As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 8(05):325-330, 2022
Copyright © 2022 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0805047
Available online at: http://www.ijmtst.com/vol8issue05.html

Credit Card Fraud Detection Critical Analysis

P.Nagaraj [1], Chenna Sriya[2], Sarikonda Mahidhar Raju[2], Saba Naazneen Kauser[2], Chintala Rakesh[2]

[1]Associate Professor, Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad
[2]Department of CSE,Sreyas Institute of Engineering and Technology .Hyderabad

To Cite this Article
P.Nagaraj, Chenna Sriya, Sarikonda Mahidhar Raju, Saba Naazneen Kauser and Chintala Rakesh. Credit Card Fraud Detection Critical Analysis. International Journal for Modern Trends in Science and Technology 2022, 8(05), pp. 325-330. https://doi.org/10.46501/IJMTST0805047

Article Info
Received: 07 April 2022; Accepted: 12 May 2022; Published: 15 May 2022.

ABSTRACT

Fraud is one of the principal ethical problems with inside the credit card industry. The essential factors are right off the bat, to distinguish the distinctive kinds of Credit card smart, and to survey discretionary strategies that have been utilized in fraud recognition. Credit card organizations must be able to once the fraudulent credit card transactions so that the customers are not charged for items that they did not merchandise. The datasets consist of exchanges made by charge cards in September 2013 by European cardholders. This dataset presents exchanges that happened in the span of 2 days, where there are 492 frauds out of 284,807 exchanges. The dataset is profoundly un equal, the positive class (frauds) represent 0.172% of all things in the considerations. It contains just numerical input features which are the consequence of a PCA Transformation. Lamentably, because of secrecy issues, they can't give the original data and more background information about the data. Features V1, V2, V28 are the foremost parts gotten with PCA. The primary capabilities that have now no longer been modified with PCA are 'Time' and 'Amount'. Feature 'Time' carries the seconds elapsed among each transaction and the number one transaction with inside the dataset.

Keywords- Fraudulent, PCA transformation, Logistic regression, decision tree, random forest

1. INTRODUCTION

Although incidences of credit card fraud are confined to 0.1% of all card transactions, they have led to massive economical losses as the fraudulent transactions have been large value transactions. In 1999, out of 12 billion transactions made annually, approximately 10 million—or one out of every 1200 transactions—turned out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts was fraudulent. Even with extreme volume and value increase in credit card transactions since then, these proportions have stayed the same or have decreased due to sophisticated fraud detection and prevention systems. Today's fraud detection systems are designed to prevent one-twelfth of one percent of all transactions processed which still translates into billions of dollars in losses.[3] In the decade to 2008, general credit card losses have been 7 basis points or lower (i.e. losses of $0.07 or less per $100 of transactions). In 2007, fraud in the United Kingdom was estimated at £535million. The essential factors are right off the bat, to distinguish the distinctive kinds of Credit card smart, and to survey elective strategies that have been utilized in fraud recognition.

2. EXISTING SYSTEM

For quite a while, there has been a solid enthusiasm for the morals of banking (Molyneux, 2007; George, 1992), just as the ethical intricacy of fake conduct

(Clarke, 1994). Extortion implies getting administrations/products and additionally cash by deceptive methods and is a developing issue everywhere throughout the world these days. Extortion manages cases including criminal purposes that, generally, are hard to recognize. Charge cards are a standout amongst the best-known focuses of extortion however by all accounts not the only ones; misrepresentation can happen with acknowledge editems, for example, individual advances, home credits, and retail. Besides, the essence of extortion has changed significantly amid the most recent couple of decades as advancements have changed and created. A basic assignment to support organizations and money-related establishments including banks is to find away to anticipate extortion and to manage it productively and adequately when it happens (Anderson, 2007). Anderson (2007) has identified and explained the different types of fraud, which are as many and varied as the financial institution's products and technologies.

DISADVANTAGES

Fraud detection and prevention software that analyzes patterns of normal and unusual behavior as well as individual transactions to flag likely fraud. Profiles include such information as IP addresses. Technologies have existed since the early 1990s to detect potential fraud. One early market entrant was Falcon; other leading software solutions for card fraud include actimize, SAS, BAE Systems Detica, and IBM. All these do not utilize modern machine learning techniques which are far more efficient and more.

3. PROPOSED SYSTEM

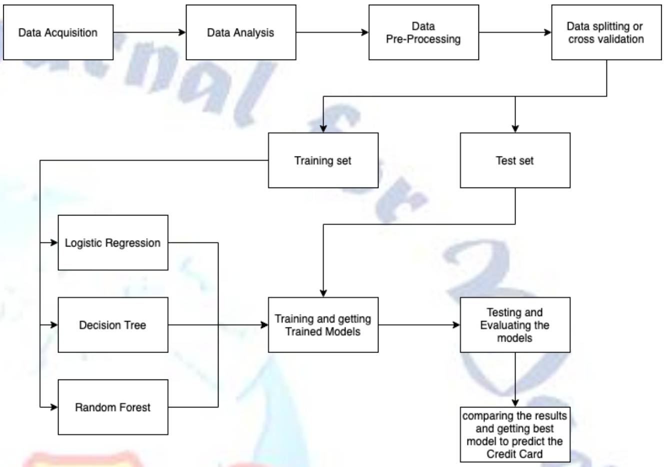
The main aims are, firstly, to identify credit card fraud and, secondly, to review alternative techniques that have been used in fraud. Here we compare two different algorithms each of different types of machine learning, one of supervised and another of unsupervised.

ADVANTAGES

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, an e-commerce website like Amazon serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the result store real relevant advertisements to them.

Explanation:

The proposed system can be explained by the below diagram. The below diagram is used for understanding the architecture of the research done in the project, it's also understood as the modules used in the project.


Data Acquisition
Data Analysis
Data Pre-Processing
Data splitting or cross validation
Training set
Test set
Logistic Regression
Decision Tree
Random Forest
Training and getting Trained Models
Testing and Evaluating the models
comparing the results and getting best model to predict the Credit Card

Data Acquisition: converting the normal data like CSV files etc. into python understandable data such as Nd-array object of NumPy or Data-Frame object of pandas.

Data Analysis: understanding the basics of the data loaded. To know several columns, several rows, their statistics. So that we can perform Data pre-processing step.

Data pre-processing: cleaning the data like removing the empty values, removing insignificant columns or balancing the data, preparing the data for giving it as input to the algorithm, etc.

Data Splitting: creating the training set and testing set so that we can train the algorithm and understand the performance of that model.

Training: Making the particular algorithm understand the train data and become intelligent in that concept.

Testing: the process used to predict the outputs for the inputs in the test set.

Comparing the metrics: the process used to measure the performance of all the algorithms and to obtain a conclusion.

DECISION TREE:

Classification is a two-step process, learning step and pre- diction step, in machine learning. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response forgiven data. A decision tree is one of the easiest and most popular classification algorithms to understand and interpret.

DECISION TREE ALGORITHM

The Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record's attribute. Based on the comparison, we follow the branch corresponding to that value and jump to the next node.

TYPES OF DECISION TREES

Types of decision trees are based on the type of target variable we have. It can be of two types:

Categorical Variable Decision Tree: A decision tree that has a categorical target variable is then called a **Categorical variable decision tree.**

Continuous Variable Decision Tree: A decision Tree has a continuous target variable then it is called a **Continuous Variable Decision Tree.**

Important Terminology related to Decision Trees

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets. **Splitting:** It is a process of dividing a node into two or more sub-nodes.

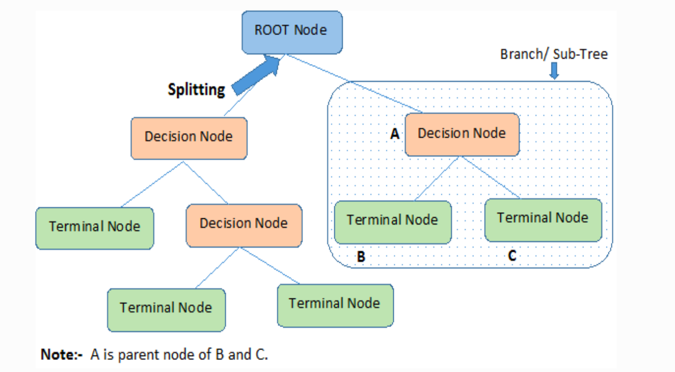
Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

Leaf / Terminal Node: Nodes that do not split is called Leaf or Terminal node. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch/Sub-Tree: A subsection of the entire tree is called a branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

Decision trees classify the examples by sorting them down he tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive.



Note:- A is parent node of B and C.

Fig: Decision Tree

Logistic Regression:

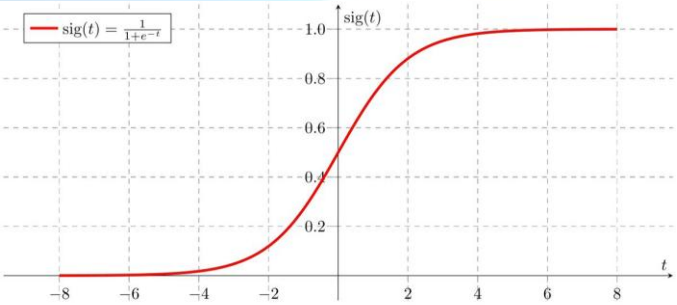
Logistic regression is used in the biological sciences in the early twentieth century. It was the nusedinm any social science applications. Logistic Regression is used when the dependent variable (target) is categorical.

For example,

To predict whether an email is a spam (1) or (0), whether the tumorism alignant (1) or not (0)

Consider a scenario where we need to classify whether an email is a spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequences in real-time.

Sigmoid function:



From this example, it can be inferred that linear regression is not suitable for classification problems. Linear regression is unbounded, and this brings logistic regression into the picture. Their value strictly ranges from 0to1.

RANDOM FOREST:

Data science provides a plethora of classification algorithms which includes logistic regression, support vector machine, naïve bayes classifier, and decision trees. But close to the top of the classifier hierarchy is the random forest classifier (there is also the random forest repressors but that is a topic for another day).In this post, we will examine how basic decision trees work, how individual decisions trees are combined to make a random. Forest, and ultimately discover why random forests are so good at what they do.

Let's quickly go over decision trees as they are the building blocks of the random forest model. Fortunately, they are pretty intuitive. I'd be willing to bet that most people have used a decision tree, knowingly or not, at some point in their lives.

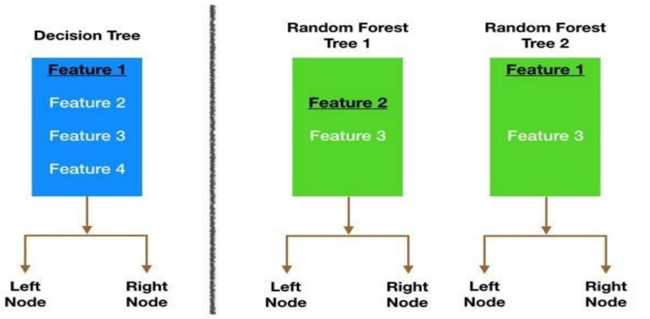
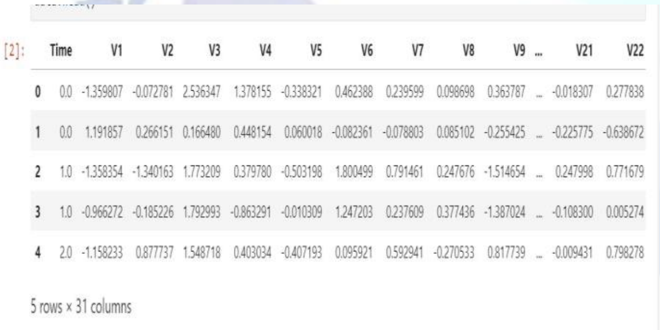


Fig: Random Forest

4. RESULTS

CSV file:

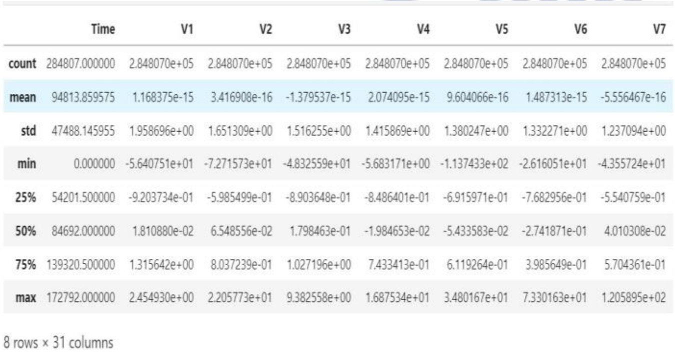


[2]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278

5 rows × 31 columns

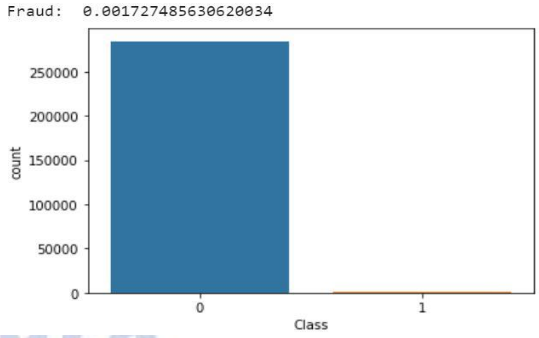
Describe:



	Time	V1	V2	V3	V4	V5	V6	V7
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05
mean	94813.859575	1.168375e-15	3.416908e-16	-1.379537e-15	2.074095e-15	9.604066e-16	1.487313e-15	-5.556467e-16
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00	1.237094e+00
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01	-4.355724e+01
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01	-5.540759e-01
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01	4.010308e-02
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01	5.704361e-01
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01	1.205895e+02

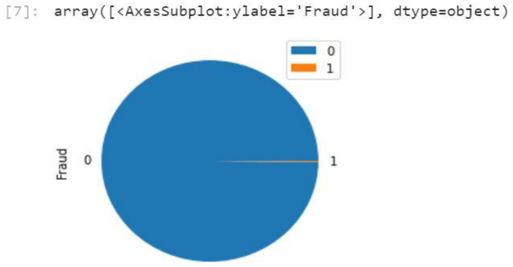
8 rows × 31 columns

Fraud by Count-plot graph



```
Fraud:  0.001727485630620034
```

Data-Frame

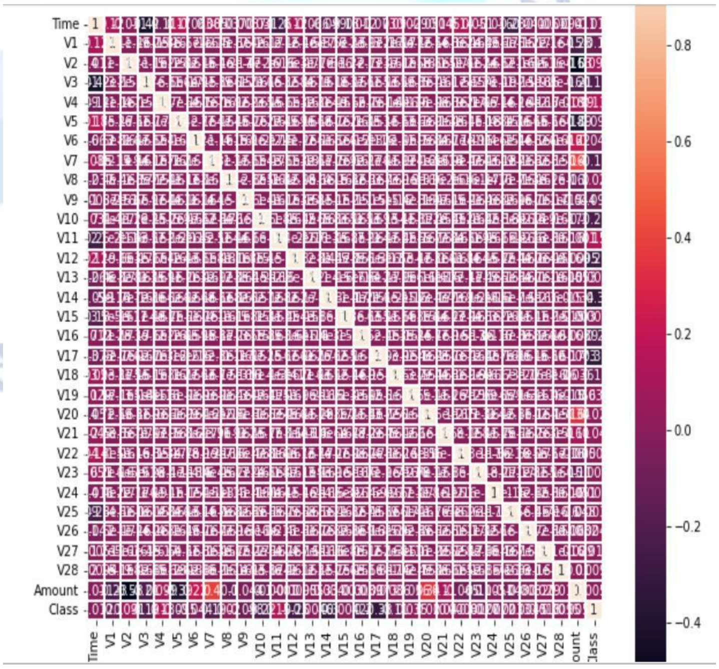


```
[7]:  array([<AxesSubplot:ylabel='Fraud'>], dtype=object)
```

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
```

Correlation Matrix:



By logistic Regression:

The testing of all the trained models provided us with insights into the performance of all three algorithms namely logistic regression, Decision Tree, and Random Forest. We used accuracy score, confusion matrix, and classification report to evaluate the three algorithms. With all the above results we can say the Random Forest algorithm works very well for the prediction of credit card fraud transactions

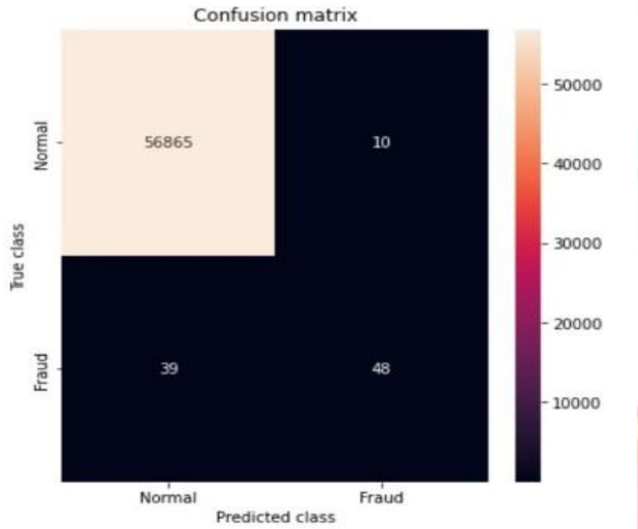


Fig: Logistic Regression

Accuracy: 0.9495435553526912

Precision: 0.9295774647887324

Recall: 0.7586206896551724

F1-Score: 0.8354430379746836

AUC score: 0.879266 3887836302

By decision trees:

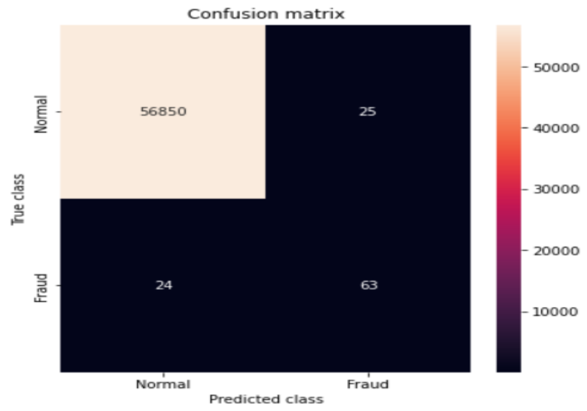


Fig: Decision Tree

Accuracy: 0.9991397773954567

Precision: 0.7159090909090909

Recall: 0.7241379310344828

F1-Score: 0.72

AUC score: 0.8618491852974611

By Random-forest Algorithm:

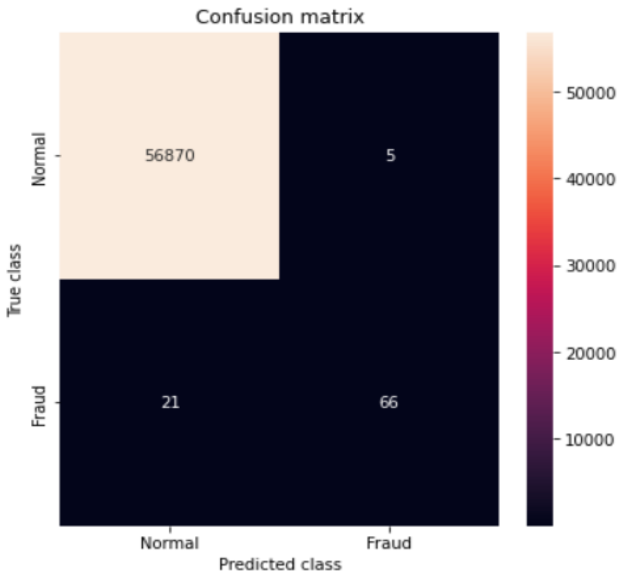


Fig: Random Forest

Accuracy: 0.9459459459459459

Precision: 0.9509803921568627

Recall: 0.8981481481481481

F1 Score: 0.9238095238095237

AUC SCORE: 0.9012548568468455

5. CONCLUSION:

The testing of the all the trained models provided us with insights of the performance of all the three algorithms namely logistic regression, Decision Tree, and Random Forest. The testing here on the data was done by supervised learning produced 99.9% of accuracy for the data set of 2.84+lakhs which was much better than the existing system. The confusion matrices generated here given us the detailed number of transactions which are normal and fraud transactions. We did correlate all the labels or columns heads of the data sets eg., encoders like V1,V2,…V28 by which we can also get the correlation of these labels which can help us to get the data set in this format itself or changes to be done in these. We used accuracy score, confusion matrix, and classification report to evaluate the three algorithms. With all the above results we can say the Random Forest algorithm works very well for prediction of credit card fraud transactions.

Future Scope:

The very nature of this project allows for multiple algorithms to be integrated together as modules and their results can be combined to increase the accuracy of the final result. This model can further be improved with the addition of more algorithms into it. However, the output of these algorithms needs to be in the same format as the others. Once that condition is satisfied, the modules are easy to add as done in the code. This provides a great degree of modularity and versatility to the project. The drawback while doing this project is that we can't determine the names of fraud and normal transactions from the given dataset.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[2] https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html

[3] Gupta, Shalini, and R. Johari". A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant." International Conference on Communication Systems and Network Technologies IEEE, 2011:22-26.

[4] Y. Gmbh and K. G. Co, "Global online payment methods: the Full year 2016, Tech.Rep" 3 2016.

[5] Bolton, Richard J., and J. H. David." Unsupervised Profiling Methods for FraudReal-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721-1732.

[6] Drummond, C., and Holte, R. C. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling. Proc of the ICML Workshop on Learning from Imbalanced Datasets II, 1–8. Quah, J. T. S., and Sri Ganesh, M. (2008).

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

[8] P.NAGARAJ, Dr.A.V.Krishna Prasad, "Survey on Swine flu Prediction", InternationalJournal of Management, Technology and Engineering, Volume IX, Issue V,May/ 2019 , ISSN No:2249-7455, Page no: 937-941.

[9] P.Nagaraj , Rajesh Banala and A.V.Krishna Prasad, "Real Time Face Recognition using Effective Supervised Machine Learning Algorithms", Journal of Physics: Conference Series 1998 (2021) 012007 IOP Publishing doi:10.1088/1742-6596/1998/1/012007

[10] P. Nagaraj and Dr A. V. Krishna Prasad, "A Novel Technique to Predict the Hotspots Swine Flu Effected Regions", THINK INDIA JOURNAL, ISSN:0971-1260, Vol-22- Issue-41-December-2019

[11] P. Nagaraj and Dr A. V. Krishna Prasad, "A Novel Technique to Detect the Hotspots Swine Flu Effected Regions", Published in: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 15 November 2021 DOI:10.1109/ICRITO51393.2021.9596422, Electronic ISBN: 978-1-6654-1703-7 CD: 978-1-6654-1702-0

[12] Nagaraj P, Dr A.V. Krishna Prasad, "A Cloud Computing Emerging Security Threats and Its Novel Trends in Knowledge Management Perception", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Special Issue 2, December 2017)

[13] Nagaraj P,Rohit Kumar K, Rajesh Banala b," Energy efficient 2 tier data aggregation scheme in-Sensor networks", Accepted 7 March 2021,https://doi.org/10.1016/j.matpr.2021.03.140 2214-7853/ 2021 Elsevier Ltd.scientific committee of the Emerging Trends in Materials Science,Technology and Engineering.

[14] P. Nagaraj, Gunta Sherly Phebe, Anupam Singh, "A Novel Technique to Classify Face Mask for Human Safety", 2021 Sixth ICIIP Published in :2021 Sixth International Conference onImage Information Processing (ICIIP),26-28 Nov. 2021, 10 February 2022 DOI: 10.1109/ICIIP53038.2021.9702607 Publisher: IEEE Conference Location: Shimla, India

[15] P.Nagaraj, Sahith Krishna Palla, Shiva Sai Putnala, Sampath Kumar Parvatham," SYSTEM FOR RECORDING ATTENDANCE USING PYTHON" published in "Journal of Information and Computational Science", ISSN: 1548-774, Volume 12 Issue 4 – 2022, www.joics.org.