ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

A PROPOSED FRAMEWORK OF DIMENSIONALITY REDUCTION TECHNIQUES TO BOOST CREDIT CARD FRAUD CLASSIFICATION

[1]Dr. JITENDRA SHEETLANI

[1]Assistant Professor, Medi-Caps University, Indore, Madhya Pradesh, India.

[2]Dr. HARSH PRATAP SINGH

[2]Assistant Professor, Medi-Caps University, Indore, Madhya Pradesh, India.

[3]Dr. PANKAJ MALIK

[3]Assistant Professor, Medi-Caps University, Indore, Madhya Pradesh, India.

[4]Dr. RAJESH KUMAR VISHWAKARMA

[4]JNS College Sohagpur district –Hoshangabad (M.P.) India.

[5]Dr. ABDUL RAZZAK KHAN QUERSHI

[5]Assistant Professor, Medi-Caps University, Indore, Madhya Pradesh, India.

ABSTRACT

Machine learning techniques are widely employed in the modern world for prediction and classification jobs. Sentiment analysis, disease detection, network intrusion detection, and many other uses fall within the broad category of machine learning classification. The main topic of this essay is the use of machine learning to detect credit card fraud. Machine learning algorithms will be used exclusively in this study's assessment and forecasting of credit card fraud. These algorithms will use feature selection approaches. The suggested study will demonstrate how feature selection could raise the classification systems' level of precision. By using dimensionality reduction approaches, this work investigates the application of enhanced Nave Bayes, K-nearest neighbour, random forest, and logistic regression on highly skewed credit card extortion data.. The feature selection method is used in this study to reduce the number of dimensions. A dataset of credit card trades containing 284,807 exchanges was given by European cardholders. A method's effectiveness is evaluated using its accuracy, affectability, precision, and specificity. The results demonstrate that K-Nearest Neighbor (KNN), Logistic Regression Classifier, Random Forest (RF), and Naive Bayes had the highest accuracy rates in the field, at 97.50%, 99.96%, and 99.95%, respectively.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Keywords: Credit Card, Fraud Detection, Machine Learning, Dimension Reduction.

1. INTRODUCTION

One of the current issues that is increasing is financial extortion, which regularly poses a threat and has serious repercussions for businesses, corporations, associations, and the government. Credit card fraud is another illegal activity involved in this financial extortion that is seriously harming the banking sector [1]. The expansion of credit card exchanges has boosted online reliability. Credit card fraud is increasing as credit card exchanges become the most popular method of payment for both online and offline transactions. Internal card extortion and external card misrepresentation are both forms of credit card fraud. External card extortion involves using a stolen credit card to obtain funds through dubious means, while internal card extortion occurs as a result of an agreement between cardholders and the bank using a fictitious identity to submit misrepresentation. Numerous investigations have been conducted about the development of external card misrepresentation, which accounts for the majority of credit card fraud. Big data problems cannot be solved by human methods since it is highly difficult and time-consuming to identify fraud using conventional techniques like manual prediction. Financial institutions have focused their attention on future computational solutions to address the credit card extortion issue. One of the most effective methods for addressing the problem of credit extortion discovery is the data mining process. [2].

Credit card classification strategies divide exchanges into two categories: authentic (verifiable) and dishonest exchanges [3]. Systems of many various kinds, including SVM,[4], data mining [5], genetic algorithms (GA) [6], decision trees (DT) [7], artificial neural networks (ANN) [8], etc., have been created for credit card fraud detection approaches. There are several more analyses that are primarily conducted using logistic regression [9] and naive Bayes [10]. These days, di-mensionality reduction techniques are combined with classical classification algorithms so that they may quickly identify credit card extortion [11], while logistic regression and neural networks are linked to the problem of identifying credit card misrepresentation. Optimal feature (variable) selection for the models, suitable measurement to evaluate strategy execution on skewed credit card misrepresentation data, and rare and horribly excessive (or skewed) credit-card exchange datasets are only a few of the issues involved with credit card discovery. Because fake conduct profiles, in particular, are dynamic, fraudulent exchanges frequently resemble legitimate ones. How credit card extortion identification is carried out depends greatly on the kind of sample method used, the variables identified, and the location technique(s) employed. This study seeks to complete a comparable investigation of credit card misrepresentation detection using naive Bayes, k-nearest neighbour, and logistic regression methods on highly skewed data that depend on accuracy, affectability, and specificity.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

2. RELATED WORK

Some well-known methods for detecting credit card fraud include logistic models, Bayesian belief networks, neural networks, and decision trees. Each of these methods offers a different approach to the challenge of locating and classifying false data. In general, support vector machines, data mining, meta-learning, and neural networks are used to detect credit card fraud. Iyer et al.'s [12] paper describes the "Credit Card Fraud Detection Approach by Employing Hidden Markov Models (HMM)". By simulating the stages involved in a credit card transaction, this model demonstrates how hidden Markov models (HMMs) can be used to identify fraudulent transactions. The normal behaviour of cardholders was used to teach this. Credit Card Fraud Detection With Neural Network (NN) is the title of an essay written by S. Ghosh, Douglas L. Reilly, et al. The neural networks used in this method were trained on a substantial sample of labelled credit card account transactions to detect credit card fraud. Testing is conducted using holdout data, which consists of all account activity for the following two months. Neural networks were used to the trained data, which comprised information pertaining to misplaced, lost, or stolen cards, application cards, counterfeit fraud, and mail-order fraud. It has significantly detected more fraud accounts with fewer false positives—almost 20% less—than rule-based fraud detection methods. provided fraud detection methods in [14]. To distinguish between authentic and fraudulent transactions, this technique first employs region-based clustering of parameter values. This Gaussian mixture model is used to simulate the probability density of credit card users' previous activity, and it can be used to calculate the probability of current user behaviour by finding anomalies in prior behaviour.

As a last stage, Bayesian networks are used to characterise the statistics of a particular user and the statistics of other fraud situations. The Hilas and Mastorocostas [15] methodology is predicated on the client giving verifiable proof. The capacity of each profile to discriminate between honest use and fraud is tested using a feed-forward neural network (FF-NN) classifier. Panigrahi et al. (2009) [16] proposed a different approach for detecting credit card fraud that integrates confirmations of the various sorts of behaviour.

Kunal Goswami, Younghee Park, and Chungsik Song [17] have created feature sets that can be compared against cutting-edge feature sets in order to detect fraud. They use both the feature set and the user's social interactions on the Yelp website when determining if a user is involved in fraud. He employed neural networks to generate the F1 score, and the result was 0.95, which is comparable to all other widely used fraud detection methods. The effectiveness of the feature set is on par with existing fraud detection techniques. Masoumeh Zareapoor and Pourya Shamsolmoali [18] investigate how various classification algorithms function for detecting credit card fraud using confusion matrix parameters.

3. MATERIALS AND METHOD

A. Feature Selection Methods

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022**

Dimensions in the context of big data are essentially features or traits, and there are a ton of them [19]. Processing high-dimensional data is particularly challenging; hence, feature selection and feature extraction methods are used [20]. These techniques shrink the dimensions while preserving the information. Processing is made simple after using this strategy, and machine learning methods frequently boost performance as well. A distinguishing feature is a distinctive, quantitative norm for a technique. When a credit card is used, the transaction details, which contain a number of details (such as the credit card ID, the transaction amount, and so on), are recorded in the administration provider's database. Unambiguously, specific qualities have an impact on how a fraud location system operates. Successful classification results from the procedure known as feature determination, which involves picking a smaller selection of traits from a larger set. The pursuit space has a size of 2N, where N is the total number of features, and all possible feature subsets are contained in it. As a result, choosing features is an NP-hard problem. [21]. The alleged critique of dimensionality is that it may further impair the classifier's ability to handle the wide inquiry space because repeated and inconsequential features are not helpful for categorization.

B. Dataset

The dataset can be found starting with ULB Machine Learning in [22]. Credit card exchanges made by European cardholders in September 2013 were incorporated into the dataset. This dataset represents a total of 284,807 transactions over two days. The positive class (fraud instances) accounts for 0.172% of the exchange statistics. The dataset is out of balance and heavily biassed in favour of the positive class. The Principal Component Analysis (PCA) feature now only accepts numerical (continuous) input variables and uses a modified set of 28 main components. After then, a total of 30 input features are used in the research. It is not feasible to discuss the specifics and background information of the attributes due to confidentiality issues. The time feature in the dataset contains the number of seconds that passed between each exchange and the main exchange. The "sum" feature is the exchange sum. The characteristic termed "class," which accepts a value of 1 for a positive case (fraud) and 0 for a negative instance, serves as the objective class for the paired classification (non-fraud).

4. PROPOSED METHOD

The proposed work is divided into two main parts: feature determination, a dimensionality reduction strategy, and classification. The initial part of the suggested strategy divides the datasets, and a thorough wrapping technique is utilized to direct the selection of the finest and most practical qualities. The classification algorithm that is connected to the preprocessed dataset and assesses whether or not the trade is fraudulent makes up the second element. View the task's proposed flowchart.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

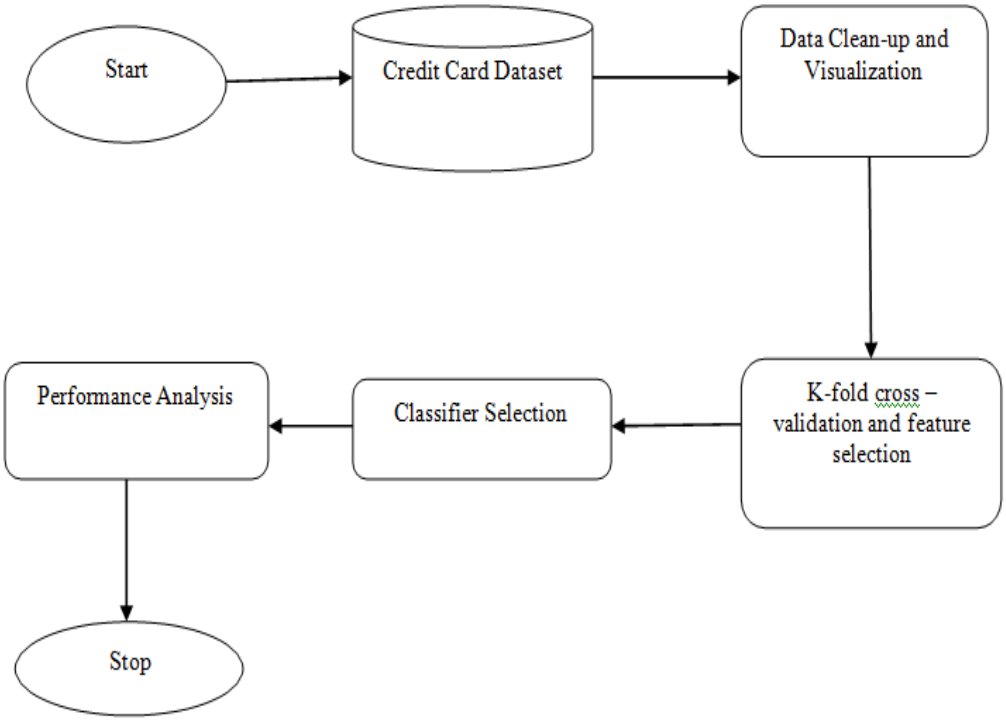


Fig.1: Process flow diagram of the proposed work

At this point, different subsets of the preparation dataset are made to provide consistency on the best features for the final research.

Steps of Algorithm

1. Import whole libraries
2. Get the dataset it means credit card.csv
3. Divide the dataset into training set and testing set as x_train, y_train, x_test and y_test
4. Now apply feature-selection method for selecting the finest feature using random forest (RF) classifier clf_rf_5 = RandomForestClassifier() clr_rf_5

= clf_rf_5.fit(x_train,y_train) importance

= clr_rf_5.feature_importances_

std = np.std([tree.feature_importances_ for tree in clf_rf.estimators_],

axis=0)

Indices = np.argsort(importance)[::-1]

5. Succeeding classification algorithms will be applied on the reduced dataset.
6. Print precision, accuracy, recall and F-score.

5. RESULT ANALYSIS

In this case, we use Python version 3.6 for the examination, with this examination as its argument. This section will include all calculations associated with the systematic arrangements,

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

both in parallel and sequential assessment. Here, the top four evaluation models are also introduced.

A. Confusion Matrix:

A classification difficulty's unique prediction outcome is a confusion matrix. The number of accurate and inaccurate predictions is totaled, weighted, and disaggregated by each class. The confusion matrix's key is this. The confusion matrix shows how your classification model can become confused while making predictions. It allows us to get closer to the errors made by individuals through a classifier as well as more clearly see the kind of errors that are being made.

	Set 1 (Yes) Predicted	**Set 2 (No) Predicted**
Set 1(Yes) Actual	TP	FN
Set 2(No) Actual	FP	TN

Here,

Set 1: Positive (Yes) Set 2: Negative (No)

Description of the Terms:

- Positive (P): Test is positive (for instance: is an apple). Negative (N): Test is not positive (for case: is not an apple).
- True Positive (TP): Test is positive, accompanied by is predicted to be positive.
- False Negative (FN): Test is positive, other than is predicted negative.
- True Negative (TN): Test is negative, accompanied by is predicted to be negative.
- False Positive (FP): Test is negative, further than is predicted positive.

B. Classification Accuracy:

Classification During the relationship, accuracy is recognized. However, there are drawbacks to accuracy. It is believed that common types of errors will have associated costs. Depending on the difficulty, 99% accuracy could be acceptable, great, mediocre, average, or even terrible.

Right now, we are using classification techniques found in a thorough library. We will first estimate the confusion matrix in order to calculate the exactness throughout using a function or any confusion metric.accuracy_score;

$$\text{Accracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

As stated by RF (Random Forest), we will demonstrate the productivity of Dataset, which is given below:

	True	**False**
True	77981	0
False	24	79

Accuracy : 99.97%

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Precision: 100.0%

Recall: 76.7%

F1 Score: 86.8%

Such as indicated by K- nearest neighbor

	True	False
True	77976	5
False	31	72

Accuracy: 98.03%

Precision: 93.5%

Recall: 69.9%

F1 Score : 80%

As stated by Logistic Regression

	True	False
True	77979	2
False	30	73

Accuracy: 99.23%

Precision: 97.3%

Recall: 70.9%

F1 Score: 82%

As stated by Naïve-Bayes determine the confusion matrix

	True	False
True	76323	1658
False	20	83

Accuracy: 97.85%

Precision: 47.6%

Recall: 80.6%

F1 Score: 49%

The accuracy scores of K-nearest neighbour classifier, Logistic Regression, Naive Bayes, Random Forest, and other classifiers were compared. The following table 1 displays the comparison of classifier results.

Table 1: Comparisons of previous & present result

Method	Base Methods	Proposed Methods
Logistic Regression	0.9824	0.9923
Naïve Bayes	0.9737	0.9785
K- nearest neighbor	0.9691	0.9803

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Random Forest	-----	0.9997

The table above shows how our recommended methods outperform the existing method. Conversely, we also applied the work based on the F1 score, recall, accuracy, and precision. The suggested approach outperforms the existing method, and the random forest classifier performs better than other classifiers like Naive Bayes, Logistic Regression, and K-Nearest Neighbor. The figures below for accuracy, precision, recall, and F1 score display a graph comparison.

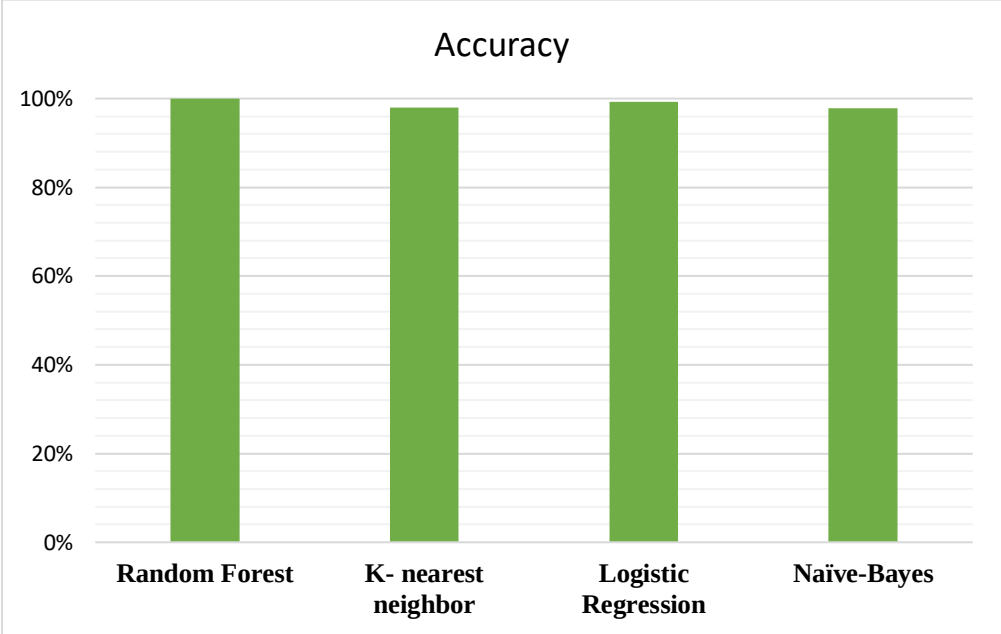
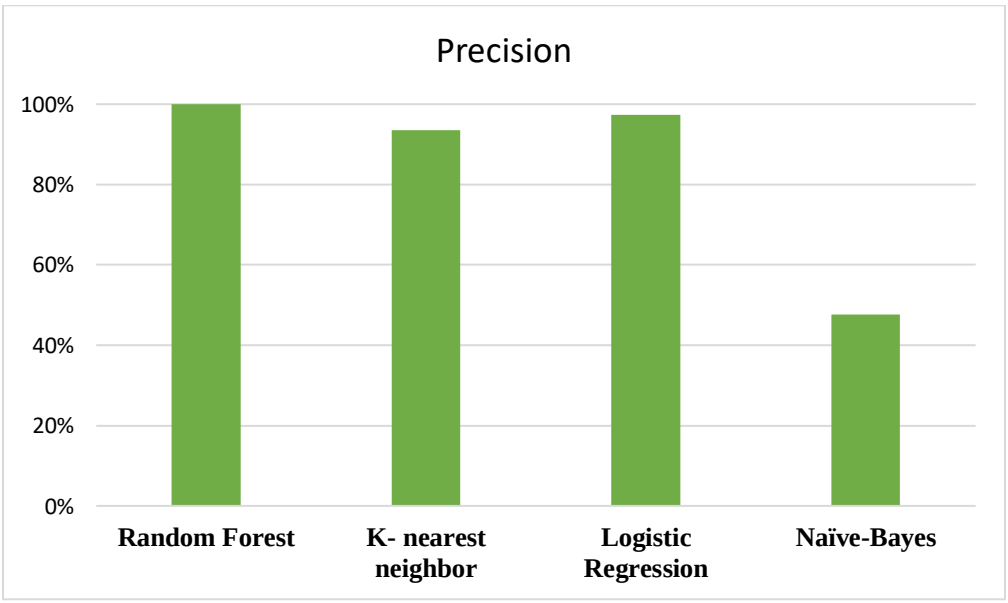


Figure 2- Accuracy comparison of different classifiers



ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

Figure 3- Precision comparison of different classifiers

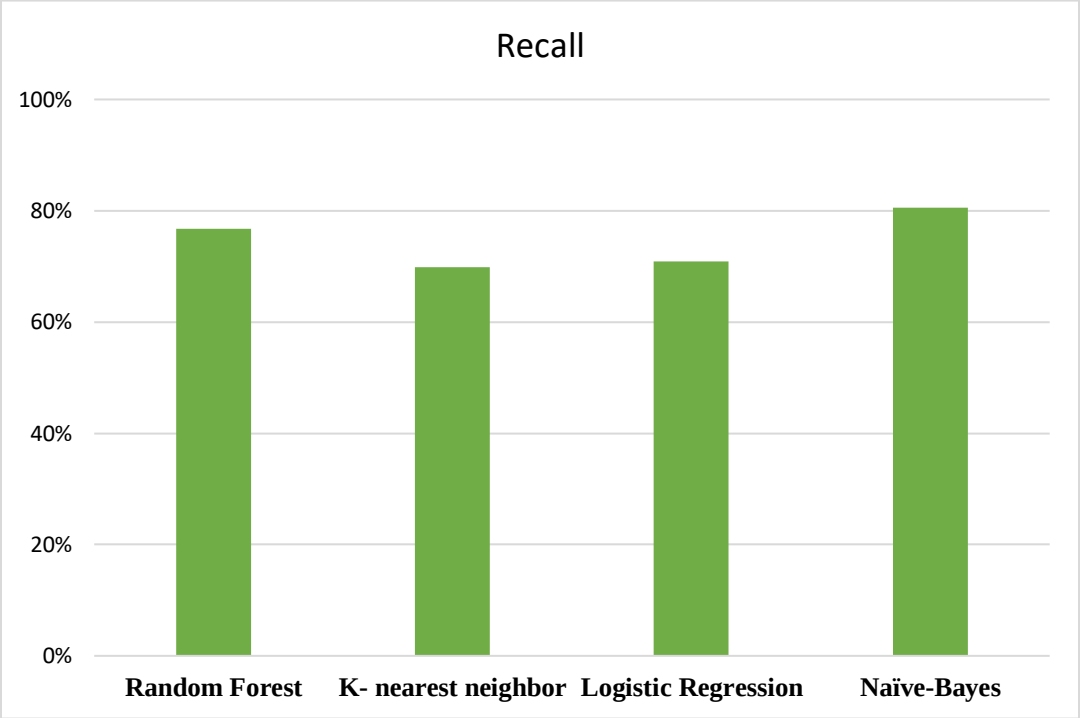


Figure 4- Recall comparison of different classifiers

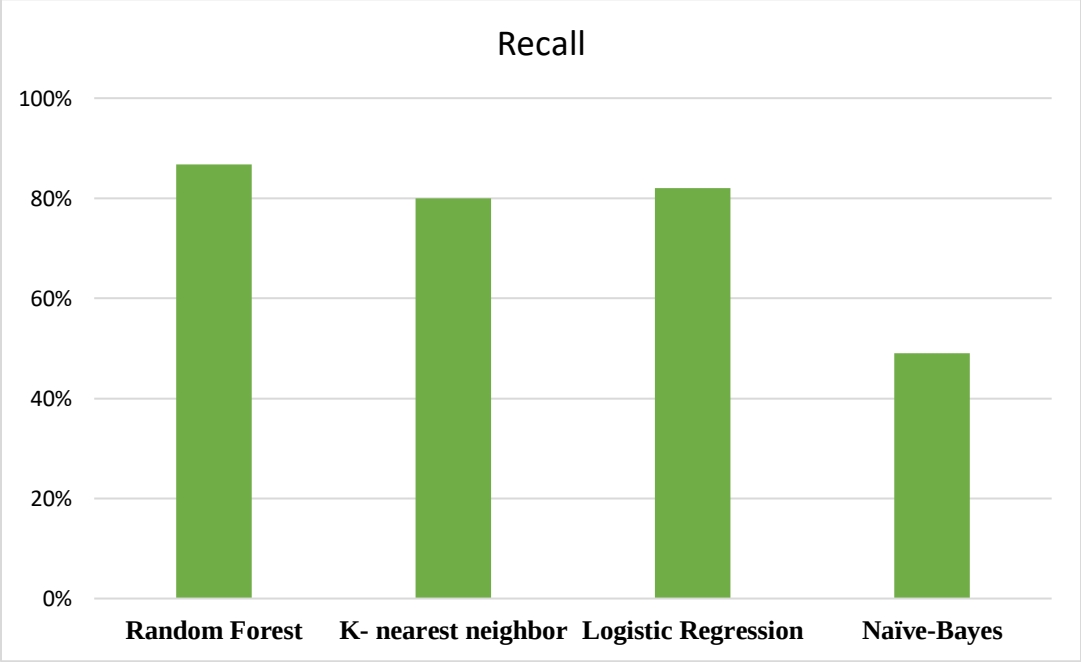


Figure 5- F1 score comparison of different classifiers

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022**

6. CONCLUSION

In this work, we have demonstrated how to classify credit cards using a feature selection method. The feature selection strategy lowered the dimensions, which improved the performance of the classifiers. This study compares the binary categorization of unbalanced credit card fraud data using Naive Bayes, K-nearest Neighbor, and Logistic Regression models. The justification for looking into these three techniques is that there have not been as many comparisons drawn to them in earlier research. However, a follow-up study employing our methodology to examine various single and ensemble methodologies is also under way. The findings of proposed method give better accuracy that the existing method which is about 100%. Future applications of deep learning techniques could enable us to do classification without the need for dimensionality reduction techniques.

7. REFERENCES

1. L. Delamaire, H. Abdou, and J. Pointon, “Credit card fraud and detection techniques: a review,” Banks Bank Syst., 2009.
2. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” Decis. Support Syst., 2011.
3. Data Mining: Practical Machine Learning Tools and Techniques. 2016.
4. D. V. and D. R., “BEHAVIOR BASED CREDIT CARD FRAUD DETECTION USING SUPPORT VECTOR MACHINES,” ICTACT J. Soft Comput., 2016.
5. K. R. Seeja and M. Zareapoor, “FraudMiner: A novel credit card fraud detection model based on frequent itemset mining,” Sci. World J., 2014.
6. E. Duman and M. H. Ozcelik, “Detecting credit card fraud by genetic algorithm and scatter search,” Expert Syst. Appl., 2011.
7. Y. Sahin, S. Bulkan, and E. Duman, “A cost-sensitive decision tree approach for fraud detection,” Expert Syst. Appl., 2013.
8. L. S. Raghavendra Patidar, “Credit Card Fraud Detection Using Neural Network,” India Int. J. Soft Comput. Eng., 2011.
9. Y. Sahin and E. Duman, “Detecting credit card fraud by ANN and logistic regression,” in INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications, 2011.
10. M. J. Islam, Q. M. J. Wu, M. Ahmadi, and M. A. Sid- Ahmed, “Investigating the performance of Naive- Baye
11. Classifiers and K- nearest neighbor classifiers,” in 2007 International Conference on Convergence Information Technology, ICCIT 2007, 2007.
12. A. Ghodsi, “Dimensionality Reduction A Short Tutorial,” Science (80-.)., 2006.
13. D. Iyer, A. Mohanpurkar, S. Janardhan, D. Rathod, and A. Sardeshmukh, “Credit card fraud detection using hidden Markov model,” in Proceedings of the 2011 World Congress on Information and Communication Technologies, WICT 2011, 2011.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, S Iss 3, Dec 2022

14. Ghosh and Reilly, "Credit card fraud detection with a neural-network," in Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS-94, 1994.
15. V. Dheepa and R. Dhanapal, "Analysis of Credit Card Fraud Detection Methods," ternational J. Recent Trends Eng., 2016.
16. C. S. Hilas and P. A. Mastorocostas, "An application of supervised and unsupervised learning approaches to telecommunications fraud detection," Knowledge-Based Syst., 2008.
17. S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning," Inf. Fusion, 2009.
18. K. Goswami, Y. Park, and C. Song, "Impact of reviewer social interaction on online consumer review fraud detection," J. Big Data, 2017.
19. M. Zareapoor and P. Shamsolmoali, "Application of credit card fraud detection: Based on bagging ensemble classifier," in Procedia Computer Science, 2015.
20. R. Nair and A. Bhagat, "A Life Cycle on Processing Large Dataset - LCPL Rajit Nair," vol. 179, no. 53, pp. 27–34, 2018.
21. M. H. ur Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman,
22. T. Y. Wah, and S. U. Khan, "Big Data Reduction Methods: A Survey," Data Sci. Eng., 2016.
23. R. Kannan and M. Karpinski, "Approximation Algorithms for NP-Hard Problems," Oberwolfach Reports, 2009.
24. A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, "Improving Credit Card Fraud Detection with Calibrated Probabilities,"