Comparative Analysis of K-Nn, Naïve Bayes, and logistic regression for credit card fraud detection

Análisis comparativo de K-NN, Naïve-Bayes y regresión logística para la detección de fraude con tarjetas de crédito

Análise comparativa de K-NN, Naïve-Bayes e regressão logística para detecção de fraudes com cartão de crédito

Kavita Arora[1]
Sonal Pathak[2]
Nguyen Thi Dieu Linh[3]

Received: March 10th, 2023
Accepted: June 23th, 2023
Available: September 22th, 2023

How to cite this article:
Arora, S. Pathak, N. Thi Dieu Linh, "Comparative Analysis Of K-Nn, Naïve Bayes, A nd Logistic Regression for Credit Card Fraud Detection," *Revista Ingeniería Solidaria,* vol. 19, no. 3, 2023.
doi: https://doi.org/10.16925/2357-6014.2023.03.05

Research article. https://doi.org/10.16925/2357-6014.2023.03.05

[1] Associate Professor, Department of Computer Applications, Manav Rachna International Institute of Research & Studies, Faridabad, Haryana, India – 121004

Email: kavita.fca@mriu.edu.in

ORCID: https://orcid.org/0000-0002-3389-3939

[2] Professor, Department of Computer Applications, Manav Rachna International Institute of Research & Studies, Faridabad, Haryana, India – 121004

Email: sonal.fca@mriu.edu.in

ORCID:https://orcid.org/0000-0003-4435-1614

[3] Professor, Hanoi University of Industry, Hanoi, Vietnam

Email: nguyen.linh@haui.edu.vn

Abstract

Introduction: This paper highlights the outcome of the comparative study of "Various Machine learning algorithms namely K-NN, Naive Bayes, and Logistic Regression for Credit Card Fraud Detection" carried out based on a dataset taken from UCI.com in 2022-23 at Manav Rachna International Institute of Research and Studies.

Problem: Credit card fraud is still rife today and the modes are increasingly varied. Quite often we hear of fraud cases that cause irreplaceable injury to banks and financial institutions which cannot be compensated in terms of costs. To avoid scams with various modes of credit cards, we must be able to identify and find out the modes often used by fraudsters. This scheme liberates such financial institutions and banks with complete and appropriate information using Machine Learning Techniques, not only about the modes that scammers or fraudsters often use but also ways to protect against such frauds.

Objective: The present paper discusses the various machine learning models based on classification and regression, namely K-Nearest Neighbors, Naïve Bayes, and Logistic Regression, which are successfully able to achieve the classification accuracy of 80% using Logistic Regression with a Precision of 78%, Recall of 100%, and F1-Score of 88% for fraudulent credit card transactions.

Methodology: The comparative analysis demonstrates that for Precision, Recall, and Accuracy parameters, the K-Nearest Neighbor is a better approach for detecting fraudulent transactions than the Logistic Regression and Naïve Bayes.

Results: The accuracy is marginal high in Logistic Regression but the False Positive parameters are not able to identify the imbalanced data; therefore, they disguise the results and accuracy of Logistic Regression and K-Nearest Neighbor deems fit for such cases.

Conclusion: This scheme depicts the automated fraud classification systems using machine learning techniques, namely K-Nearest Neighbor, Logistic Regression, and Naive Bayes, to produce a model that can distinguish valid and invalid credit card transactions.

Originality: Through this research, the most relevant features are used to go through the visualization of accuracy with the confusion matrix, and accuracy calculations are obtained from the dataset used.

Limitations: Deep learning techniques could have been used to fetch even better results.

Keywords: Fraud Detection, Machine Learning, Naïve Bayes, K-Nearest Neighbor, Logistic Regression.

Resumen

Introducción: este artículo muestra el resultado de un estudio comparativo de "varios algoritmos de machine learning, a saber, K-NN, Naïve-Bayes y regresión logística para la detección de fraudes con tarjetas de crédito", realizado con base en un conjunto de datos tomado de UCI.com en 2022-23 en el Instituto Internacional de Investigaciones y Estudios Manav Rachna.

Problema: el fraude con tarjetas de crédito está muy extendido hoy en día y las modalidades son cada vez más variadas. A menudo, se oye hablar de casos de fraude que causan daños irreparables a bancos e instituciones financieras, que no pueden ser compensados en términos de costos. Para evitar estafas con diversos modos de tarjetas de crédito, se debe poder identificar y descubrir los modos que suelen utilizar los estafadores. Este esquema proporciona a dichas instituciones financieras y bancos información completa y adecuada utilizando técnicas de machine learning, no solo sobre los modos que suelen utilizar los estafadores o defraudadores, sino también sobre las formas de protegerse contra dichos fraudes.

Objetivo: el presente artículo analiza los diversos modelos de machine learning basados en clasificación y regresión, a saber, K-Nearest Neighbors (K-NN), Naïve Bayes y regresión logística, que pueden lograr con éxito una precisión de clasificación del 80% utilizando regresión logística con una precisión de 78%, Retiro del 100% y F1-Score del 88% para transacciones fraudulentas con tarjeta de crédito.

Método: el análisis comparativo muestra que, para los parámetros de precisión, recuperación y exactitud, el K-NN es un mejor enfoque para detectar transacciones fraudulentas que la regresión logística y el Naïve Bayes.

Resultados: la precisión es marginalmente alta en la regresión logística, pero los parámetros de falso positivo no pueden identificar los datos desequilibrados; por lo tanto, disfrazan los resultados y la precisión de la regresión logística y el K-NN se considera adecuado para tales casos.

Conclusión: este esquema describe los sistemas automatizados de clasificación de fraude que utilizan técnicas de machine learning, a saber, K-NN, Regresión logística y Naïve Bayes, para producir un modelo que pueda distinguir transacciones con tarjetas de crédito válidas e inválidas.

Originalidad: a través de esta investigación, se utilizan las características más relevantes para visualizar la precisión con la matriz de confusión y se obtienen cálculos de precisión a partir del conjunto de datos utilizado.

Limitaciones: se podrían haber utilizado técnicas de Deep learning para obtener mejores resultados.

Palabras clave: detección de fraude, K-Nearest Neighbor, Naïve Bayes, machine learning, regression logística.

Resumo

Introdução: Este artigo apresenta o resultado de um estudo comparativo de “vários algoritmos de aprendizagem automática, nomeadamente K-NN, Naïve-Bayes e regressão logística para detecção de fraude de cartão de crédito”, realizado com base num conjunto de dados retirados da UCI. com em 2022-23 no Instituto Internacional de Pesquisa e Estudos Manav Rachna.

Problema: As fraudes com cartões de crédito são hoje muito difundidas e as modalidades são cada vez mais variadas. É frequente ouvirmos falar de casos de fraude que causam danos irreparáveis a bancos e instituições financeiras, que não podem ser compensados em termos de custos. Para evitar fraudes com vários tipos de cartões de crédito, você deve ser capaz de identificar e descobrir os métodos que os golpistas costumam usar. Este esquema fornece a estas instituições financeiras e bancos informação completa e adequada através de técnicas de aprendizagem automática, não só sobre os métodos que os burlões ou fraudadores costumam utilizar, mas também sobre as formas de se protegerem contra tais fraudes.

Objetivo: O presente artigo discute os vários modelos de aprendizado de máquina baseados em classificação e regressão, nomeadamente K-Nearest Neighbours (K-NN), Naïve Bayes e regressão logística, que podem atingir com sucesso uma precisão de classificação de 80. % usando regressão logística com uma precisão de 78%, saque de 100% e pontuação F1 de 88% para transações fraudulentas com cartão de crédito.

Método: A análise comparativa mostra que para parâmetros de precisão, recall e exatidão, K-NN é uma abordagem melhor para detectar transações fraudulentas do que a regressão logística e Naïve Bayes.

Resultados: A precisão é marginalmente alta na regressão logística, mas os parâmetros falsos positivos não conseguem identificar dados desequilibrados; portanto, disfarçam os resultados e a precisão da regressão logística e o K-NN é considerado adequado para tais casos.

Conclusão: Este esquema descreve sistemas automatizados de classificação de fraude que utilizam técnicas de aprendizagem automática, nomeadamente K-NN, Regressão Logística e Naïve Bayes, para produzir um modelo que pode distinguir transações de cartão de crédito válidas e inválidas.

Originalidade: Através desta pesquisa, os recursos mais relevantes são utilizados para visualizar a precisão com a matriz de confusão e os cálculos de precisão são obtidos a partir do conjunto de dados utilizado.

Limitações: Técnicas de aprendizagem profunda poderiam ter sido utilizadas para obter melhores resultados.

Palavras-chave: detecção de fraude, K-Nearest Neighbor, Naïve Bayes, aprendizado de máquina, regressão logística.

1. INTRODUCTION

E-commerce dealings in India are increasing due to the enhancement in financial growth in accumulation to the growing middle class. The World Bank affirms that 56.5 percent of India's inhabitants, or around 134 million people, are in the middle-class group with a spending rate of 12-120 US dollars every day. This middle-class crowd has a moderately high income as they are moderately educated and are constantly interacting with the internet. The expansion of tools and communication apparatus has collided with the booming world of online commerce. In 2009, in India, only 13 percent of internet users used to shop online. However, now the number has reached 36 percent of internet users and this number continues to grow. According to www.statista.com's comprehensive review, more than 85 percent of the world's online inhabitants have utilized the internet for purchases [1].

In India, social media and E-Commerce platforms are pitched (49.2%) for buying products ranging from trendy clothes, to electronics, to reading, to household gadgets. With these burgeoning E-commerce platforms, junctures for unsocial actions and fraudulent behavior have opened, which were earlier thought to be non-viable. Fraudulent behavior is another face of cybercrime, which includes all actions and their modes of operation, carried out using technology. The crime that occurs very frequently is credit card theft or credit card fraud or carding. According to [2], carding is credit card fraud if the offender is aware of the validity of somebody's credit card number, then they make a purchase online and the bill is routed to the primary credit card holder; the offender is termed as a carder. In this crime, the credit card owner will lose his money because it has been used by another person. Such theft is carried out by breaking into the security of online shops that have made transactions and if the online shops do not have strong security, credit card accounts can then be hijacked by the carders [2].

Misdemeanors carried out using computers and networks, termed as a cyber offense in India, is on the high-rise. For credit card break-in cases (credit card fraud/carding) alone, based on research results from www.statista.com, an information technology (IT) company based in the USA, in 2020, India was in the 2nd position as the country of origin for most carders in the world after Ukraine. The results of this research solidify the impression that India did not do much to make changes from 2018 to 2020 when its "new" position was second only to Ukraine. In fact, at that time the image of the Indian Internet had already been highlighted by mass media abroad such as Time and Business Week magazines, which also quoted E-Commerce research results at the time. Not only that, until now, almost all users of the well-known auction site eBay.com are very "afraid" of making a transaction with someone who asks to have their goods sent to an address in India. For them, addresses in India have been included in their black-list records [3].

Cybercrime also falls into the transnational crime category. Transnational crime networks are not a new problem, indeed, transnational crime operations have existed for a long time, but it is only in the last two decades that transnational forms of crime have shown increased activity, are more organized and move more effectively, and are able to carry out criminal operations without significant legal impediments. The manifestation of cybercrime that has occurred so far is very detrimental to people's lives or the interests of a nation and State in international relations. Today's cybercrime is experiencing rapid development without recognizing the boundaries of the state (borderless state), with the technological advances used by the perpetrators being quite sophisticated. Even in developing countries, law enforcers, especially the police, are unable to prevent and overcome them, due to limited human resources, technology facilities, and infrastructure.

1.1 Literature Review

Authors in [4] have done a survey on credit card fraud detection while taking into consideration major fraud-related domains, namely corporate, bank, and insurance frauds, along with the mode of the transaction, either virtual or physical. The authors have also discussed various Machine Learning techniques such as Regression, Classification, Logistic Regression, K-Nearest Neighbor, Naïve Bayes, and Genetic Algorithms, along with certain data-mining approaches. As per the researchers, every individual machine learning technique provides a different accuracy rate for detection purposes, and enterprises are looking for the best technique which can enhance the profit rate and decrease the cost incurred. Among these, the classification algorithm K-Nearest Neighbor Classifier (KNN) is also called 'lazy Learner" because training data gets delayed in this modeling procedure until it classifies the examples by necessary labeling [5]

In [6] the authors have proposed a fraud detection model which is based on a decision tree and a combination of Luhn's and Hunt's algorithms and decision trees. Luhn's algorithm ascertains if an incoming transaction is fraudulent and for this, it vouches the number of a credit card. The parameters like Address Mismatch and Degree of Outlines are applied to appraise the deviation on the arrival of every transaction. Eventually, the credence is reinforced or debilitated by making use of the Bayes Theorem, ensured by using the values of probability.

Authors in [7] have applied numerous methodologies so as to ascertain the best-performing model to recognize crooked transactions. The research work was carried out using approaches like Neural Networks, Bayesian Networks, Support Vector Machines (SVM), and K-Nearest Neighbor (KNN). A comparison table has been shown to exhibit that Bayesian Network has been swift in unearthing the crooked transactions and also with a high degree of precision. On the other hand, Neural Networks performed the same task with a medium degree of precision and KNN's speed was high with a medium degree of precision, and eventually Support Vector Machine could attain the minimum total with low speed and medium degree of precision. In terms of the cost factor, every model was proven to be over-priced.

In [8], the authors proposed a model for credit card fraud detection with the help of KNN and Outlier detection while using oversampled data. Here KNN was the most befitting approach to identify and establish the commended aberration with the memory impediment. Furthermore, the storage and computation incumbent in the case of the Outlier detection method is significantly lower, despite its sprite and finer operation procedure. The authors' work and inference proclaimed that KNN was highly explicit and coherent.

In [9], the authors compared three widely used Machine Learning methods for credit card fraud detection: first KNN, second Naïve Bayes, and third Logistic Regression. During their research work, they explored divergent dispensations from the perspective of viewing innumerable results. The highest degree of precision, with 1:9 dispensations in the case of Naïve Bayes, turned out to be 97.5%; KNN turned out to be 97.1%, and Logistic regression performed unsuccessfully with a 36.4% degree of precision. Another dispensation contemplated was 34:66, in which KNN managed to top the chart with a modest improvement in the precision of 97.9%, followed by Naïve Bayes with 97.6%, and Logistic Regression moving up to a 54.8% degree of precision.

Table 1. Comparative Analysis (source: self-created)

Sr. No.	Author/s	Title of the Paper	Conclusion and Findings
1.	Alenzi, H. Z., & Aljehane, N. O.	Fraud detection in credit cards using logistic regression. International Journal of Advanced Computer Science and Applications (2020)	A comparison was performed between their model and two other Classifiers: Voting and KNN. Voting Card classifier scored 90% Accuracy, 88% Sensitivity and 10% Error Rate. KNN scored 93% Accuracy, 94% Sensitivity and 7% Error Rate.

(continúa)

(viene)

Sr. No.	Author/s	Title of the Paper	Conclusion and Findings
2.	Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A.	Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis. International Conference on Computing Networking and Informatics (ICCNI) (2017)	They used 3 techniques namely KNN, Naïve Bayes and Logistic Regression. Results: Accuracy with Naïve Bayes: 97.5% Accuracy with KNN: 97.1% Accuracy with Logistic Regression: 36.4%
3.	Jain, Y., Namrata Tiwari, S., & Jain, S.	A comparative analysis of various credit card fraud detection techniques. International Journal of Recent Technology and Engineering (2019)	The authors used Machine learning techniques to detect Credit Card Fraud using SVM, ANN, KNN. To compare the outcome of each model, they calculated True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). ANN: 99.71% Accuracy, 99.68% Precision SVM: 94.65% Accuracy, 85.45% Precision KNN: 97.15% Accuracy, 96.84% Precision
4.	Adepoju, O., Wosowei, J., lawte, S., & Jaiman, H.	Comparative evaluation of credit card fraud detection using machine learning techniques. Global Conference for Advancement in Technology (GCAT). (2019)	The authors used Logistic Regression, SVM, Naïve Bayes and KNN on distorted dataset. The Accuracy rate with Logistic Regression is 99.07%, 95.98% with Naïve Bayes, 96.91 with KNN, and 97.53% with SVM
5.	Safa, M. U., & Ganga, R. M.	Credit Card Fraud Detection Using Machine Learning. International Journal of Research in Engineering, Science and Management (2019)	They investigated Logistic Regression, KNN and Naïve Bayes techniques on exceptionally distorted credit card dataset. The Accuracy rate with Logistic Regression is 97.69%, 83% with Naïve Bayes, and 54.86 with KNN.
6.	Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M.	Credit card fraud detection using Naïve Bayes model based and KNN classifier. International Journal Of Advance Research, Ideas And Innovations In Technology (2018)	The outcomes of the research work showed that Naïve Bayes performed better than KNN with an Accuracy rate of 95% and 90% respectively.
7.	Saheed, Y. K., Hambali, M. A., Arowolo, M. O., & Olasupo, Y. A.	Application of ga feature selection on Naive Bayes, random forest and SVM for credit card fraud detection. International Conference on Decision Aid Sciences and Application (DASA) (2020)	This paper focuses on detection of credit card frauds with Genetic Algorithm as a feature selection technique. The feature selection data is splitted in 2 parts; first priority features and second priority features. The Machine Learning techniques used were Naïve Bayes, Support Vector Machine, and Random Forest with an Accuracy rate of 94.3%, 96.43% and 96.40% respectively.
8.	Itoo, F., Meenakshi, & Singh, S.	Comparison and analysis of logistic regression, Naïve Bayes and KNN Machine Learning Algorithms for credit card fraud detection. International Journal of Information Technology (2020)	The authors used Logistic Regression, Naïve Bayes and KNN for the detection of Credit card frauds. The mentioned techniques came out with an Accuracy rate of 91.2%, 85.4% and 66.9% respectively.
9.	Dighe, D., Patil, S., & Kokate, S.	Detection of credit card fraud transactions using machine learning algorithms and Neural Networks: A comparative study. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (2018)	The authors worked on KNN, Naïve Bayes, Neural Networks and Logistic Regression techniques and the Accuracy rate was 99.13%, 96.98%, 96.40% and 96.27% respectively.

Source: own work

2. METHODOLOGY

2.1 Techniques Used for Fraud

(i) *Carding*

Carding is credit card deception if the offender, on recognizing a person's valid credit card number, does online buying and the bill is addressed to its actual possessor [10]. The other name for such crimes is cyber offense [11]. Carding crimes have two categories, National and International. Nationally, carding actors perpetrate inside the purview of a nation. International carding is when the offense is committed across national borders. According to [12], credit card abuse can be done in two ways:

1. Credit cards are valid but are not being used as per the by-laws mentioned in the agreement.
2. Incapacitated/fake cards that are being used unlawfully.

In addition, carding is a terminology commonly used by hackers for fraud-related acts using credit cards. This is indicated by several definitions of carding. According to Doctor Crash, which wrote an article in the hackers' bulletin, the definition of carding is: "A way of obtaining the necessary goods without paying for them." The nature of carding in general is non-violent, the chaos it causes is not seen directly, but the impact it can have is very large. One example can be using someone else's account number to shop online for the sake of enriching yourself. Previously, the perpetrator (carder) stole the account number from the victim.

Even though, the prevention of carding is very difficult to overcome, not as in conventional cases, preventive steps must still be taken. This is intended so that the space for carding actors can be narrowed. Here are some of the methods commonly used by carders:

(ii) *Extrapolation*

It is widely known that the 16-digit credit card number has a certain algorithmic pattern. Extrapolation is carried out on a credit card which is commonly referred to as a master card so that other credit card numbers can be obtained which will be used for transactions. However, this method is arguably out of date, due to the development of today's security devices [13].

(iii) *Hacking*

This method of piracy is done by breaking into an online store website that has a weak security system. A hacker will hack an online store website, to then retrieve customer data. Carding with this method is not only detrimental to credit card users but also detrimental to the store because its image will be damaged, so customers will choose to shop in other safer places [14].

(iv) *Sniffer*

This method is done by sniffing and recording transactions made by a credit card user using the software. This can only be done in the same network, such as in an internet cafe or hotspot. The perpetrator uses sniffer software to intercept transactions made by someone who is on the same network so that the perpetrator will get all the data needed for further carding. The prevention of this method is that the e-commerce website will implement a functioning SSL (Secure Socket Layer) system to encode databases from customers [15].

(v) *Phishing*

Carding actors will send random and bulk emails on behalf of an agency such as a bank, shop, or service provider, which contain a notification and an invitation to log into the agency's website. However, the site that is notified is not the original site, but a site that is made very similar to the original site. Furthermore, the victim is usually asked to fill in the database on the site. This method is the most dangerous method because the hijacker can get complete information from the credit card user himself. The information obtained is not only the user's name and credit card number, but also the date of birth, identity number, credit card expiration date, and even height and weight if the carding actor wants it. The impact of this carding crime includes [16]:

I. Lost money mysteriously
II. Credit Card Extortion and Draining
III. People's unrest in using credit cards
IV. The loss of public trust in financial services in this country

Artificial intelligence (AI) is brainpower supplemented by machines and can be cited in a scientific context as the intelligence of a scientific entity. The artificially intelligent machine has the capability to decipher data, precisely learn this data, and apply that learning in order to accomplish outcomes. Intelligence is generated and catered to in devices so as to make them work like humans. Thereafter, Machine learning is a branch of Artificial Intelligence that includes designing and developing algorithms which in turn permit machines to develop deportment on the basis of empirical data. The machine further makes use of this data to grab traits that are required from the underlying probability. In 1959, Arthur Samuel defines machine learning as a field of study that gives the ability to study without being programmed explicitly. Learning abilities that become dominant are determined by the ability of the software or the algorithm. Machine learning can work for adapting to a new situation, as well as to detect and predict a pattern. The Algorithm in machine learning can be grouped by the expected input and output of the algorithm. Machine Learning algorithms are categorized as:

1. Supervised learning: an approach wherein we have expected output beforehand which is used for learning. This approach to learning is subdivided into Classification and Regression.
2. Unsupervised learning is learning that is not supervised and does not require a target output. The idea of this approach is to group the same output units in one particular zone. This learning is found to be appropriate for pattern classification. Unsupervised learning is subdivided into Association and Clustering.
3. Reinforcement learning algorithms carry on with learning from the environment iteratively. During this, the agent learns on the basis of experiences. Machine Learning actually requires data to be learned by training data. It describes several processes for building a Machine Learning system, namely:

    - Collect data
    - Prepare input data
    - Analyze data input (Analyze input data)
    - Include human involvement (Human involvement)
    - Training the algorithm (Train algorithm)
    - Testing the algorithm (Test algorithm)
    - Using the algorithm (Use it)

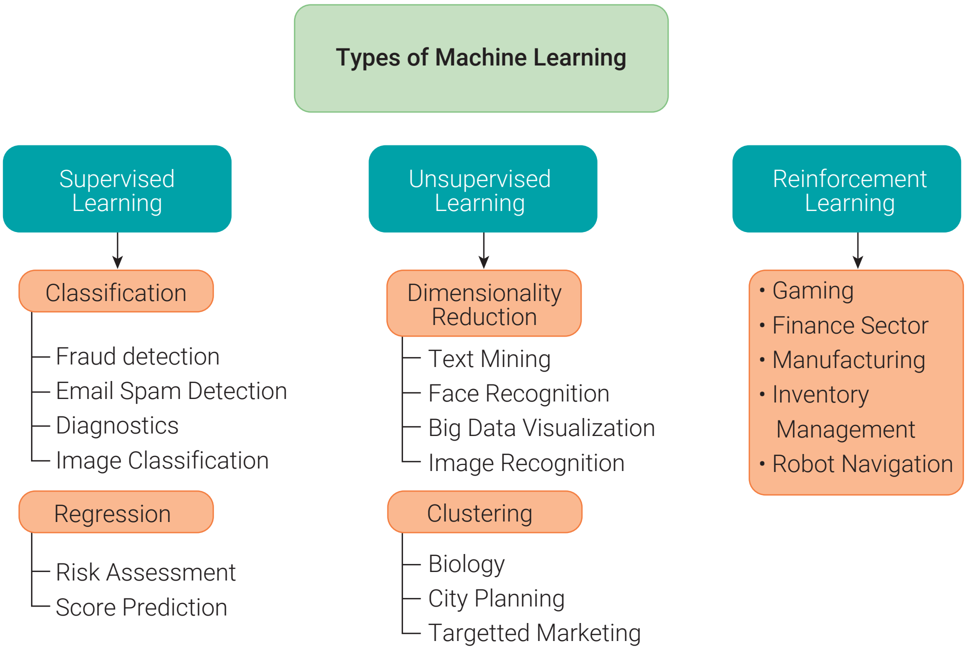


Figure 1. Types of Machine Learning Techniques
source: [11]

2.2 Machine Learning Tools for Credit Card Fraud Detection

(i) *K- Nearest Neighbor Classifier*

KNN algorithms aspire to classify new data objects. This approach is based on unearthing several k data objects (training data) that are nearest to given test data, subsequently choosing the class with a maximum number of votes. The steps to implement the algorithm are:

1. Ascertain the k parameter (number of closest neighbors).
2. Calculate the square of the object's Euclidean distance to training data
3. Sort the results in ascending order.
4. Collect the nearest neighbor classified based on the k-value
5. Use the nearest neighbor category.

k in the k-nearest neighbor algorithm is the quantum of adjacent neighbors that are used as points to classify new data or objects. In determining the number of k values, odd numbers should be used. To calculate the distance between data objects in this algorithm, the Euclidean Distance method can be used. Here's the formula for this:

$$D(i,j) = \sqrt{\sum_{k=1}^{n} \square\, (x_i k - y_i k)^2} \qquad (eq. 1)$$

Where:
i, j: the matrix to measure distance
n: the amount of data on the matrix
x: matrix value

K-Nearest Neighbor (KNN) uses neighborhood classification as a predictive value for the new query instance. As an illustration of the application of the KNN algorithm, if there is data from a survey using a questionnaire to enquire into peoples' opinions about the test of attributes namely acid resistance and strength for classification of quality of tissue paper as to Good or bad. The following four training data can be used for this purpose. If the density is low, the cells will grow larger-, but will stop after entering a region that has a high density. Therefore, to assist the space the function is:-

$$d_i = \{0\ 1 \quad \frac{if\ P(x_i)=y_i}{besides} \qquad (eq. 2)$$

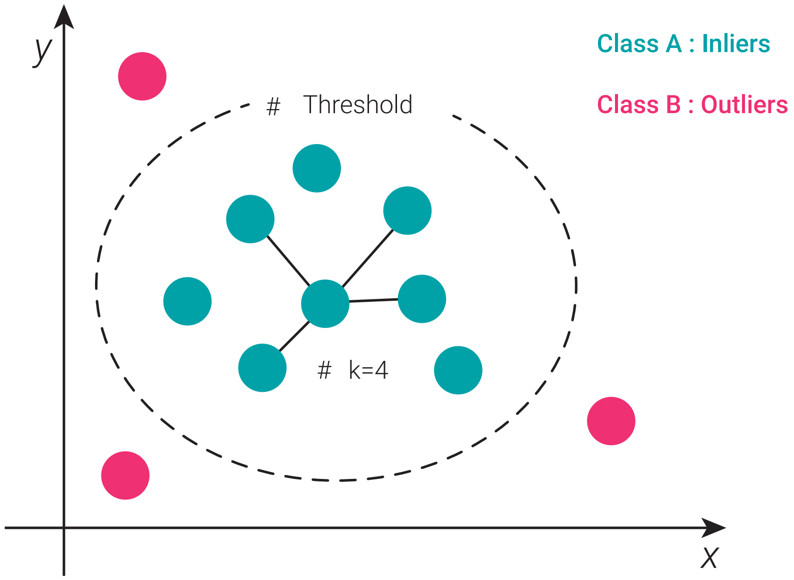


Figure 2. K-NN evaluating Threshold using class A and B
Source:[14]

(ii) *Logical Regression*

Regression is how one variable, namely the dependent variable, is affected by one or more other independent variables with the aim of predicting the mean value. The main aim of regression is a prediction of the dependent variable's value done on the basis of one or more independent variables. The specific form of the logistic regression model is:

$$\pi(x) = \frac{e^{\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}{1+e^{\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}} \qquad (eq.3)$$

Where,
$\pi\,(x)$: predicted output,
$\beta 0$ **:** bias
$\beta 1$: coefficient of an input (x).

Every column of input data is analogous to coefficient β.

(iii) *The Naive Bayes Classifier*

The Naive Bayes Classifier algorithm carries out data mining techniques by putting in the Naive Bayes method in classifying data. This algorithm was given by Thomas Bayes and uses the Bayes theorem to compute a number of probabilities for events that have an impact on observed results. This theorem describes the association between the probability of events A and Z, described as:

$$P(x) = \frac{P(H) * P(H)}{P(x)} \qquad (eq.4)$$

Or

$$P(x) = \frac{P(H) * P(H)}{P(x|H)P(H) + P(H)P(H)} \qquad (eq.5)$$

For data sample class x whose label is unknown, and H is a hypothesis, sample data x is transferred to a particular special class c. $P\ (H/x)$ is the posterior probability, $P\ (H)$ is the probability H before the sample is used,

3. RESULTS

3.1 Proposed Scheme

The proposed scheme talks about credit card fraud detection systems with the help of comparative analysis of KNN and Logistic Regression algorithms in addition to classification and regression algorithms, aiming to procure optimal elucidation as time progresses. Here, we aim to diminish false alerts with the help of a Machine Learning algorithm while optimizing a group of interval-valued parameters. For that reason, through this work, we have tried to evolve a fraud detection system using K-NN and Logistic Regression algorithm. By using this proposed scheme, we can detect malicious activities and can raise false alerts while making credit card transactions. The parameters considered for comparative analysis are precision, recall, and accuracy. We have mentioned here a pseudo code of Logistic Regression and KNN along with the proposed workflow and description of data.

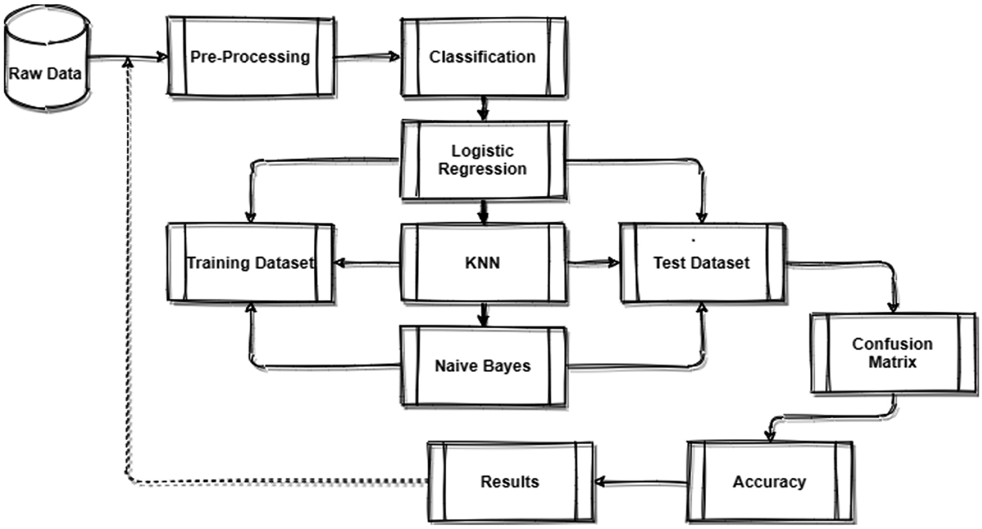
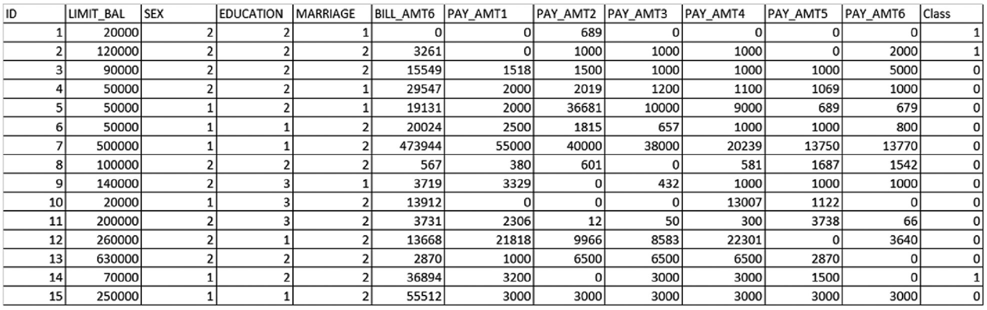


Figure 3. Proposed Methodology using K-Nearest Neighbors, Logistic Regression and Naïve Bayes
Source:[7]

The respective data has been obtained from https://archive.ics.uci.edu/ml/machine-learning-databases/00350/ comprising a 3000 raw data set, shown below:



ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	Class
1	20000	2	2	1	0	0	689	0	0	0	0	1
2	120000	2	2	2	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	19131	2000	36681	10000	9000	689	679	0
6	50000	1	1	2	20024	2500	1815	657	1000	1000	800	0
7	500000	1	1	2	473944	55000	40000	38000	20239	13750	13770	0
8	100000	2	2	2	567	380	601	0	581	1687	1542	0
9	140000	2	3	1	3719	3329	0	432	1000	1000	1000	0
10	20000	1	3	2	13912	0	0	0	13007	1122	0	0
11	200000	2	3	2	3731	2306	12	50	300	3738	66	0
12	260000	2	1	2	13668	21818	9966	8583	22301	0	3640	0
13	630000	2	2	2	2870	1000	6500	6500	6500	2870	0	0
14	70000	1	2	2	36894	3200	0	3000	3000	1500	0	1
15	250000	1	1	2	55512	3000	3000	3000	3000	3000	3000	0

Figure 4. Example Data used from uci.com

3.2 Description of Preprocessed Data

Preprocessing Data stages are accomplished to get data that is ready for use. Then, at last, cleaning data that is achieved is ready to be used as taken in this study, as many as 30000 good records, with 23 attributes. We can see the data before preprocessing in Figure 4.

K-Nearest Neighbor (KNN) Algorithm

The K-Nearest Neighbor (KNN) algorithm is an approach used for classifying objects contingent on learning data. The principle of KNN is to perceive the nearest interspace between data eventually to be evaluated and the closest k neighbors in training data. The KNN algorithm uses the supervised learning method. The best k value here counts upon data. We can opt for a good k value by exploiting parameter optimization such as cross-validation. The particular instance in which classification is envisaged using closest learning data or k = 1, is known as the k-nearest neighbor algorithm. The Impetus of using the KNN algorithm is to categorize new objects on the basis of samples collected, taking into consideration their attributes and training. KNN algorithm implies the neighborhood classification of the predictive value of the new test sample. The ranking for k closest neighbors on the basis of similarity value is computed using the Euclidean distance:

$$(X,Y) = \sqrt{\sum_{k=1}^{n} \quad (X_i - Y_i)^2} \qquad (eq.6)$$

With

$D(X, Y)$: Euclidean Distance

X_i: sample data

Y_i: test data

n: dimension data

k: variable data

Pseudocode

K-NN pseudocode can be written as:

```
Step 1 Ascertain $k$◌, nearest neighbor
Step 2 Compute the distance of input data with training data Distance measure is the Euclidean distance
```

$$D(X,Y) = \sqrt{\sum_{k=1}^{n} \square \, (X_i - Y_i)^2}$$

```
Step 3 Sort the distance from the nearest Step 4 Check the nearest neighbor:: class
Step 5 New data class = closest neighbor majority class.
```

Logistic Regression Algorithm

Logistic regression is one of the most frequent classification methods used. Binary logistic regression is used when the dependent variable consists of dichotomous variables. Multinomial logistic regression was used at variable times. The dependent variable is a categorical variable with more than two categories. Generally, logistic regression models are:

$$\pi(x) = \frac{e^{\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}{1+e^{\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}$$

Where $\pi (x)$ is the probability value of $0 \leq \pi (x) \leq 1$, which means that the logistic regression describes a probability. By transforming $\pi (x)$ in the above equation with the logit transformation $g (x)$,

where: $(x) = In\left(\frac{\pi(x)}{1-\pi(x)}\right)$

then the logistic form is obtained:

$$\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i .$$

To obtain estimates from logistic regression parameters, it can be done by using the Maximum Likelihood Estimation (MLE) as follows: Estimating parameters in the logistic model using Maximum Likelihood with the following steps:

1. The likelihood function of Y

$$L(\beta) = \prod_{i=1}^{n} [p]^{y_i}[q]^{1-y_i} \qquad (eq.7)$$

The ln-likelihood function

$$In\ L(\beta)\ In\{\prod_{i=1}^{n} [p]^{y_i}[q]^{1-y_i}\}$$

$$n\ L(\beta)\ In\{\prod_{i=1}^{n} [p]^{y_i}[1-p]^{1-y_i}\}$$

$$In\ L(\beta)\ In\left\{\prod_{i=1}^{n} \left[\frac{exp\left(g(x)\right)}{1+exp\left(g(x)\right)}\right]^{y_i}\left[1-\frac{exp\left(g(x)\right)}{1+exp\left(g(x)\right)}\right]^{1-y_i}\right\}$$

$$In\ L(\beta) = \sum_{s=0}^{q} \left\{\left[\sum_{i=1}^{n} y_i x_q\right]\beta_1 - \sum_{i=1}^{n} In\left[1 + exp \sum_{s=0}^{q} \beta_q x_q\right]\right\}$$

Evaluating the regression function is done by dividing data into 2 parts. The first part will be used as a training set, which is treated to form a logistic regression classification model. Next, the second part will be used as a validation set, which serves as a cross-validationof the logistic regression function. Classifying is expected to minimize misclassification or minimize the average adverse effect of misclassification.

The steps to perform Logistic Regression are as follows:

(i) Divide the data into 90% training data and 10% testing data.
(ii) Conducting an independence test using training data.
(iii) Forming a Logistic Regression model using training data.
(iv) Testing the significance of parameters individually and as a whole
(v) Validate the prediction accuracy of the model with the data testing.
(vi) Calculating the value of Accuracy, Sensitivity, Specificity, and G-Mean using logistic regression model formed using likelihood estimations.

Naïve Bayes Algorithm

A Bayesian classifier is built on the Bayes theorem. Bayes' decision is a statistical perspective that is fundamental in pattern recognition. Let X be the attribute set data and h class variable and if the class has a relationship with attributes it requires X and h as a random variable and captures the relationship: odds *P (h | X X)* are posterior odds for prio opposite perior *P (h)*. However, Naive Bayes Classifier evaluates the probability of a conditional class on assumption that the attribute is independent subject to the condition, given the class label. The conditional independent hypothesis can be expressed as:

$$P(X, Y = y) = \frac{P(y)\prod_{i=1}^{q} P(X_i|Y=y)}{P(x)} \qquad (eq. 8)$$

where each set of attributes X = {X1, X2, X3, , Xn} consists of d attributes.

Steps in the Naive Bayes algorithm:

1. Prepare training data
2. Present data as an n-dimensional vector, namely X = {X1, X2, X3, , Xn}
3. *n* is a description of the size made in the test of n attributes, namely A1, A2, A3,
   An
4. *M* is a collection of categories, namely C1, C2, C3, Cm
5. Given the *X* test data whose category is unknown, the classifier envisions that *X* appertains to the category with the maximum posterior probability based on condition *X*
6. Naive Bayes classifier indicates that the unknown *X* test is from category *C1* only in cases where *P (Ci | X)> P (Cj | X) for 1 ≤ j ≤ m, j ≠ i.*
7. Maximize P (C_i | X) $P(X) = \frac{P(C_i).P(C_j)}{P(X)}$
8. Where *x* is the attribute value in sample *x* and the probability *P (x1 | Ci), P (x2 | Ci), ... P (xn | Ci)*, can further be appraised using a training dataset.

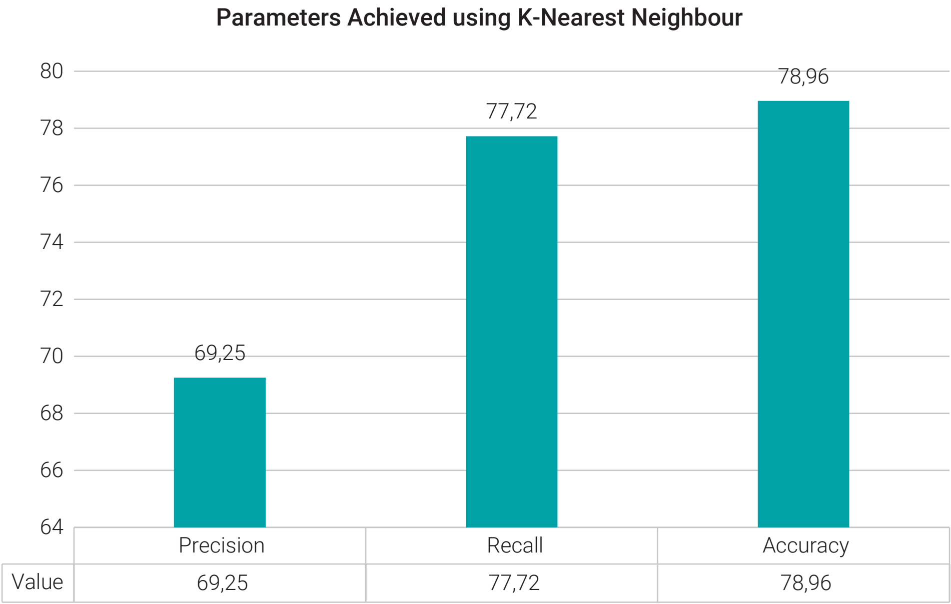


Figure 5. Precision, Recall and Accuracy Achieved using KNN
Source: own work

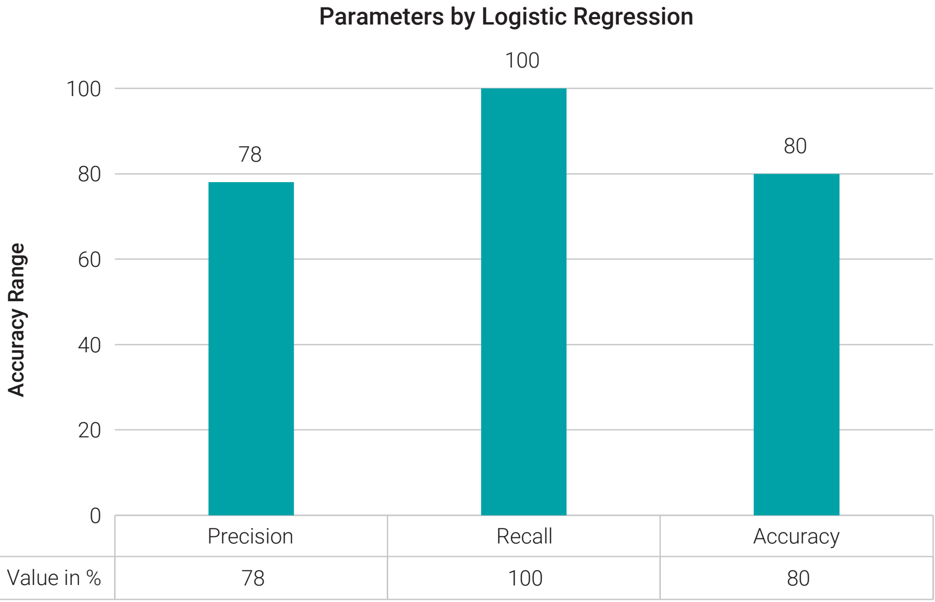


Figure 6. Precision, Recall and Accuracy Achieved using Logistic Regression
Source: own work

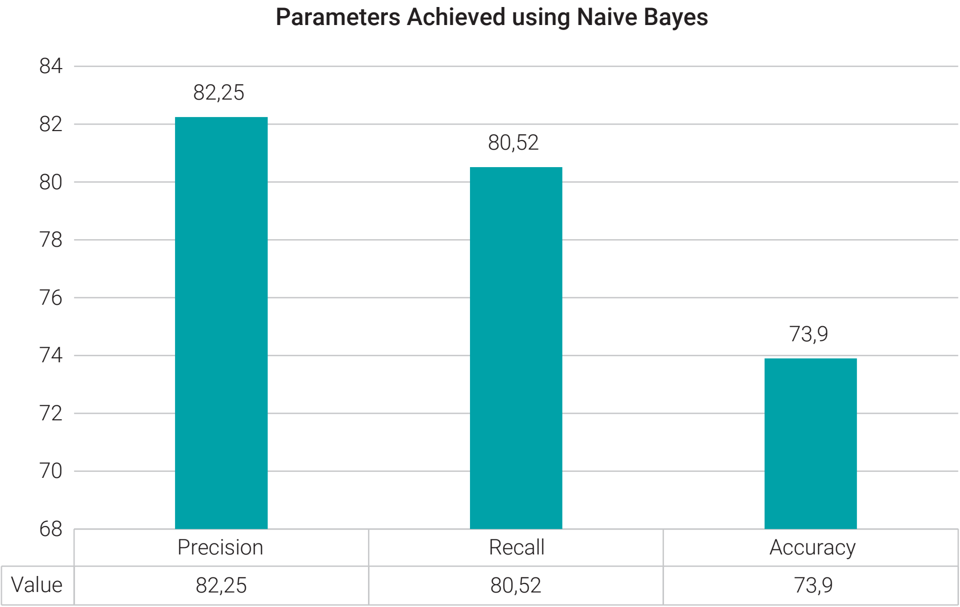


Figure 7. Precision, Recall, and Accuracy Achieved using Naïve Bayes Algorithm
Source: own work

4. DISCUSSION

The search carried out by various authors proves that many researchers are carrying out efforts to resolve the issues of credit card frauds using different data splitting ratios to generate different accuracy levels. Based on a series of works of literature on the various machine learning models that have been developed for credit card fraud detection, the authors are not able to identify a mechanism where the above-mentioned dataset is used for the analysis. Therefore, it can be concluded that the most relevant features are used in this scheme to go through the visualization of accuracy with the confusion matrix, and accuracy calculations are obtained for the said dataset. Thus, it can be inferred that the proposed classifiers, with the developed model, are capable of performing classification analysis of criminal acts of credit card fraud.

5. CONCLUSIONS

One form of data manipulation in the field of e-commerce is credit card fraud. Credit card transactions are the most common payment method in recent years. However, if fraud cannot be prevented, it must be detected as early as possible, and necessary measures must be taken against it. Classification of fraudulent transactions is the process of detecting whether a transaction is valid or not. An automated fraud classification system is necessary, especially given the large traffic of transaction data, and it is impossible for humans to manually check each transaction. This scheme depicts the automated fraud classification systems using machine learning techniques, namely KNN, Logistic Regression, and Naive Bayes, to produce a model that can distinguish between valid and invalid credit card transactions consequently achieving accuracy of 79%, 80%, and 73.90% respectively. However, as per the results, the Logistic Regression can be considered as the best model based on performance and accuracy for credit card fraud detection.

As per exploration of output deduced and discussion, we reached the following conclusions:

1. Classification precision of Logistic Regression is 78% with veracity of 80%, and Recall of 100%.
2. Classification precision of KNN is 69.25% with veracity of 78.96%, and Recall of 77.72%.
3. Classification precision of Naïve Bayes is 82.85% with a veracity of 73.90%, and Recall of 80.52%.

The authors wish to provide several suggestions for further research:

1. Increasing the number of samples and using similar industries such as banking companies.
2. Techniques based on Deep Learning like LSTM and GRU can be used to achieve more accuracy.

6. REFERENCES

[1] S. L. Vailshery, “Wide-area and short-range IoT device installed base Worldwide 2014-2027,” *Technology & Telecommunications*. [Online] Available: https://www.statista.com.

[2] Credit card fraud, [Online] Available: https://en.wikipedia.org/wiki/Credit_card_fraud.

[3] S. Okoro, “Combatting Cybercrime, Tools and Capacity Building for Emerging Economies”, 2017. [Online], Available: https://documents1.worldbank.org/curated/en

[4] Lookerstudio, 2018. [Online] Available: https://www.indiacode.nic.in/bitstream/123456789 /1999/3/A2000-21.pdf

[5] A. Rashmi, "Predictive Analysis Of Breast Cancer Using Machine Learning Techniques," *Revista Ingeniería Solidaria*, vol. 15, no. 3, 2019. doi: https://doi.org/10.16925/2357-6014.2019.03.01

[6] R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection," *Knowledge-Based Systems*. vol.13, pp.93–99. [Online]. Available:https://isiarticles.com/bundles/Article/pre/pdf/17658.pdf

[7] Y.K. Saheed, Hambali, "Application of feature selection on Naive Bayes, random forest, and SVM for credit card fraud detection," International Conference on Decision Aid Sciences and Application. (DASA), 2020. doi: https://doi.org/10.1109/DASA51403.2020.9317228

[8] H. Najadat, O. Altiti, "Credit card fraud detection based on machines and Deep Learning," International Conference on Information and Communication Systems.2020. doi: 10.1109/ICICS49469.2020.239524

[9] R. Sailusha, V. Gnaneswar, R. Ramesh, G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 1264-1270. doi: https://doi.org/10.1109/ICICCS48265.2020.9121114.

[10] A. Gupta, M.C. Lohani, "Financial fraud detection using naive Bayes algorithm in highly imbalance data set," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 24, no. 5, pp. 1559–1572, 2021.

[11] D. Dighe, S. Kokate, "Detection of credit card fraud transactions using machine learning algorithms and Neural Networks: A comparative study," International Conference on Computing Communication Control and Automation. doi: https://doi.org/10.1109/ICCUBEA.2018.8697799

[12] Y. Jain, S. Jain, "A comparative analysis of various credit card fraud detection Techniques," *International Journal of Recent Technology and Engineering,* vol. 7, no.52, pp.402-407, 2019.

[13] Maniraj, S. Sarkar, "Credit card fraud detection using machine learning and Data Science," *International Journal of Engineering Research and Technology*, vol. 8, no. 9, 2019. doi: https://doi.org/10.17577/IJERTV8IS090031S.

[14] S. Kiran, J. Guru, "Credit card fraud detection using Naïve Bayes model based and KNN classifier," *International Journal of Advance Research, Ideas And Innovations In Technology,* vol. 4, no. 3, pp.44 - 47, 2018.

[15] S. Maes, K. Tuyls, "Credit card fraud detection using Bayesian and neural networks," *International naiso congress on neuro fuzzy technologies.* pp. 261-270.2002

[16] M. Zareapoor, K. Seeja, "Analysis on credit card fraud detection techniques: Based on certain design criteria," *International Journal of Computer Applications,* vol. 52, no. 3, pp. 35–42, 2012