e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM

Bhargav M[*1], Gaurav G[*2], Harsith S[*3], Suhail Ahmed Sayyed[*4],

Mrs. Belji T[*5]

[*1,2,3,4]Department Of Computer Science And Engineering, K.S. School Of Engineering
And Management, Bengaluru, Karnataka, India.

[*5]Asst. Prof., Department Of Computer Science And Engineering, K.S. School Of Engineering
And Management, Bengaluru, Karnataka, India.

ABSTRACT

Every year fraud cost generated in the economy is more than $4 trillion internationally. Financial institutions such as commercial and investment banking operations are increasingly bring targeted. Users can use credit card as it provides an efficient and it is easy to use. Due to the increase of usage of credit cards ,the credit card misuse has been enhanced. Fraud detection means a collection of activities to avoid collecting money by misleading pretensions. The main aim to detect such frauds, including the accessibility of public data, the changes in the fraud nature and high rates of false alarm. A machine learning algorithm was first applied to dataset, which improve the accuracy of the detection of frauds to some extent. A number of sectors are today using fraud detection which includes ecommerce and banking agencies. The mode of payment has moved from cash to digital settlements such as debit/credit card, online wallet payment, online banking etc. As the result financial fraud is increasing at rapid rate for personal gain. The algorithms used are K Nearest Neighbors, Random Forest, Decision Tree and Logistic Regression.

Keywords: KNN, Random Forest, Decision Tree, Random Forest, Logistic Regression.

I. INTRODUCTION

Fraud is defined as a wrongful or criminal deception which is aimed to bring financial or personal gain. Two mechanisms are used to avoid fraud and losses due to fraud. They are Fraud Prevention and Fraud Detection. Fraud Prevention is a proactive method where it stops fraud from being happening. Fraud Detection is used when a fraudulent transaction is attempted by the fraudster.

Users can use credit card as it provides an efficient and it is easy to use. Due to the increase of usage of credit cards ,the s credit card misuse has been enhanced. Fraud detection means a collection of activities to avoid collecting money by misleading pretensions. The main aim to detect such frauds, including the accessability of public data, the changes in the fraud nature and high rates of false alarm.

A machine learning algorithm was first applied to dataset, which improve the accuracy of the detection of frauds to some extent. A number of sectors are today using fraud detection which includes ecommerce and banking agencies. The mode of payment has moved from cash to digital settlements such as debit/credit card, online wallet payment, online banking etc. As the result financial fraud is increasing at rapid rate for personal gain. The algorithms used are KNN, Random Forest, Decision Tree and Logistic Regression.

With the growing prevalence of electronic payment systems, credit card fraud has become a significant concern for both merchants and consumers. Machine learning techniques can be applied to analyze large volumes of transaction data and identify patterns that indicate fraudulent behavior, such as unusual spending patterns, geographic anomalies, and other factors that can be indicative of fraud. By detecting fraudulent transactions in real-time, credit card companies can prevent financial losses and minimize the impact of fraud on consumers. The use of machine learning algorithms can improve the accuracy and speed of fraud detection.

The scope of credit card fraud detection using machine learning is quite broad and covers various aspects of electronic payment systems. Here are some of the key areas where machine learning techniques can be applied to prevent credit card fraud Firstly transaction monitoring where machine learning algorithms can analyze large volumes of transaction data in real-time and detect fraudulent activities. Followed by Risk assessment where machine learning models can be trained to assess the risk associated with specific transactions, users, or merchants. This can help credit card companies to prioritize their fraud prevention. It is also used in user behavior analysis where machine learning algorithms can analyze user behavior patterns and detect suspicious activities such as login attempts from unusual locations, changes in spending patterns, or multiple account registrations with the same device. Another key application is Fraud trend analysis where machine learning models can be trained to identify emerging fraud patterns and trends. This can help credit card companies to proactively prevent future fraud attempts and minimize losses.

II. METHODOLOGY

The flowchart above shows the flow of our application. The different classification techniques we have applied in this study for fraud detection purposes are logistic regression, decision tree, random forest & KNN. Their performances are compared to see which model can better extract the relationship between the features and detect fraudulent transactions. After training all the classifiers, a new ensemble model will be applied as a voting classifier to combine all the other classification techniques. The objective is to reduce the errors of single models, which helps the ensemble model make better predictions compared with the individual classifiers. If all the classifiers are considered as $C1$, $C2$,$C3$, $C4$, $andC5$, then the final classifier will take the votes as the majority of votes as the final prediction or Ct . $Ct = Majority\{C1, C2, C3, C4 , C5\}$.The next steps are Data Collection where in credit card transactions are recorded and stored in a database. The data includes information such as the transaction amount, the date and time of the transaction, the merchant ID, and the cardholder information. Data pre-processing where the raw data is processed and prepared for analysis. This may involve removing irrelevant or duplicate data, converting data into a common format, and standardizing data fields .Data analysis. Statistical and machine learning techniques are applied to the pre-processed data to identify fraudulent transactions. These techniques may include anomaly detection, clustering, and classification algorithms. Model building is where a model is developed based on the data analysis results. The model is trained to identify patterns of fraudulent activity and can be used to detect future fraudulent transactions. Model validation where in the model is tested on a separate set of data to ensure its accuracy and effectiveness.. The next step is Deployment where the validated model is deployed in a production environment to monitor credit card transactions in real time. The final steps are Monitoring and updates is the deployed model is continuously monitored to ensure it remains effective.

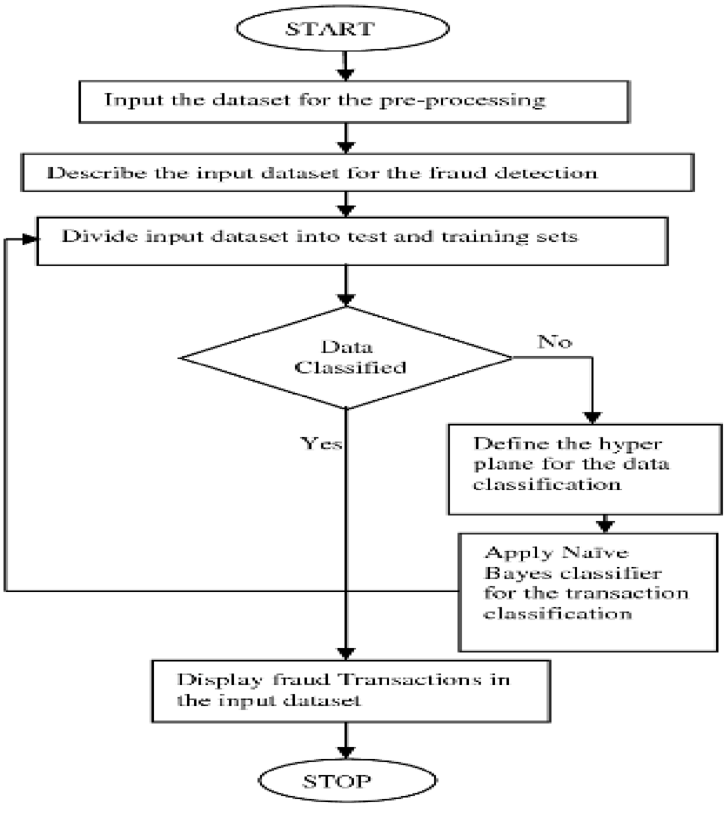
START
Input the dataset for the pre-processing
Describe the input dataset for the fraud detection
Divide input dataset into test and training sets
Data Classified
No
Yes
Define the hyper plane for the data classification
Apply Naïve Bayes classifier for the transaction classification
Display fraud Transactions in the input dataset
STOP

Fig 1: Proposed System

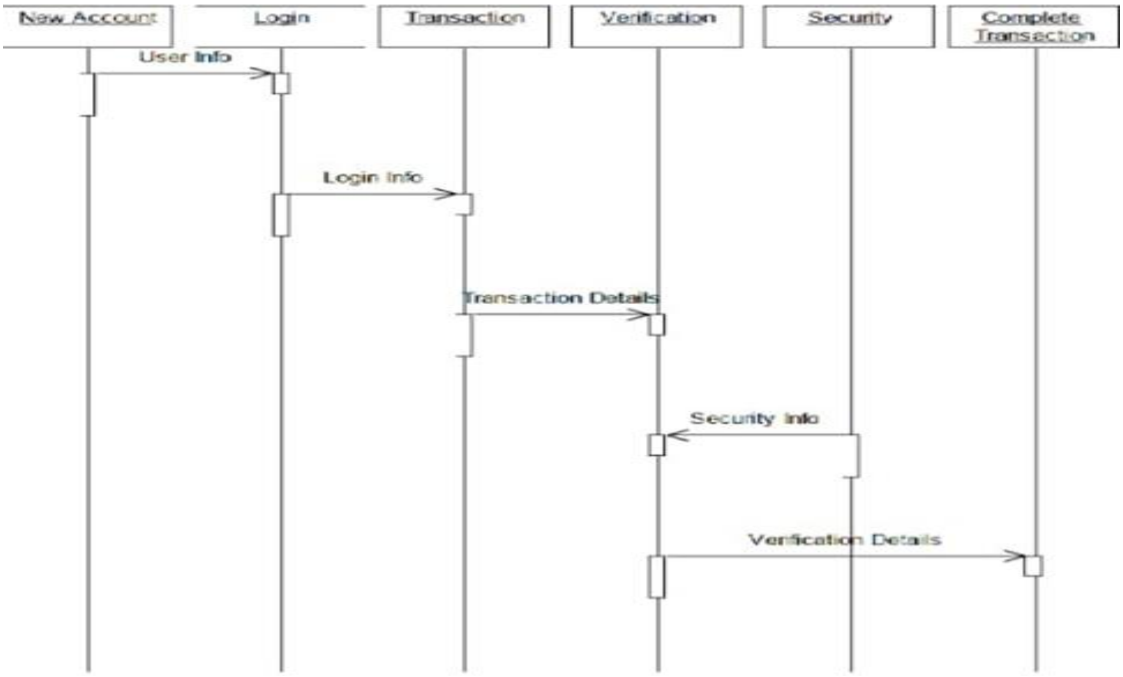

Fig 2: Sequence Diagram

A sequence diagram is a type of UML diagram that illustrates the interactions between objects or components in a software system. In the context of credit card fraud detection, a sequence diagram can be used to visualize the flow of information and activities between the different components involved in the fraud detection process. For example, the diagram might show how the credit card information is captured and validated, how it is compared to known fraud patterns, and how alerts are generated and sent to relevant parties. By providing a clear visual representation of the system's behavior, a sequence diagram can help developers and stakeholders better understand the fraud detection process and identify areas for improvement.

In addition to helping developers and stakeholders understand the system's behavior, a sequence diagram can also be used to document the system's design and architecture. By mapping out the interactions between components, the diagram can serve as a blueprint for future development and maintenance. For example, if a new component is added to the system, the sequence diagram can be updated to show how it interacts with the existing components. Similarly, if a component is modified or removed, the diagram can be used to ensure that the change does not impact the system's overall behavior. Overall, a sequence diagram is a valuable tool for both understanding and documenting the behavior of a credit card fraud detection system. A sequence diagram is a powerful tool for visualizing the interactions between components in a software system, especially in the context of credit card fraud detection. The diagram can help developers and stakeholders understand how the system works by showing how different components interact with each other. For example, a typical sequence diagram for credit card fraud detection might show how a credit card transaction is captured by a payment gateway, how it is passed to a fraud detection engine, and how alerts are generated and sent to relevant parties such as the bank and the cardholder. By using a sequence diagram, developers can easily identify areas of the system that need improvement or optimization, and stakeholders can better understand how the system works and what steps are involved in detecting and preventing fraud.

III. INCORPORATED PACKAGES

Python

Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems.

Jupyter Notebook

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access accessed through the internet.

Machine Learning

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm.

Logistic regression

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type. The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one.

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

IV. MODELING AND ANALYSIS

The figure below shows the block diagram of our application. In this credit card fraud detection system project.

We analyze the given dataset and apply various classification techniques like KNN, Random Forest Model to classify the values as fraudulent or not. We plot a graph and correlation matrix based on this data. We also compare the various machine learning algorithms to find out which is the best classification technique. The correlation matrix graphically gives us an idea of how features correlate with each other and can help us predict what are the features that are most relevant for the prediction.A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) produced by the model on the test data.

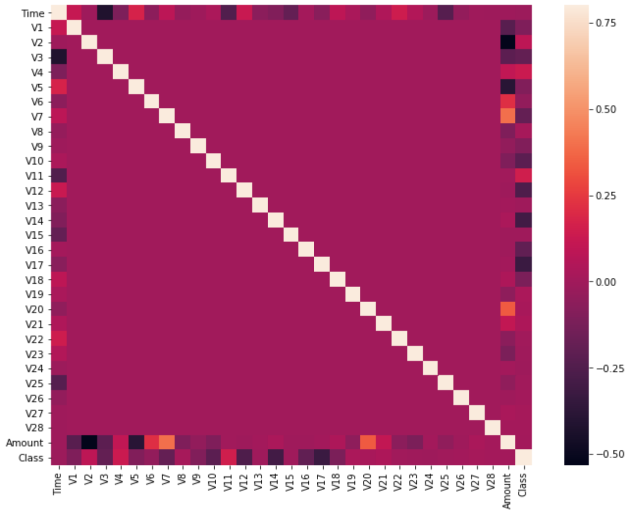

Fig 3: Correlation Matrix

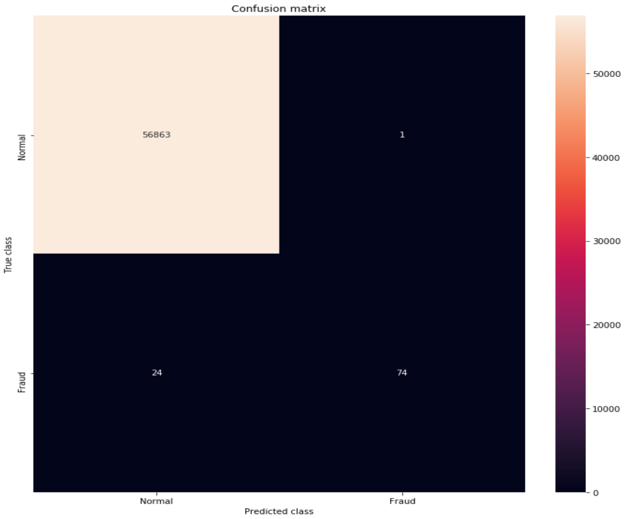

Fig 4: Confusion Matrix

V. RESULTS AND DISCUSSION

In this part, we show the classified result from two prediction models. We used different parameters for make comparison with different models; the parameters i.e. Accuracy, Precision and Recall.

Precision: It is the estimation analysis of true positive to the aggregate value of true positive and false positive rate.

Precision = (TP)/(TP+FP)

Recall: It is the estimation analysis of true positive rate to the aggregate value of the true positive and false negative rate.

Recall = (TP)/(TP+FN)

The experiments were carried out in two folds. In the first step, a classification process was conducted using F={v1,v2,v3,v4,v5}.For each feature vector in F, the following methods were trained and tested: RF,DT,ANN,NB and LR. The results are depicted. As an initial validation of the proposed method, were an further experiments using the full feature vector and a feature vector that was generated using a random approach random_ vector = {V2, V3, V4, V5, V6, V7, V8, V9, V11, V12, V13, V16, V17,V18,V19,V20,V21,V22,V23,V25,V26,V28,Amount}. Furthermore, we computed the AUC of each vector in F. These results are depicted. The best performing models in terms of the quality of classification are the RF, NB, and LR with the AUCs of 0.96, 0.97, and 0.97, respectively. In the instance of v5 , the RF and NB obtained the highest AUCs of 0.95 and0.96. Moreover, a comparison analysis is presented in Table 7. This comparison reveals that the GA feature selection approach presented in this paper as well as most of the proposed ML methods that were implemented outperformed the existing techniques that are proposed.

Credit card fraud detection is a critical task for financial institutions and merchants. With the increasing volume of credit card transactions and the growing sophistication of fraudsters, it is becoming more challenging to detect fraudulent transactions accurately. Therefore, various algorithms have been developed to detect and prevent credit card fraud. In this article, we will compare some of the most commonly used algorithms for credit card fraud detection.

Rule-based algorithms: Rule-based algorithms are simple and easy to implement, making them a popular choice for small-scale fraud detection. They are based on defining a set of rules for identifying fraudulent transactions based on certain patterns or criteria. While rule-based algorithms can be quick and effective, they may not be able to capture more complex patterns of fraud and may generate false positives.

Clustering algorithms: Clustering algorithms involve grouping transactions into clusters based on their similarity, and then using various metrics to identify anomalous clusters that may contain fraudulent transactions. Clustering algorithms can be effective for identifying unusual patterns of activity, but may not work as well for detecting low-level fraud. They may also struggle with large datasets or noisy data.

Decision trees: Decision trees are a popular machine learning algorithm for credit card fraud detection that involves breaking down a problem into smaller sub-problems and then recursively building a decision tree to predict the outcome of each sub-problem. Decision trees can be effective for detecting complex patterns of fraud, but can be sensitive to overfitting and may require a large amount of training data.

K-Nearest Neighbor (KNN): KNN is a simple and easy-to-understand algorithm that is based on the concept of finding the K closest data points to a new point and using the class of those points to predict the class of the new point. KNN is effective when there is a clear separation between classes and can work well in detecting fraud in small datasets. However, KNN may struggle with larger datasets, and its performance can be affected by noisy data.

Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs to make a final decision. Random Forest is more complex than KNN and can handle larger datasets with noisy data. It can also handle missing values and non-linear relationships between features. Random Forest can be effective in detecting fraud, but it may be difficult to interpret the decision-making process due to its complexity.

In conclusion, there is no single algorithm that is the most effective for credit card fraud detection. The choice of algorithm depends on the specific problem and dataset at hand, and it is recommended to experiment with different algorithms and choose the one that provides the best results for the given problem.

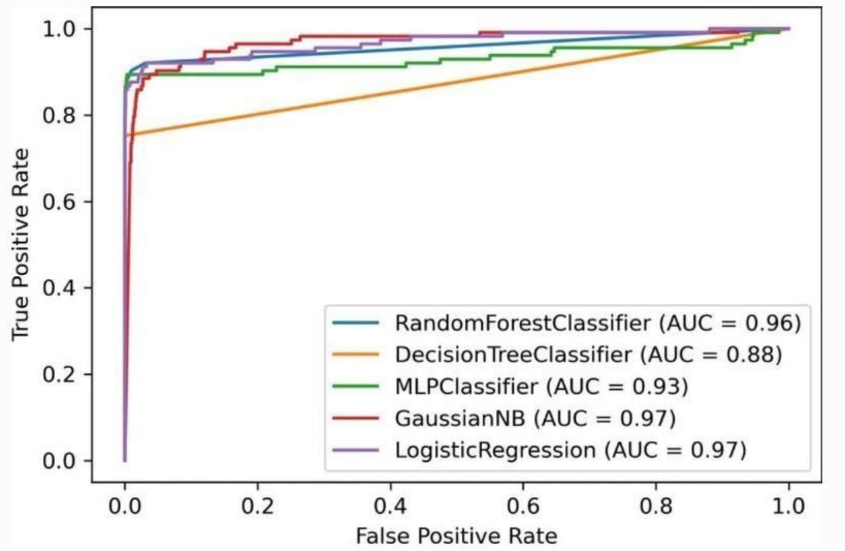

Fig 5: Comparision of algorithms

VI. CONCLUSION

In conclusion, credit card fraud detection is a critical process for financial institutions to protect their customers and prevent financial losses. By implementing a robust fraud detection system, financial institutions can identify and prevent fraudulent transactions in real-time. This involves collecting and processing transaction data, analyzing the data using statistical and machine learning techniques, building and validating a fraud detection model, and continuously monitoring and updating the system. With the growing sophistication of fraudsters and the increasing number of online transactions, credit card fraud detection systems must remain vigilant and adapt to new fraud patterns to provide effective protection for customers.

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions. Since the entire dataset consists of only two days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency overtime as more data is put in to it.

VII. REFERENCES

[1] Research Article Fraud Miner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset by K.R.Seeja and Masoumeh Zarea poor.

[2] D. Sanchez, M.A. Villa, L. Cerda & J.M. Serrano," Association rules applied to credit card fraud detection", Expert Systems with Applications, vol.36,no.2,pp.3630-3640.

[3] Baesens, B., Höppner, S., Ortner, I., Verdonck,T.,2021a.robROSE:arobustapproachfordealing with imbalanced data in fraud detection.Stat. Methods Appl. doi:10. 1007/s10260-021-00573.

[4] Baesens, B., Vlasselaer, V.Van., Verbeke, W, 2015. Fraud Analytics Using Descriptive. Predictive. and Social Network Techniques: A Guide to Data Science For Fraud Detection. John Wiley \& Sons .Inc, Hoboken. NJ. USAdoi: 10.1002/9781119146841.

[5] Carrasco,R.S.M.,Sicilia-Urbán,M.A.,2020.Evaluation of deep neural networks for reduction of credit card fraud alerts. IEEEAccess8,186421186432.doi:10.1109/ACCESS.2020.3026222

[6] Kaminski, ´ B., Jakub czyk, M., Szufel, P., 2018.A framework for sensitivity analysis of decision trees. Cent. Eur. J. Oper.Res.26,135–159.doi:10.1007/s10100-017-0479-6.

[7] Sadgali, Imane, Sael, Nawal, Benabbou, Faouzia, 2021. Human behaviors coring in credit card fraud detection. IAES Int. J. Artif. Intell. 10,698. doi:10.11591/ijai.v10.i3.pp698-706,IJ-AI.

[8] Sun, J., Li, Y., Chen, C., Lee, J., Liu, X., Zhang,Z., Xu, W., 2020. FD Helper: assist unsuper vised fraud detection experts with interactive feature selection and evaluation. In: Paper presented at the Conference on Human Factors in Computing Systems Proceedings doi:10.1145/3313831.3376140.

[9] Trisanto, D., Rismawati, N., Mulya, M., Kurniadi, F.,2020. Effectiveness under sampling method and feature reduction in credit card fraud detection. Int. J. Intell. Eng. Syst. 13,173181.
doi:10.22266/ijies2020.0430.17.

[10] Vinod Jain, Mayank Agrawal, Anuj Kumar, "Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection, 2020 at 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2022.