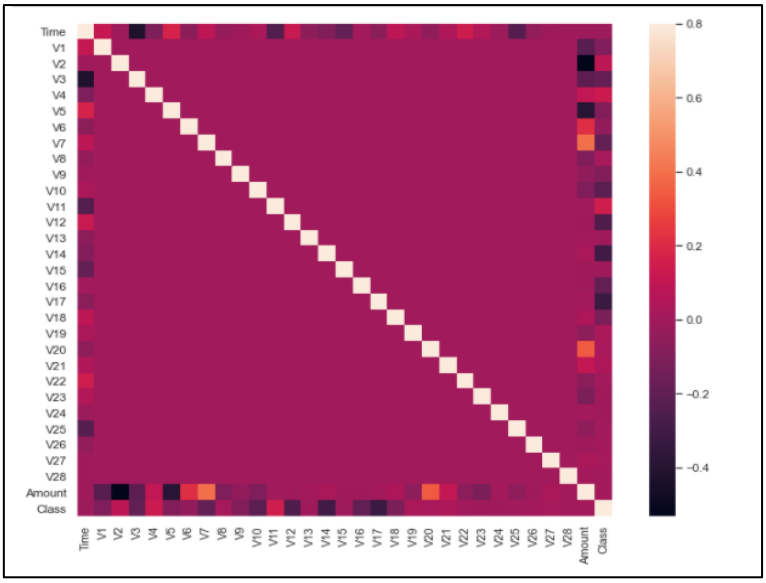

Fig. 4 Correlation Matrix

c. Analyze Target classes:

Examine the number of misrepresentation and non-extortion cases exist in our dataset. We should also calculate the probability of extortion occurrences in the absolute number of exchanges. We find that only 492 misrepresentation cases exist among the examples, representing 0.17 percent of the aggregate. Subsequently, we might express that the information we're working with is for the most part uneven and requires cautious taking care of while modelling and analysing.

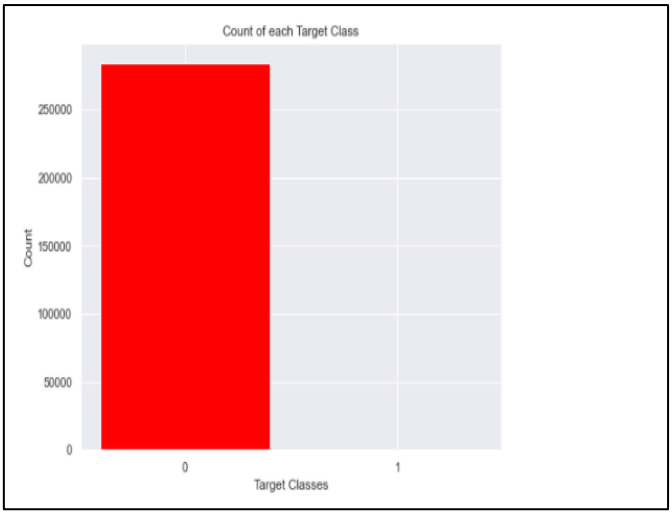

Fig. 5 Target Class

d. Splitting of data:

We will depict the free assortment (X) and conditions during this step (Y). The data is partitioned into readiness set and test set using the variables given, which will be used for exhibiting and testing. Using Python's 'train test split' approach, we can decide the data quickly.

B. Modelling of data:

We will raise two distinct kinds of separation models in this progression: Decision Tree and Logistic Regression. Notwithstanding the way that there are a lot more models that can be utilized, these are the most regularly involved models for order challenges. The insights gave by the scikit-learn package can be utilized to make this huge number of models.

ALGORITHMS

A. Decision Tree Classifier:

The algorithm called Decision Tree is a strategy for pushing toward discrete-regarded objective limits, with the learned limit showed by a decision tree. These statics are regular in proficiency guidance and should be visible in various exercises. The ID3 decision tree is inspected. Root, branch, and leaf are the three hubs of a choice tree. Each inner hub relates to a trait test, each branch to an experimental outcome, and each leaf hub to a class mark. The root node is the tree's most noteworthy node. Choice trees coordinate occasions by arranging them from the root to a leaf hub, which characterizes the grouping of the occurrence. Each branch that plummets from a hub in the tree addresses a test trait test, and every hub in the tree addresses a attribute test.

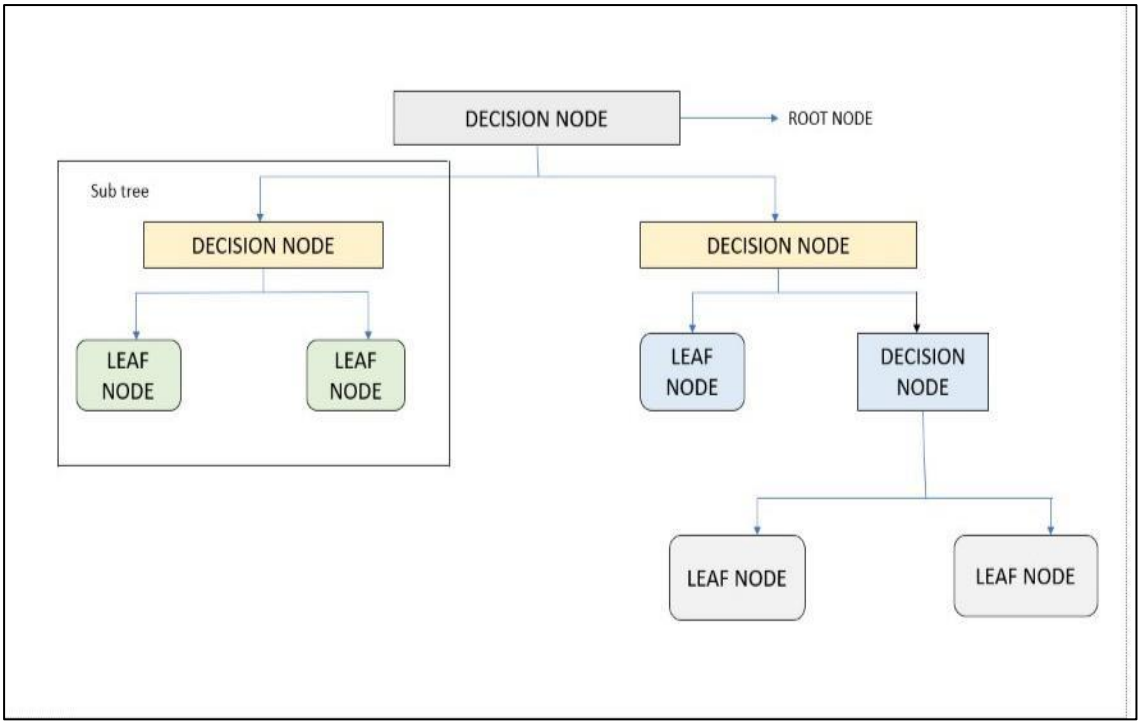

Fig. 6 Decision Tree Classifier

B. Logistic Regression Algorithm:

The algorithm called Logistic Regression is the most notable man-made intelligence computations that is utilized for grouping. Albeit the term 'relapse' shows up in the name, it's anything but a relapse calculation. . Logistic upgrade has your name as it is based on one of the most well known AI algorithm, utilized for retrofitting issue. Everything considered, the estimate is communicated with regards to the likelihood of the result of each stage. The information factors (x) with weights are anticipated by genuine esteemed yields in the linear regression model. To be more precise we are considering a single data 'x' and a variable which is dependent 'y'. The direct backslide of hypothesis is imparted as as equation y=a0+a1*x where a0 is called a tendency term, and a1 is also liable for the adaptability of a solitary data x. These loads are dominated during the planning. For this present circumstance, the value of the hypothesis can be under at least 0 imperative than 1. Vital backslide moreover utilize especially direct condition. Nevertheless, since it should predict the probability of the consequence of having a spot with each class, it uses a sigmoid limit or a determined limit which, to pound the expected real characteristics between the extent of 0 and 1.The figure 7. beneath means the working of logistic regression.Logistic withdrawal rate is found in the Linear Regression measure. The numerical advances calculated regression measurements utilizing the numerical advances gave underneath. the straightline equation is composed as:

$$y = b_0+b_1x_1+b_2x_2+b_3x_3+...+b_nx \quad ------1$$

If we consider Logistic Regression, the middle of 0 and 1 will be y, lets divide the equation 1 by (1-y):

$$y/1-y \text{ ; 0 for y=0,and infinity for y=1} \qquad \text{------2}$$

The middle of negative infinity to the positive infinity is taken as the range and again lets consider the logarithm of the equation ,

$$\log[y/y-1] = b_0+b_1x_1+b_2x_2+b_3x_3+...+b_nx_n \qquad \text{------3}$$

Equation 3 is the last equation for the Logistic Regression algorithm.

4.EVALUATION AND RESULT ANALYSIS:

A. Data:

The data base contains creditcard misrepresentation information from an European credit card firm. The dataset was accumulated utilizing Kaggle. credit card transactions from September 2013 to September 2014 are remembered for the information. Exchanges that were finished inside two days are remembered for the information base. In data gathering, there are 284,807 exchanges, 492 of which are phony. Just 0.172 percent of the whole income is represented by this fake exchange. Datasets are changed over into mathematical factors by means of PCA transformation. This is finished for attentiveness. Utilizing PCA, you might transform information bases into mathematical factors. This is finished to safeguard your security. The 'Time' and 'Values' highlights don't take into account PCA alteration. The distinction in seconds between the current and earlier exchanges is addressed in the 'Time' section. 'Sum' is a sort of item. The monetary exchange that occurred is displayed in this field.

B. Evaluation Criteria:

We really want to examine measures like F1 score,precision,recall and accuracy to analyze various techniques . The confusion matrix plot is included. The lattice address of the confusion matrix is utilized to compute the awareness, accuracy and the rate of error. At that point we will see what is the best suit for finding a credit card misrepresentation.

C.Result Analysis:

For both techniques, the confusion matrix is displayed. The results will be varied if the dataset is used with different algorithms. The dataset is applied to the Decision Tree Classifier which yields the below results:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	85308
1	0.86	0.72	0.78	135
accuracy			1.00	85443
macro avg	0.93	0.86	0.89	85443
weighted avg	1.00	1.00	1.00	85443

Fig. 7. Output for Decision Tree Classifier

The scrutinization measures are described in the figure 7. The review, f1 score and the accuracy for non-misrepresentation occurrences are equivalent to fraud cases, yet vary for fraud cases.

	predicted fraud	predicted normal
is Fraud	97	38
is Normal	16	85292

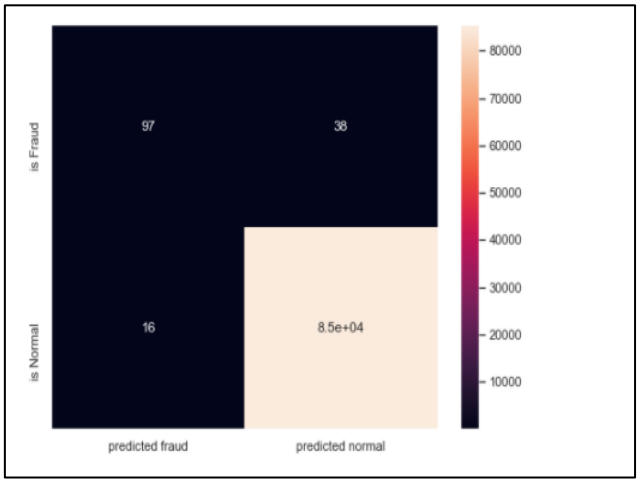

Fig. 8 Decision Tree Classification based Confusion Matrix.

The Database Tree Database is calculated. The result is the same as that of the Logistic Regression algorithm.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	85308
1	0.83	0.56	0.67	135
accuracy			1.00	85443
macro avg	0.92	0.78	0.83	85443
weighted avg	1.00	1.00	1.00	85443

Fig. 9 Output for Logistic Regression

The constraints of evaluation have been explained in Fig9.The f1 score, precision and recall are different for non-fraud cases and is same for fraud cases.

	predicted fraud	predicted normal
is Fraud	75	60
is Normal	15	85293

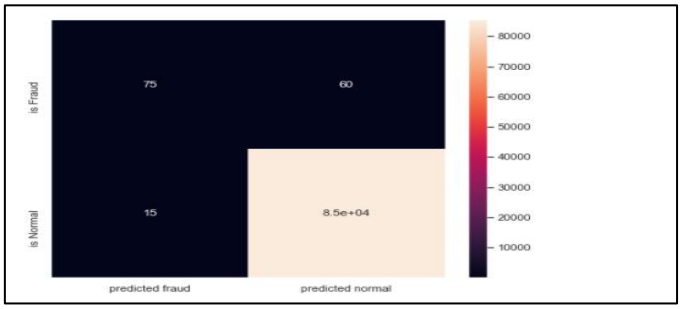

Fig. 10 Confusion Matrix for Logistic Regression

```
Precision of the Decision Tree is
0.8584070796460177
Precision of the Logistic Regression is
0.8444444444444444
Recall of the Decision Tree is
0.7185185185185186
Recall of the Logistic Regression is
0.562962962962963
F1 Score of the Decision Tree is
0.7822580645161291
F1 score of the Logistic Regression is
0.6755555555555556
```

Fig. 11 F1score ,precision and recall of Logistic Regression and Decision Tree Classifier

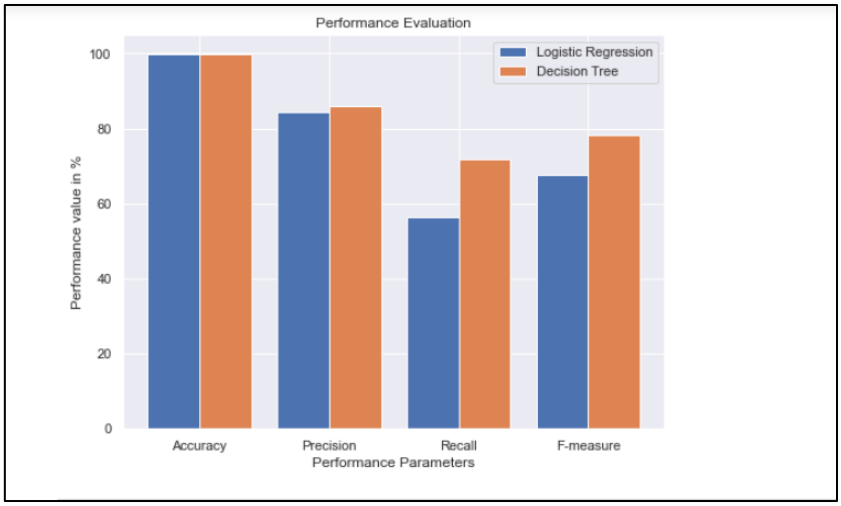

Fig.13 .Performance Evaluation of both the algorithms.

Table I

	Decision Tree Classifier	Logistic Regression
Accuracy	99.94	99.91
Precision	0.8584	0.8444
Recall	0.7185	0.5629
F1 Score	0.7822	0.6755

The above table represents the precision, recall and the f1 score of both the algorithms.

5. CONCLUSION

We have analysed ML applications like Logistic regression and Decision tree category to show that fraud is proven to reduce the number of false alarms. If these algorithmic methods are implemented into bank credit card fraudulent finding system, the prospect of fraudulent. When a transaction is made, the transactions are usually predicted immediately. Anti-fraud measures are being used to block the bank. from big loses and this will also help to reduce the risks of frauds. The performance of the system is scrutinized with the help of the factors like precision, accuracy and support which are proposed in this paper. The two algorithms are compared where Logistic Regression came out to be the best algorithm.

REFERENCES

[1]. Fabiana Fournier, Ivo carriea, Inna skarbovsky, The Uncertain Case of Credit Card Fraud Detection, The ACM International Conference On Distributed Event Based Systems(DEBS15) 2015.

[2]. D. Olszewski, ‘‘Fraud detection using self-organizing map visualizing the user profiles,’’ Knowl.-Based Syst., vol. 70, pp. 324–334, Nov. 2014.

[3]. Dinesh L. Talekar, K. P. Adhiya, Credit Card Fraud Detection System-A Survey, International journal of modern engineering research(IJMER) 2014.

[4]. SamanehSorournejad, Zahra Zojaji, Reza Ebrahimi Atani, Amir Hassan Monadjemi, A Survey of credit card fraud detection techniques: Data and techniques oriented perspective.

[5]. Lakshmi S V S S, Selvani Deepthi Kavila, Machine learning for credit card fraud detection system, International Journal Of Applied Engineering Research ISSN 2018.

[6] R. Bolton and D. Hand, ‘‘Statistical fraud detection: A review,’’ Stat. Sci., vol. 17, no. 3, pp. 235–249, Aug. 2002.

[7] P. A. Dal, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, ‘‘Credit card fraud detection: A realistic modeling and a novel learning strategy,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 8, pp. 3784–3797, Sep. 2017.

[8] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, ‘‘Data mining for credit card fraud: A comparative study,’’ Decis. Support Syst., vol. 50, no. 3, pp. 602–613, Feb. 2011.

[9] N. Sethi and A. Gera, ‘‘A revived survey of various credit card fraud detection techniques,’’ Int. J. Comput. Sci. Mobile Comput., vol. 3, no. 4, pp. 780–791, Apr. 2014.

[10] A. O. Adewumi and A. A. Akinyelu, ‘‘A survey of machine-learning and nature-inspired based credit card fraud detection techniques,’’ Int. J. Syst. Assurance Eng. Manage., vol. 8, no. S2, pp. 937–953, Nov. 2017.

[11] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, ‘‘Credit card fraud detection using machine learning techniques: A comparative analysis,’’ in Proc. ICCNI, Lagos, Nigeria, Oct. 2017, pp. 1–9.

[12] M. Carminati, R. Caron, F. Maggi, I. Epifani, and S. Zanero, ‘‘BankSealer: A decision support system for online banking fraud analysis and investigation,’’ Comput. Secur., vol. 53, no. 1, pp. 175–86, Sep. 2015.

[13] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, ‘‘Detection of financial statement fraud and feature selection using data mining techniques,’’ Decis. Support Syst., vol. 50, no. 2, pp. 491–500, Jan. 2011.

[14] E. Kirkos, C. Spathis, and Y. Manolopoulos, ‘‘Data mining techniques for the detection of fraudulent financial statements,’’ Expert Syst. Appl., vol. 32, no. 4, pp. 995–1003, May 2007.