Turkish Journal of Engineering

https://dergipark.org.tr/en/pub/tuje

e-ISSN 2587-1366

Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets

Vahid Sinap *[1]

[1] *Ufuk University, Department of Management Information Systems, Türkiye, vahidsinap@gmail.com*

Cite this study: Sinap, V. (2024). Comparative analysis of machine learning techniques for credit card fraud detection: Dealing with imbalanced datasets. Turkish Journal of Engineering, 8 (2), 196-208

https://doi.org/10.31127/tuje.1386127

Keywords
Credit card fraud
Fraud detection
Data mining
Machine learning
Imbalanced datasets

Research Article

Received: 04.11.2023
Revised: 30.11.2023
Accepted: 03.12.2023
Published: 07.04.2024

Abstract

The main objective of this research is to evaluate the performance of machine learning algorithms in the field of credit card fraud detection and then compare them according to various performance metrics. Seven different supervised classification algorithms including Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors and Support Vector Machine were used. The performance of these algorithms was measured through a comprehensive evaluation of metrics including Accuracy, Precision, Recall, F-Score, AUC and AUPRC values. Furthermore, ROC curves and confusion matrices were used to evaluate these algorithms. The data preparation phase is critical in this study. The data imbalance problem arises as an unequal distribution between fraudulent and non-fraudulent transactions. Addressing this imbalance is imperative for successful model training and subsequent reliable results. Various techniques, such as Scaling and Distribution, Random Under-Sampling, Dimensionality Reduction, and Clustering, are employed to ensure an accurate evaluation of model performance and its ability to generalize effectively. As a result, the "Random Forest" and "K-Nearest Neighbors" algorithms exhibit the highest performance levels in this research with 97% accuracy rates. This study contributes significantly to the ongoing fight against financial fraud and provides valuable guidance for future research efforts.

1. Introduction

Credit card fraud is a type of crime in which credit card information is obtained without permission and misused. The history of credit card fraud begins with the emergence of credit cards. Credit cards first began to be used in the USA in the 1950s [1]. At that time, the security measures of credit cards, which consisted of paper or metal plates instead of plastic cards, were very weak. For this reason, credit card fraud and theft became common problems from the outset. Credit card fraud causes both financial and ethical harm to credit card holders and financial institutions, constituting a significant problem worldwide. In 2019, the USA experienced a loss of $28.65 billion due to credit card fraud [2]. In Türkiye, 1.2 billion TL was lost due to credit card fraud in 2020 [3]. Therefore, combating credit card fraud has become an important field of work.

Both legal and technological measures are taken to combat credit card fraud. From a technological perspective, magnetic strips were added to credit cards in the 1960s, signature panels in the 1970s, holograms in the 1980s, and chips in the 1990s [4]. Today, methods such as biometric authentication, virtual credit cards, one-time passwords, tokenization, and blockchain have been developed. Legally, laws and regulations have been established and implemented at national and international levels to address credit card fraud [5]. These measures enhance the security of credit cards and make it more challenging for fraudsters to carry out their activities. However, fraudsters have also developed new methods and identified vulnerabilities in these security measures. In this context, fraudulent techniques such as copying magnetic stripe cards, using signature panel cards with forged signatures, imitating hologram cards, and compromising chip cards have emerged [6]. These fraudulent methods can be described in broader terms as follows:

- Card theft or loss: If a credit card is physically stolen or lost, the individual who discovers or steals it can use the card. This approach is the most straightforward and ancient.
- Card copying (skimming): Thieves engage in card copying by utilizing a specialized device to read and

copy the data from the magnetic strip of a credit card. This technique is commonly utilized at ATMs, gas stations, or restaurants.

- Phishing: Phishing involves duping individuals into divulging their credit card number, expiration date, and security code through deceitful emails, phone calls, or websites, usually perpetrated through emails or calls masquerading as the bank or institution.

The rise of online shopping, fueled by the extensive utilization of the internet, has emerged as a pivotal juncture in credit card fraud, presenting both convenience and peril to credit card holders. Fraudsters have ingeniously employed a multitude of tactics, such as counterfeit websites, phishing emails, malevolent software, and the exploitation of security vulnerabilities in wireless networks, to acquire credit card information. Moreover, as there is no requirement to physically present the credit card when making online purchases, stolen or lost credit cards can be readily utilized [7].

2. Credit card fraud detection

With the increasing number of cases of credit card fraud, the detection of this type of fraud has become of great importance. Credit card fraud detection is the process of classifying credit card transactions as normal or abnormal. This process is important to prevent or reduce losses for both credit card holders and banks. Various quantitative and statistical methods have been used in credit card fraud detection from past to present. Some of these methods are as follows:

- Behavioral analysis: This method aims to identify transactions that deviate from the norm by monitoring the shopping habits, spending amounts, frequency, and locations of credit card holders [8].
- Rule-based analysis: This method attempts to detect suspicious transactions by evaluating credit card transactions according to certain rules [9]. These rules can be criteria such as the transaction amount exceeding a certain threshold, the location of the transaction being far from the place of residence of the credit card holder, the frequency of the transaction exceeding a certain limit.
- Scoring analysis: In this method, certain points are assigned to credit card transactions, and it is aimed to measure the risk levels of transactions. Scores can be calculated based on variables such as transaction amount, transaction location, transaction time and transaction type. When the score of the transaction exceeds a certain threshold, the transaction is considered suspicious.

These methods have advantages as well as disadvantages. Advantages include simplicity, understandability, and applicability. The disadvantages include high false alarm rate, low accuracy, lack of flexibility and inability to adapt to new fraud methods [10]. With the developments in information technologies, the disadvantages of traditional detection methods are tried to be eliminated or reduced with current technologies such as machine learning and data mining. The applications of these technologies in credit card fraud detection are outlined as follows:

- Machine learning methods: Machine learning methods are automatically classifying credit card transactions based on patterns or features in the data. These methods include algorithms that can cope with the complexity and volume of data and discover hidden relationships in the data. Machine learning methods include various algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Decision Trees (DT) [11].
- Data mining methods: Data mining is the classification of credit card transactions according to statistical or logical rules in the data. These methods enable data to be transformed into meaningful and useful information. Among data mining methods, algorithms like K-Nearest Neighbor (KNN), Logistic Regression (LR), and Bayes' theorem are commonly employed [12].

There are significant challenges when using both machine learning techniques and traditional methods for detecting credit card fraud. One of these challenges is the imbalance and scarcity of data. While the vast majority of credit card transactions are considered normal, a small percentage can be classified as abnormal or fraudulent. While this makes it relatively straightforward for machine learning models to learn normal transactions, it becomes challenging to distinguish abnormal transactions. Additionally, credit card fraud data is often not shared due to privacy concerns, or there are limited datasets available, which hinders the provision of sufficient data for training and testing machine learning models. Various methods have suggested to tackle the issue of data instability and scarcity; these include:

- Data resampling methods: These involve the adjustment of normal or abnormal operation numbers to attain data balance. Utilizing techniques such as the Synthetic Minority Oversampling Technique (SMOTE) allows for an increase in minority class instances, specifically abnormal transactions [13]. Similarly, methods like random subsampling or near neighbor-based subsampling when employed significantly reduce majority class quantities; these represent normal operations [14].
- Transfer learning methods: This technique enable data acquired from different sources to be transferred to the target dataset. For instance, credit card transaction data from different banks or regions can be processed for the target bank or region. In this way, the amount of data can be increased, and model performance can be improved [15].
- Deep learning methods: With deep learning methods, complex and high-dimensional data can be processed and hidden patterns and relationships in the data can be discovered. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can learn the temporal and spatial characteristics of credit card transaction data and detect fraudulent transactions [16].

Another challenge of using machine learning methods to detect credit card fraud is related to the way these methods work. While machine learning models classify credit card transactions as normal or abnormal, they may not be able to explain how and why they make these decisions. This can undermine the trust of both credit card holders and banks. Moreover, machine learning models may produce false positive or false negative results. This may increase the financial and moral losses

of both credit card holders and banks. Various methods have been proposed to solve the problem of explainability and transparency of the decision process. Some of them are as follows:

- Decision tree-based methods: With these methods, credit card transactions are classified as normal or abnormal and the decisions taken are shown in simple and understandable rules. For instance, techniques like Random Forests (RF) yield results through the consensus of multiple DT, and the rules of these DT can be examined to understand the outcome [17].
- Sensitivity analysis-based methods: Sensitivity analysis-based methods show the importance of the input variables that are effective in these decisions when making classification. For example, approaches like LR calculate coefficients for input variables, indicating the impact of these variables on the outcome. To see this effect, the values of the input variables can be changed to see how the result changes [16].
- Comparative analysis-based methods: They are methods that show similar or different operations that play a role in these decisions during classification. For example, algorithms such as KNN consider the k closest transactions when deciding whether a transaction is normal or abnormal. To see how these transactions are selected, distance measures between transactions are examined [17].

Machine learning techniques for credit card fraud detection have become a popular research area in recent years due to their benefits. Studies in this area are evaluated with different datasets, techniques, performance measures and application scenarios. Some of these studies demonstrate the effectiveness and advantages of machine learning techniques in credit card fraud detection [11,16-17]. Another part of the work presents the challenges faced by machine learning techniques in credit card fraud detection and proposed solutions to overcome them [13-15].

In this research, models were created using various machine learning algorithms and a prediction was performed for the detection of credit card fraud and suspicious transactions. The aim of the research is to identify the machine learning algorithms that give the best results in credit card fraud and suspicious transaction detection. In addition, another aim of the research is to reveal which data preprocessing processes can be used when working on imbalanced datasets with machine learning algorithms.

3. Material and Method

In this section, explanations of the machine learning algorithms used in the research, performance metrics employed for comparing the algorithms, characteristics of the dataset, and information about the data preparation process are included.

3.1. Algorithms utilized

Machine learning is a sub-branch of artificial intelligence and is the ability of computers to make intelligent decisions by learning from data. Machine learning uses various methods according to the nature of the data and objective variables. When these methods are analyzed, three categories emerge as (1) supervised learning, (2) unsupervised learning and (3) reinforcement learning [18]. Supervised learning is an approach to learning in which a machine learning model tries to learn the relationship between a given input and a target output [19]. Supervised learning is typically divided into two fundamental categories: classification and regression. Classification basically aims to predict whether an input belongs to a certain category or not. Regression, on the other hand, focuses on predicting a continuous numerical output associated with input data. It is used to predict a specific value or a continuous function of an output variable [20]. In this research, supervised classification algorithms are used because credit card fraud detection requires real-time intervention and requires identification between two main classes: fraudulent and non-fraudulent transactions. The real-time nature of credit card transactions requires fast and accurate decision making. Supervised classification algorithms, which learn from historical data, are capable of instantly categorizing each transaction as it occurs, thus playing a crucial role in timely detection and prevention of fraudulent acts. Therefore, the choice of supervised classification algorithms is in line with the rigorous requirements of credit card fraud detection and provides an efficient and agile system for the protection of monetary transactions.

3.2. Supervised classification algorithms

In this research, a total of seven supervised classification algorithms were employed to detect credit card fraud and suspicious transactions, which included Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine.

3.2.1. Logistic regression

LR is a classification algorithm whose main purpose is to estimate the probability that an object or event belongs to a certain class [21]. This estimation is based on the relationship between the independent variables or features and the weights. Each feature has a separate weight, and these weights are learned during the training phase of the model. Predictions are made using a sigmoid function, which converts real numbers into a probability value between 0 and 1. The higher the probability, the more likely it is that the object belongs to a certain class. The accuracy of the model is determined by the error rate on the training data and this model can be used to classify new data and estimate probabilities [22].

The formula of logistic regression is given in Equation 1. When the formula is interpreted in terms of credit card fraud detection, $P(Y=1)$ represents the probability that an event or transaction belongs to a certain class (for example, fraud); $X_1, X_2, \ldots X_k$ denotes the independent variables related to the credit card transaction (such as transaction amount, transaction location, etc.); and b_0, b_1, $b_2, \ldots b_k$ represent the weights learned during the model's training, signifying the impact level for each feature.

$$P(Y=1)=\frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_kX_k)}} \quad (1)$$

3.2.2. Decision trees

DT are a machine learning method employed for addressing classification and regression problems. These trees analyze a dataset, creating a hierarchical structure where decisions are made in a step-by-step manner. This hierarchical tree begins with the root node and branches into different paths, with each node representing the value of a specific feature or variable [18]. The dataset undergoes processing within this tree structure, with data being divided into subgroups, and decisions being made at each node. These decisions are utilized to classify data points or predict target variables. DT employ various criteria to select attributes that best encapsulate the information within the data. Some of these criteria include concepts like entropy, gain rate, and the Gini index [23]. DT can perform effectively, particularly when dealing with nonlinear data or complex relationships.

3.2.3. Random forest

RF is considered a collective machine learning technique, effectively used in data mining and classification problems. RF primarily comprises decision trees. The operational principle of RF involves partitioning the dataset into random subsets and creating a distinct decision tree for each subset. These decision trees are trained independently and are utilized for classifying data points or predicting target variables. The final forecast is made by averaging the votes or forecasts of these decision trees [24].

One of the key advantages of RF is that it can handle the complexity of the dataset and model non-linear relationships. This allows RF to work particularly well with complex and high-dimensional datasets. Moreover, RF can reduce the noise in the data and avoid the problem of overfitting. Therefore, RF is a widely preferred algorithm for classification and regression problems [25].

3.2.4. Gradient boosting decision tree

The Gradient Boosting Decision Tree (XGBoost) is a collective learning method that successfully solves both classification and regression problems. This algorithm is built by combining multiple decision trees, but these trees are not randomly generated; rather, each tree focuses on correcting the errors of the previous trees [26]. This method involves an iteration process in which trees are built sequentially. Each new tree is trained to reduce the errors of the predictions of the previous trees. XGBoost is especially preferred for problems that require high performance and work with large datasets. This method provides high accuracy in processing complex data and provides an effective regularization mechanism to prevent overfitting [27].

3.2.5. Naïve bayes

Naive Bayes (NB) is a machine learning algorithm based on Bayes theorem and used to solve classification problems. This algorithm is basically a probabilistic classifier, meaning that it uses probability rules to calculate the probability that an object or data belongs to a certain class. Naive Bayes is called "naive" (pure or simple) because it is an algorithm whose assumptions are quite simple and independent.

One of the advantages of the algorithm is that once trained, it can accurately classify data even on small datasets. This is especially important in applications with limited data. The fundamental principle underlying NB assumes that each data feature independently influences the class label, and the product of these influences provides the class prediction [28].

3.2.6. K-Nearest neighbor

KNN algorithm is a supervised learning algorithm for classifying an observation or predicting a value. The basic idea is that the class to which an observation belongs is determined based on the classes of its nearest neighbors. This closeness is typically calculated by Euclidean distance, shown in Equation 2, or other similar distance metrics.

The logic of the algorithm is quite simple. First, for a given observation, K nearest neighbors need to be found. These neighbors are determined by the distance metric around the observation. KNN looks at the class labels of these neighbors and assigns the class label of the majority as the predicted class.

KNN is the algorithm of choice for small datasets or entry-level classification problems due to its simple working principle and easy comprehensibility. However, it should be applied with caution when used on large datasets due to its computational complexity [29].

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\sqrt{(y_i-x_i)^2}} \quad (2)$$

3.2.7. Support vector machine

SVM is an efficient classification and regression algorithm for high-dimensional datasets. Its basic principle is based on separating and classifying data by creating a hyperplane (bounding) in an n-dimensional space. This hyperplane has the same dimension as the number of independent variables (n) and acts as a classification discriminator by dividing data points into two classes [30].

A good SVM model tries to follow this hyperplane, which has the greatest distance to the nearest training data points. This distance affects the generalization ability of the model. The greater the distance, the more likely the model is to perform better on new data. Therefore, SVM aims to optimally separate data points.

Furthermore, SVM offers the ability to transform data into another space using kernel functions. This is useful for handling situations where data cannot be separated linearly or where a space is needed that will provide a better separation.

3.3. Data validation method

In this research, training and test data validation method is used to detect credit card fraud and suspicious

transactions. This method aims to evaluate the performance of the model by dividing the dataset into two main parts. The first part, the "Training Dataset", is used in the learning phase of the model. The model is trained on this dataset and learns the patterns and relationships between the data. The second part, the "Test Dataset", is used to evaluate the performance of the model. It shows how the model reacts to data outside this dataset.

Figure 1 shows the flowchart of the model created in accordance with the training and test validation method. The aim is to identify issues with the model and assess its performance on real-world data, thereby contributing to the acquisition of more reliable results through the maintenance of a balanced approach between the training and testing phases [31].

3.4. Performance metrics

Conventional techniques for evaluating machine learning classifiers employ measurements that establish a connection between the level of confusion and the disparities between the ground truth data and the model's predictions, with TP, TN, FP, and FN representing true positives, true negatives, false positives, and false negatives, in that order.

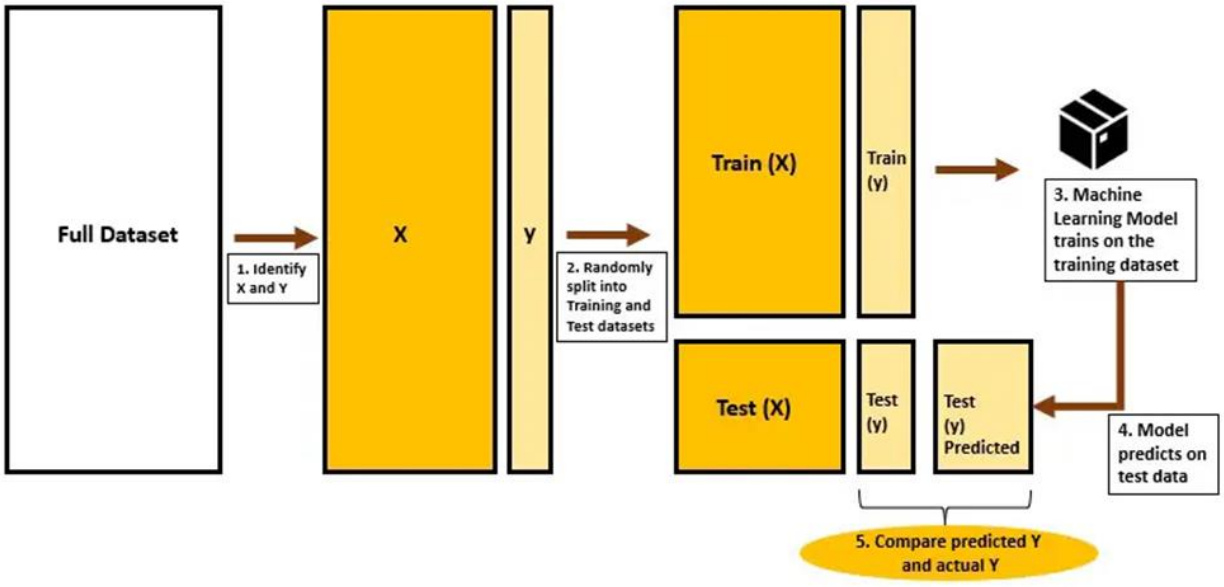

Figure 1. Training and testing validating method.

3.4.1. Accuracy

Accuracy is employed to assess the effectiveness in the retrieval and processing of data in the evidence domain. The Equation 3 can be used to express the proportion of correctly classified outcomes as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

3.4.2. Precision

Precision is a performance metric that quantifies the proportion of accurately identified positive cases among the total number of identified positive cases. This can be illustrated as shown in Equation 4.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

3.4.3. Recall

Recall, often referred to as sensitivity, represents the ratio of relevant instances successfully retrieved out of the total number of retrieved instances. This can be described as shown in Equation 5.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

3.4.4. F-Measure/F1-Score

The f-measure considers both precision and recall. It can be thought of as the weighted average of all values, as shown in Equation 6.

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (6)$$

3.4.5. Receiver operating characteristic curve

The Receiver Operating Characteristic Curve (ROC) is a graphical representation of the performance of a classification model at all classification thresholds. The ROC curve has two parameters: True Positive Rate (TPR) and False Positive Rate (FPR). Lowering the classification threshold results in more items being classified as positive [32].

TPR is calculated using the formula expressed in Equation 7.

$$TPR = \frac{TP}{TP + FN} \quad (7)$$

FPR, the False Positive Rate, is defined as shown in Equation 8.

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

3.4.6. The area under the curve

The Area Under the Curve (AUC) of the ROC Curve is expressed as the area beneath the ROC Curve. The AUC value varies between 0 and 1. A model with a 100% prediction error rate has an AUC of 0. A model with 100% correct predictions has an AUC of 1 [33].

3.4.7. Area under the precision-recall curve

Area Under the Precision-Recall Curve (AUPRC) is an important metric used to evaluate the performance of a classification model, especially in scenarios with imbalanced classes. It quantifies the trade-off between precision (the ratio of true positives to all positive predictions) and recall (the ratio of true positives to all actual positives) as the classification threshold changes. AUPRC represents the area under the precision-recall curve, which ranges from 0 to 1. A higher AUPRC value indicates better model performance, with better classification of both positive and negative samples. This metric helps fine-tune models to achieve a balance between precision and recall, ultimately improving overall classification performance.

3.5. Dataset

The dataset consists of credit card transactions made by European cardholders in September 2013. It spans two days and covers a total of 284,807 transactions, of which 492 were fraudulent [34]. Notably, this dataset exhibits a significant class imbalance, with fraudulent transactions representing only 0.172% of the total. Machine learning is faced with obstacles when dealing with imbalances in datasets, as there is a notable disparity in the distribution of classes. This discrepancy can create bias towards the dominant class when training models using these datasets, resulting in inadequate identification of patterns related to the minority class. This can hinder the ability to apply learned patterns to new data, particularly for the less represented class, and standard accuracy measurements may provide deceptive results.

The dataset consists solely of numerical input variables obtained through Principal Component Analysis (PCA), with features V1 to V28 representing PCA-derived principal components. “Time” and “Amount” are the only features that have not been transformed using PCA. “Time” represents the elapsed time in seconds since the first transaction, while “Amount” denotes the value of the transaction. The “Class” feature serves as the response variable, taking a value of 1 for fraud cases and 0 for non-fraud cases. Given the class imbalance, accuracy is best assessed using the AUPRC. To comprehend the data and understand the significance of each feature in relation to the model's predictions, a bar chart is used. This bar chart, depicted in Figure 2, provides insight into the features that have a more significant impact on the outcomes of the model, helping to understand the relationships between the features and the performance of the model.

The correlation table, illustrating the relationships between each feature in the dataset, is presented in Figure 3.

Table 1 provides a list of features and their descriptions, which are crucial in the detection of transactions classified as fraudulent, as evident from the information presented in Figure 2 and Figure 3.

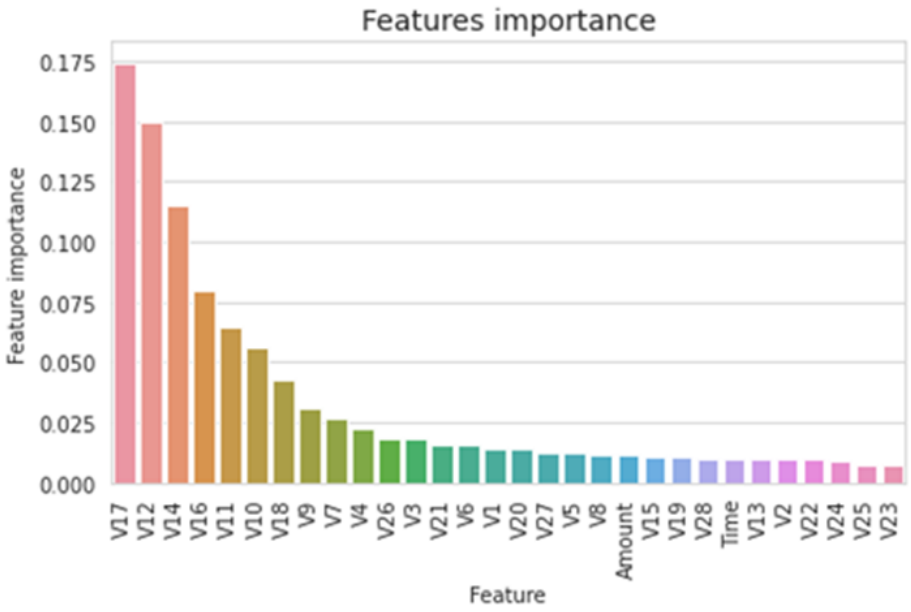

Figure 2. Features importance.

3.6. Data preparation

If the dataset is imbalanced, meaning one class has significantly more examples than the other, model training can become biased and yield misleading results. Therefore, data balancing is necessary. In the dataset used for this research, the overwhelming majority of transactions are non-fraudulent. Consequently, some initial data preprocessing was performed.

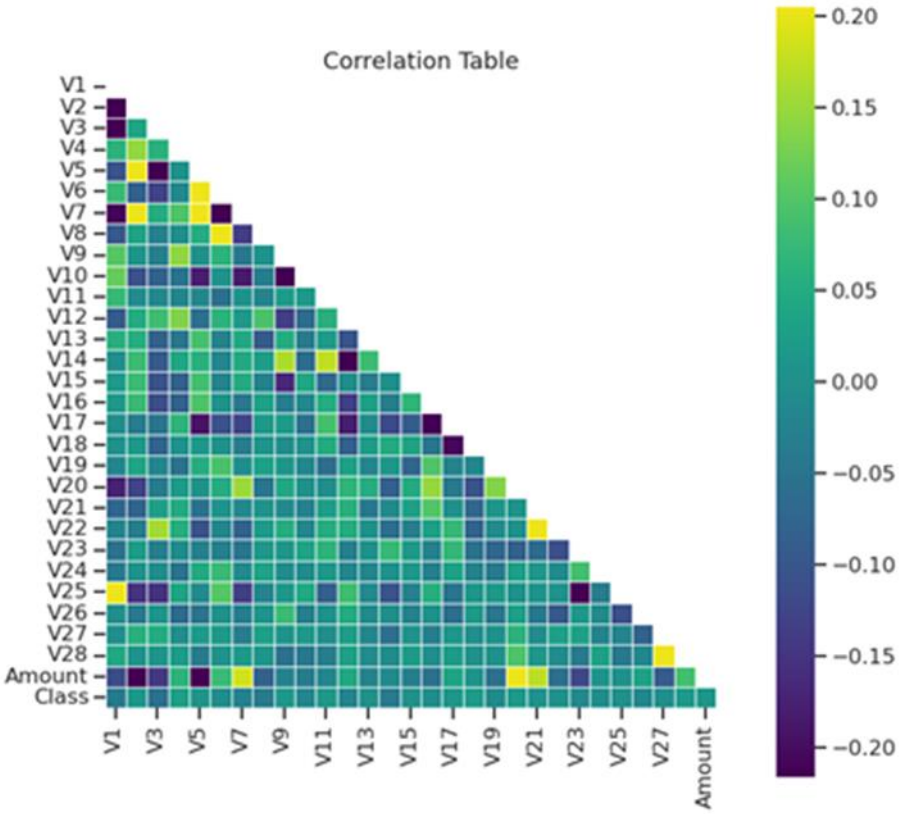

Figure 3. Correlation table.

Table 1. Features and descriptions.

Features	Description
ID	A distinct identifier for each row
Time	The duration in seconds between the current transaction and the first transaction in the dataset
V1-V28	Characteristics resulting from dimensionality reduction, applied to safeguard user identities and sensitive features
Amount	The transaction value
Class	Target category (1 for fraudulent transactions, 0 for legitimate transactions)

3.6.1. Scaling and distribution

In the specified phase of our study, the 'Time' and 'Amount' columns were standardized to the same scale as the other columns, employing Z-score standardization. Z-score standardization, also referred to as Z-score normalization or zero-mean normalization, is an essential statistical technique utilized to convert a numerical variable into a standard normal distribution with a mean of 0 and a standard deviation of 1. This method proves highly valuable in the realm of data preprocessing for machine learning and statistical analysis. The process of Z-score standardization can be described in the following steps, which guarantee the features are on a unified scale, enabling direct comparison and avoiding the dominance of any individual variable in the modeling process.Data collection: Starting with the raw data, which includes the “Time” and “Amount” columns, as well as other relevant features.

1. Compute the mean and standard deviation: Calculate the mean (μ) and standard deviation (σ) for both the “Time” and “Amount” columns. The mean denotes the average value, while the standard deviation quantifies the spread and variability of the data.
2. Z-score transformation: The Z-score (z) for each data point (x) in the "Time" and "Amount" columns is calculated using the Z-score (Equation 9):

$$Z - score(z) = \frac{\pi - \mu}{\sigma} \quad (9)$$

Equation 9 utilizes z as the standardized representation, x as the initial data point, and μ (mu) and σ (sigma) as symbols for the column's mean and standard deviation respectively. By applying this computation to every data point, novel figures are generated for the columns of “Time” and “Amount”, guaranteeing their means approximate 0 and their deviations approximate 1. The collective influence of feature scaling ultimately leads to heightened model efficacy and increased precision in identifying fraudulent activity.

3.6.2. Random under-sampling

At this stage, the 'Random Subsampling' technique is used to address the problem of class imbalance in the dataset. The primary goal is to prevent overfitting in machine learning models and ensure their effective performance. Below is a step-by-step description of the process:

1. The degree of class imbalance in the dataset was first assessed by determining the number of samples for each class label, for both fraudulent transactions (Fraud = "1") and non-fraudulent transactions.
2. Once the number of samples for fraudulent transactions was determined, the dataset was balanced by setting the number of non-fraudulent transactions equal to the number of fraudulent transactions to achieve a 50/50 balanced ratio. The aim is to ensure that if there are 492 cases of fraud, there are an equal number of 492 cases of non-fraudulent transactions.
3. To apply this technique, a sub-sample of the dataset was created to ensure a balanced 50/50 ratio between the two classes. This sub-sample consists of only 492 fraud instances and 492 non-fraudulent transaction instances.
4. Performing data shuffling is crucial to guarantee the reliability of the model's performance. Shuffling helps to eliminate possible biases or patterns in the dataset that could affect the training and evaluation of the model.

By following these steps, not only is the problem of class imbalance addressed, but the data is organized in such a way that machine learning models can generalize effectively and make accurate predictions.

3.6.3. Dimensionality reduction and clustering

The process of dimensionality reduction and clustering is performed to increase the understanding of the underlying structure of the dataset, to streamline the data and to distinguish patterns or clusters within it. At this stage, the t-distributed stochastic neighbor embedding (t-SNE) method is used, which serves as a technique for data visualization and dimensionality reduction. This method takes high-dimensional data points and transforms them into a lower-dimensional space to improve our understanding of the structure of the data. t-SNE is particularly used for visualizing data by preserving similarities and dissimilarities between data points. The primary objective is to simplify the dataset and, in doing so, improve the applicability of various machine learning techniques for fraud detection. Specifically, t-SNE excels in clustering both fraudulent and non-fraudulent cases, as evidenced by the results depicted in Figure 4. This observation remains consistent across various scenarios, even after randomizing the dataset. Essentially, this implies that applying additional predictive models is likely to be successful in effectively differentiating between fraud and non-fraud cases.

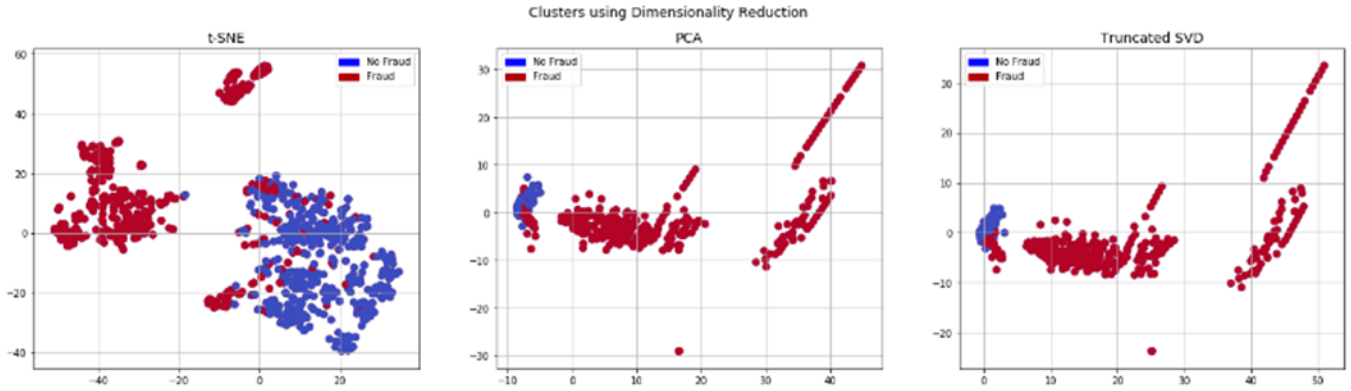

Figure 4. Clusters using dimensionality reduction.

4. Experimental study and results

In this study, supervised classification algorithms such as LR, DT, RF, XGBoost, NB, KNN, and SVM algorithms were employed to detect fraudulent transactions in credit card transactions. When creating the model, the dataset was divided into two parts: 80% for training and 20% for testing. In all the algorithms used, the random state was set to 42. The number of trees in the RF was defined as 12. In KNN, the number of neighbors was set to 5. In the XGBoost algorithm, the learning rate was determined as 0.01, the number of predictors was set to 10, and the number of seeds was 25. In the SVM, the kernel used was the Radial Basis Function (RBF), and the C parameter was set to 2. In SVM optimization, the C parameter indicates the degree to which misclassification of each training sample will be avoided. Figure 5 and Figure 6 presents the confusion matrices for the utilized algorithms.

When examining Figure 5 and Figure 6, confusion matrices were employed as a crucial tool to analyze the performance of each algorithm in a detailed manner. Notably, the RF algorithm excelled in correctly identifying non-fraudulent transactions, demonstrating the highest True Negative value. On the KNN algorithm exhibited remarkable performance in correctly classifying transactions carrying signs of fraud, boasting the highest True Positive value. On the other hand, the NB algorithm displayed a tendency to have higher False Negative values, which could potentially increase the risk of failing to identify fraudulent transactions. Other algorithms exhibited similar performance characteristics.

When examining Figure 7, the graph presents the AUC-ROC (Area Under the Receiver Operating Characteristic) and AUPRC scores of different machine learning algorithms. These scores serve as critical metrics for evaluating the performance of classification models. AUC-ROC measures the trade-off between the false positive rate and the true positive rate. It reflects a model's ability to accurately classify non-fraudulent transactions while minimizing false positives. A higher

AUC-ROC score indicates a better overall classification performance. On the other hand, AUPRC assesses the balance between recall and precision and is particularly useful for imbalanced classification tasks. High AUPRC scores indicate a model's ability to effectively detect fraudulent transactions and minimize false positives.

The graph shows that the RF algorithm exhibits the highest AUC-ROC and AUPRC scores, demonstrating its outstanding performance in accurately identifying legitimate transactions and effectively detecting fraudulent ones. Similarly, the KNN algorithm shows remarkable proficiency in correctly classifying fraudulent transactions, as evidenced by its superior AUC-ROC and AUPRC scores. In contrast, the NB algorithm shows relatively lower AUC-ROC and AUPRC scores, indicating potential limitations in accurately detecting fraudulent transactions.

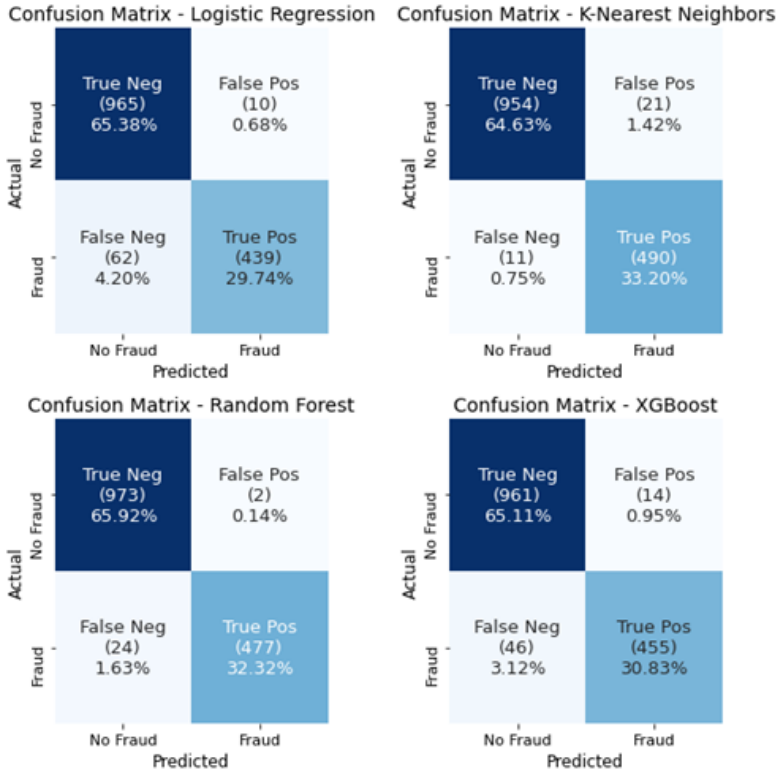

Figure 5. Confusion matrices.

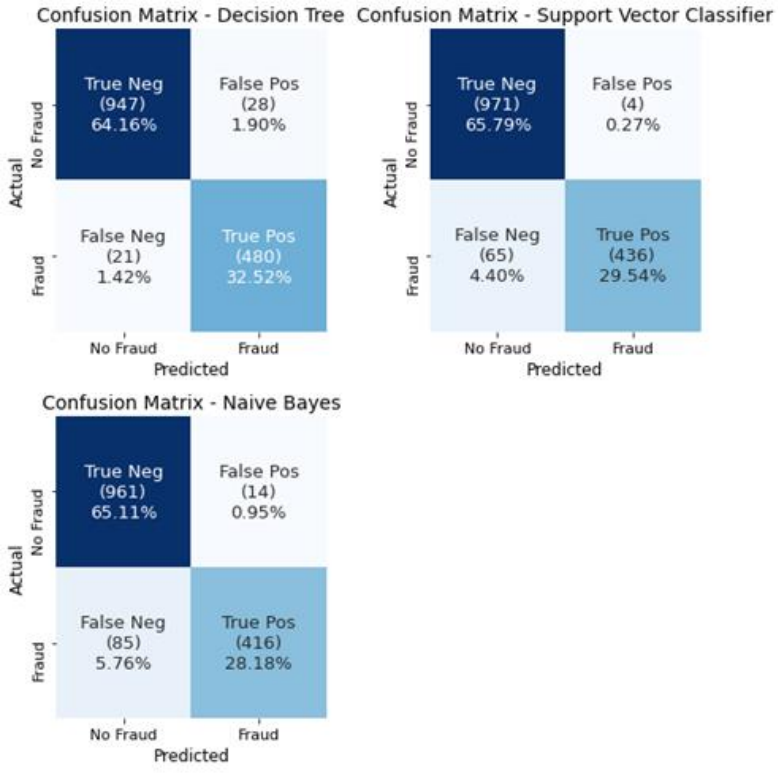

Figure 6. Confusion matrices.

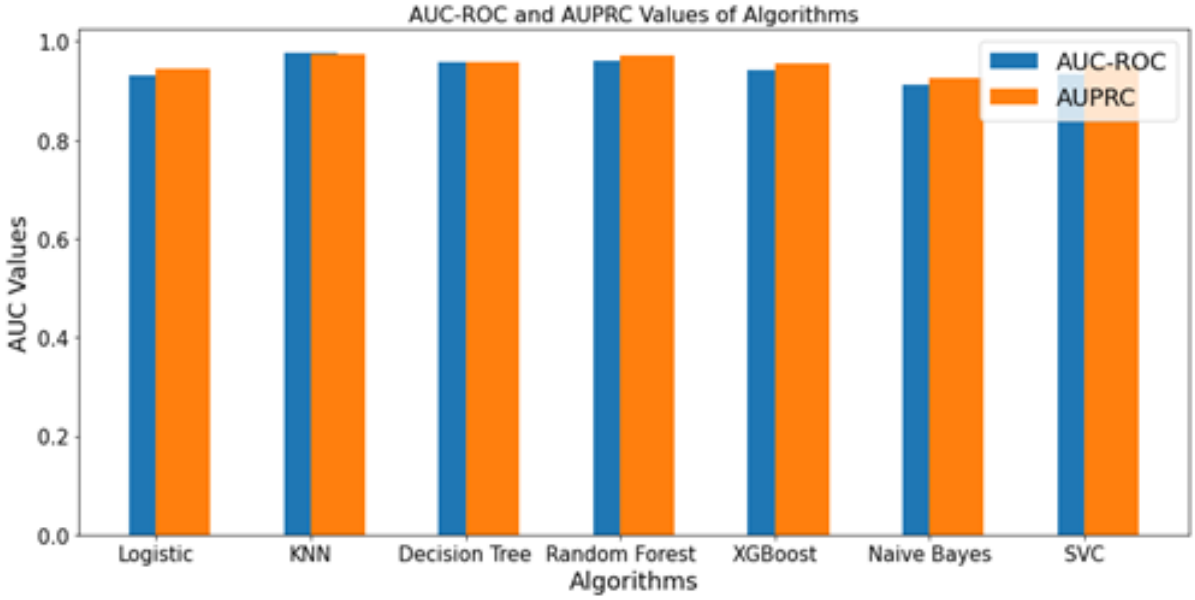

Figure 7. AUC-ROC and AUPRC values of algorithms.

Table 2 shows that two algorithms, RF and KNN, show strong potential in this area. However, some other algorithms also show performance metrics close to RF and KNN, indicating that there is no single solution for credit card fraud detection and that the choice of algorithm may depend on specific operational requirements and priorities.

Table 2. Performance metrics.

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUPRC
LR	0.95	0.97	0.87	0.92	0.93	0.94
KNN	0.97	0.96	0.96	0.96	0.97	0.97
DT	0.96	0.94	0.95	0.95	0.96	0.95
RF	0.97	0.99	0.94	0.96	0.97	0.98
XGBoost	0.96	0.98	0.91	0.94	0.94	0.95
NB	0.94	0.96	0.86	0.90	0.91	0.92
SVC	0.95	0.98	0.88	0.93	0.93	0.95

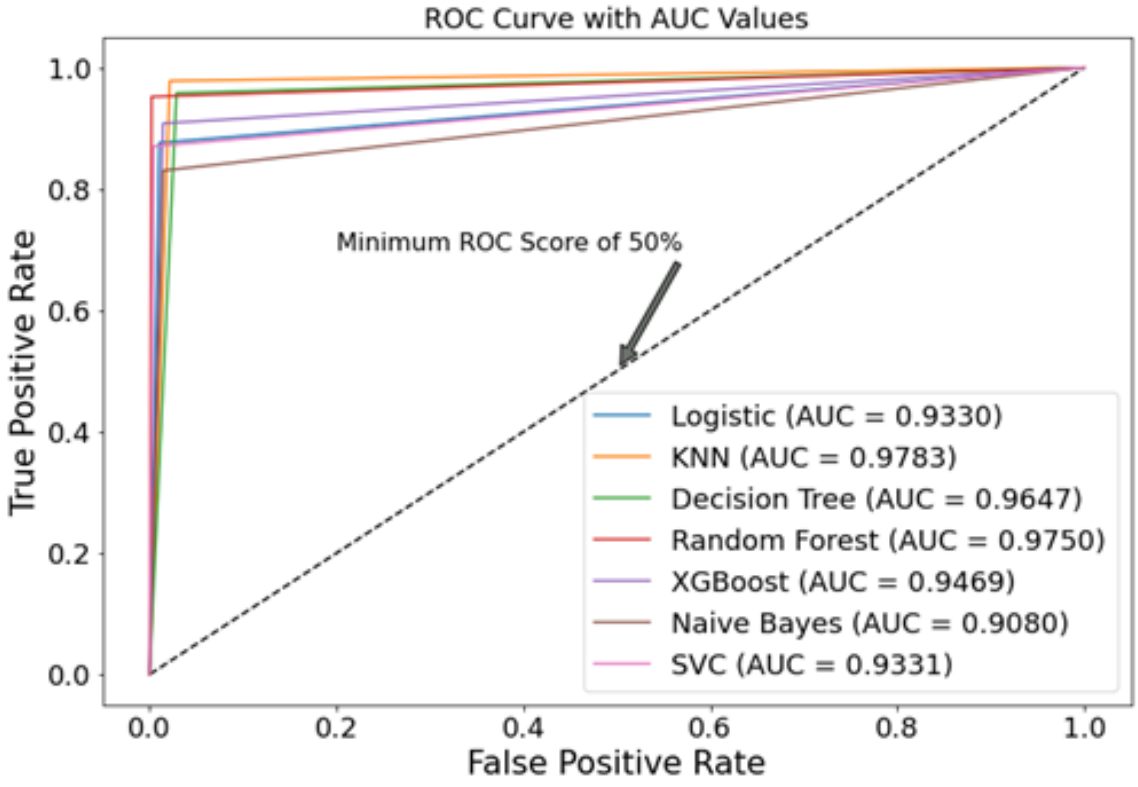

Figure 8. ROC curve with AUC values.

RF exhibits an accuracy rate of %97, indicating its flexible classification capability. This corresponds to an AUC-ROC score of 0.97 and an AUPRC score of 0.98, highlighting its expertise in distinguishing between fraudulent and non-fraudulent transactions (F1-Score = 0.96). These AUC values underline the likely utility of the model in detecting credit card fraud. Moreover, the concordant performance indicators suggest that RF can deliver reliable predictions even in situations involving imbalanced datasets, which is a common obstacle in fraud detection.

KNN excels in correctly distinguishing fraudulent transactions with a 97% accuracy rate and AUC-ROC score of 0.97 and AUPRC score of 0.97 (F1-Score = 0.96). The high accuracy rate and high AUC values emphasize that KNN is a powerful algorithm for credit card fraud detection.

It is worth noting that the NB algorithm shows relatively lower performance metrics compared to the other algorithms, with 94% accuracy. The precision (0.96) and recall (0.86) values, while not significantly lower, are relatively less competitive when considered in the context of the task (F1-Score = 0.90). This highlights the limitations of this algorithm in achieving accuracy levels comparable to other algorithms, especially in situations where higher precision and recall rates are crucial.

When the ROC curve plot in Figure 8 is analyzed, it is seen that the classifiers exhibit a high performance. The ROC curve provides a visual representation of how a model's sensitivity (true positive rate) and specificity (true negative rate) change at various thresholds.

The two prominent algorithms in this context are KNN and RF. Both algorithms achieved very high AUC scores, emphasizing their proficiency in data classification. KNN shows an AUC score of 0.97, emphasizing its effectiveness in correctly identifying cases while maintaining the balance between precision and recall. RF, on the other hand, exhibits an equal AUC score of 0.97, emphasizing its superior performance in classification.

4. Discussion

Researchers in the fraud detection field are constantly exploring and debating the effectiveness of different methodologies. Within the framework of this discourse, this section explores the comparison of the current study findings with previous studies. Employing preprocessing techniques such as scaling, random under-sampling, dimensionality reduction, and clustering can enhance fraud detection rates when comparing study findings with research on the same dataset using classification algorithms [35] and [36]. Additionally, classifying algorithms were proven to be as successful as deep learning algorithms, consistent with papers [37] and [38]. However, papers [39] and [40] advocate for deep learning algorithms as optimal, but the decision should be situation dependent. Deep networks outperform with larger datasets and exhibit versatility across various domains, thereby conferring them with distinct advantages. Nevertheless, in contexts where the emphasis lies on interpretability and cost-effectiveness rather than the intricacy of deep learning models, classification algorithms arise as a more pragmatic option. The straightforwardness and efficient utilization of resources by classification algorithms render them highly suitable for scenarios with constrained computational resources or where a transparent comprehension of the decision-making process is imperative [41]. These insights actively contribute to ongoing discussions regarding the selection of appropriate methodologies in the field of fraud detection, providing valuable considerations for real-world applications.

5. Conclusion

This study meticulously examined seven supervised classification algorithms, namely LR, DT, RF, XGBoost, NB, KNN, and SVM, to identify the most effective machine learning algorithms in the field of credit card fraud detection. After a thorough review, the evaluation of these algorithms to detect instances of credit card fraud and suspicious transactions yielded remarkable results and successfully achieved the goal of the study.

The data preparation phase in this study played a crucial role in guaranteeing the reliability and precision of the subsequent analysis. To address the issue of class imbalance, it becomes imperative to perform a preliminary data balancing process to overcome the uneven distribution between non-fraudulent and fraudulent transactions. Improving model performance and accuracy in fraud detection requires significant adjustments in scale and distribution, as well as the creation of a balanced sub-sample. A technique called "random subsampling" was used to ensure fair representation of both types of transactions and to minimize class imbalance and overfitting. Furthermore, the application of clustering and dimensionality reduction techniques increased our understanding of the structure of the dataset and facilitated the implementation of various machine learning algorithms. Among these algorithms, the t-SNE algorithm exhibited exceptional precision in categorizing fraudulent and non-fraudulent samples.

The RF algorithm proved to be the most efficient algorithm based on performance measurements. It exhibits exceptional accuracy, as evidenced by its remarkable AUC-ROC and AUPRC scores. It demonstrates a remarkable ability to correctly identify legitimate transactions while minimizing false positives, an important consideration in fraud detection. Alongside RF, KNN demonstrates exceptional proficiency in accurately classifying transactions that show signs of fraud, as evidenced by the highest true positive values. Its precise identification of fraudulent transactions positions it as an important algorithm for credit card fraud detection. Furthermore, the NB algorithm is relatively less competitive when considered in the context of the task.

In a broader context, these findings play an important role in ongoing efforts to combat financial fraud and guarantee the protection of customer and corporate assets. The selection of an algorithm holds immense significance, driven by data characteristics, the

delicate equilibrium between sensitivity and accuracy, and the comparative expenses of false positives and false negatives. Depending on specific needs and priorities, the optimal approach for a company and its customers varies. If the main goal is to reduce false positives and effectively detect legitimate transactions, it is advantageous to use the RF algorithm. Alternatively, if precise identification of fraudulent transactions is crucial, particularly in cases where fraud indicators are present, the KNN algorithm may be better suited for this purpose. As the financial sector progresses, such research will provide invaluable perspectives and direction for the establishment of resilient fraud detection systems. It is evident that advancements are being made towards more secure financial transactions, and these outcomes denote a substantial progression in that trajectory. This has the potential to unlock opportunities for further exploration and pragmatic applications in the industry.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Akers, D., Golter, J., Lamm, B., & Solt, M. (2005). Overview of recent developments in the credit card industry. FDIC Banking Review, 17, 23-35.
2. Heggestuen, J. (2020). Credit-card fraud surges 35% as coronavirus freezes the economy and wipes out jobs. Business Insider. https://markets.businessinsider.com/news/stocks/credit-card-account-fraud-skyrockets-coronavirus-pandemic-recession-economy-layoffs-2020-5-1029246107
3. Çalışkan, M. A. (2021). Credit card fraud in Turkey increased by 25% in 2020. Hürriyet. https://www.hurriyet.com.tr/haberleri/kredi-karti-dolandiriciligi
4. Bhatla, T. P., Prabhu, V., & Dua, A. (2003). Understanding credit card frauds. Cards Business Review, 1(6), 1-15.
5. Şenel, S. A., & Arslan, Ö. (2019). The role of forensic accounting profession in preventing the accounting scandals. Cumhuriyet University Journal of Economics and Administrative Sciences, 20(1), 293-308
6. Tripathi, K. K., & Pavaskar, M. A. (2012). Survey on credit card fraud detection methods. International Journal of Emerging Technology and Advanced Engineering, 2(11), 721-726.
7. Sevli, O. (2022). Kredi kartı dolandırıcılığının yapay sinir ağları kullanılarak tespiti. 11th International Conference on Applied Sciences, 233-240. Academy Global Publishing House.
8. Joo, S. H., Grable, J. E., & Bagwell, D. C. (2003). Credit card attitudes and behaviors of college students. College Student Journal, 37(3), 405-420.
9. Fogarty, T. C., Ireson, N. S., & Battle, S. A. (1992). Developing rule-based systems for credit-card applications from data with the genetic algorithm. IMA Journal of Management Mathematics, 4(1), 53-59. https://doi.org/10.1093/imaman/4.1.53
10. Raj, S. B. E., & Portia, A. A. (2011). Analysis on credit card fraud detection methods. In 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET), 152-156. https://doi.org/10.1109/ICCCET.2011.5762457
11. Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. Procedia Computer Science, 165, 631-641. https://doi.org/10.1016/j.procs.2020.01.057
12. Yee, O. S., Sagadevan, S., & Malim, N. H. A. H. (2018). Credit card fraud detection using machine learning as data mining technique. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(1-4), 23-27.
13. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953
14. Jha, S., Guillen, M., & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. Expert Systems with Applications, 39(16), 12650-12657. https://doi.org/10.1016/j.eswa.2012.05.018
15. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555
16. Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915-4928. https://doi.org/10.1016/j.eswa.2014.02.026
17. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613. https://doi.org/10.1016/j.dss.2010.08.008
18. Pulat, M., & Deveci, I. (2021). Bibliometric Analysis of Theses Published on Machine Learning and Decision Trees in Turkey. Journal of Management and Economics, 28(2), 287-308.
19. Albayrak, A. S., & Yilmaz, S. K. (2009). Veri Madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama. Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences, 14(1), 31-52.
20. Akça, M. F., & Sevli, O. (2022). Predicting acceptance of the bank loan offers by using support vector machines. International Advanced Researches and Engineering Journal, 6(2), 142-147. https://doi.org/10.35860/iarej.1058724
21. Bircan, H. (2004). Logistic regression analysis: An application on medical data. Kocaeli University Journal of Social Sciences, 8, 185-208.
22. Yavuz, A., & Çilengiroğlu, Ö. V. (2020). Lojistik regresyon ve CART yöntemlerinin tahmin edici performanslarının yaşam memnuniyeti verileri için karşılaştırılması. Avrupa Bilim ve Teknoloji Dergisi, (18), 719-727.

https://doi.org/10.31590/ejosat.691215
23. Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014). An application on computer and internet security with decision tree algorithms in data mining. Journal of Industrial Engineering, 25(3), 2-19.
24. Türk, S. T., & Balçık, F. (2023). Rastgele orman algoritması ve Sentinel-2 MSI ile fındık ekili alanların belirlenmesi: Piraziz Örneği. Geomatik, 8(2), 91-98. https://doi.org/10.29128/geomatik.1127925
25. Akar, Ö., & Güngör, O. (2012). Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, 1(2), 139-146. https://doi.org/10.9733/jgg.241212.1t
26. Alshari, H., Saleh, A. Y., & Odabaş, A. (2021). Comparison of gradient boosting decision tree algorithms for CPU performance. Journal of Institue of Science and Technology, 37(1), 157-168.
27. Şahin, E. M., Sahin, S., & Tanağardıgil, İ. (2021). Battery State of Health and Charge Estimation Using Machine Learning Methods. Avrupa Bilim ve Teknoloji Dergisi, (26), 389-394. https://doi.org/10.31590/ejosat.959630
28. Zhang, H., & Li, D. (2007). Naïve Bayes text classifier. In 2007 IEEE international conference on granular computing (GRC 2007), 708-711. https://doi.org/10.1109/GrC.2007.40
29. Yong, Z., Youwen, L., & Shixiong, X. (2009). An improved KNN text classification algorithm based on clustering. Journal of Computers, 4(3), 230-237.
30. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their applications, 13(4), 18-28. https://doi.org/ 10.1109/5254.708428
31. Polyzotis, N., Zinkevich, M., Roy, S., Breck, E., & Whang, S. (2019). Data validation for machine learning. Proceedings of Machine Learning and Systems, 1, 334-347.
32. Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: point estimates and confidence intervals. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13, 451-466. https://doi.org/10.1007/978-3-642-40994-3_29
33. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 4(11), 218-225. https://doi.org/10.21037/atm.2016.03.37
34. MLG-ULB. (2017). Credit Card Fraud Detection. Kaggle. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
35. Mishra, A., & Ghorpade, C. (2018). Credit card fraud detection on the skewed data using various classification and ensemble techniques. In 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 1-5. https://doi.org/10.1109/SCEECS.2018.8546939
36. Navamani, C., & Krishnan, S. (2018). Credit card nearest neighbor based outlier detection techniques. International Journal of Computer Techniques, 5(2), 56-60.
37. Kazemi, Z., & Zarrabi, H. (2017). Using deep networks for fraud detection in the credit card transactions. In 2017 IEEE 4th International conference on knowledge-based engineering and innovation (KBEI), 630-633. https://doi.org/10.1109/KBEI.2017.8324876
38. Dhankhad, S., Mohammed, E., & Far, B. (2018). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In 2018 IEEE international conference on information reuse and integration (IRI), 122-125. https://doi.org/10.1109/IRI.2018.00025
39. Wang, C., Wang, Y., Ye, Z., Yan, L., Cai, W., & Pan, S. (2018). Credit card fraud detection based on whale algorithm optimized BP neural network. In 2018 13th international Conference on Computer Science & Education (ICCSE), 1-4. https://doi.org/10.1109/ICCSE.2018.8468855
40. Pumsirirat, A., & Liu, Y. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. International Journal of Advanced Computer Science and Applications, 9(1), 18-25.
41. Sarızeybek, A. T., & Sevli, O. (2022). Makine Öğrenmesi Yöntemleri ile Banka Müşterilerinin Kredi Alma Eğiliminin Karşılaştırmalı Analizi. Journal of Intelligent Systems: Theory and Applications, 5(2), 137-144. https://doi.org/10.38016/jista.1036047