IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analytical Study of Machine Learning Techniques for Credit Card Fraud Detection

[1]Mr. V. Vinay Kumar, [2]Susreeja Diddi, [3]Nikita Panchadhar, [4]V. Sushniv

[1]Assistant Professor, [2]Student, [3]Student, [4]Student

[1,2,3,4]Department of Computer Science and Engineering,

[1,2,3,4]Matrusri Engineering College, Hyderabad, India.

Abstract: With the widespread use of credit cards, fraud has become a major problem in the credit card industry. As a result of theft, businesses and banks are reluctant to disclose the amount of money lost. Another issue with calculating credit card fraud losses is that it can only estimate the number of frauds that have been detected. It becomes harder to monitor the behavior and pattern of such transactions. Therefore, the assessment of fraud is necessary on a regular basis. The technologies like machine-learning, data-mining, and other artificial technologies are used to resolve this. In this regard, enhancing effective fraud detection using machine-learning methods is critical for empowering fraud investigators with these losses. This research paper explores how several Machine Learning models are being used to diagnose suspicious transactions. Machine learning contributes to the critical function of detecting credit card fraud during transactions. The dataset sampling methodology, selection of variables, and methods used for identification always have a massive influence on fraud detection efficiency. This system provides the necessary characteristics for monitoring both illegal and legal transactions. Classification Techniques like Decision Tree, K-nearest Neighbors algorithm, Logistic Regression, Gaussian Naive Bayes, and Artificial Neural Networks are demonstrated to identify fraudulent transactions. By training these techniques and evaluating them on various factors such as precision, accuracy, recall, and visualized the ROC curve. Based on the criteria of different methods, the best approach for detecting credit card fraud is chosen. As opposed to other algorithms, this paper shows that Decision Tree is the best optimal approach and can be used to detect further frauds.

Index Terms **- Artificial Neural Networks; Decision Tree; K-Nearest Neighbors; Logistic regression, Gaussian Naive Bayes.**

I. INTRODUCTION

Credit card payments are one of the most popular forms of electronic payment. A credit card allows users to purchase goods without having to pay cash. The majority are magnetic stripe cards with an EMV chip for use with card readers. Each card is assigned a distinct number [6]. The client may purchase merchandise or accommodations using this number and other information on the card (such as the validity date or a code). The money is then sent to the vendor by the card issuer [1]. The person who uses the card receives a credit. The user has a certain amount of time to pay off their credit card bill [14][13]. Globalization and increased use of the Internet for online transactions have resulted in a significant increase in credit card payments around the world. People are reliant on online transfers and believe in going cashless. The credit card has made online transactions more convenient and affordable. Therefore, a significant increase in credit card sales causes an increase in illegal practices. Credit card security is based on the physical defense of the card and the anonymity of the credit card number. Per year, criminal credit card charges result in billions of dollars in losses. Fraud is as ancient as humanity, and it can take on a wide range of various ways. As a result, it is unquestionably necessary to resolve the issue of credit card fraud identification. Furthermore, the advancement of emerging technology has provided additional avenues for scammers to operate. Credit cards are widely used in today's world, and these frauds have been on the rise in recent years. Huge financial damages have resulted from illegal activities affecting retailers and banks and individual credit consumers. Fraud may also damage a merchant's name and image, contributing to non-financial losses [8].

Fraud detection is the mechanism of examining a cardholder's transaction behavior to establish if an incoming transaction is genuine and accepted or not. If it is not, it would be flagged as fraudulent [8]. Various fraudulent activity identification approaches have been applied in credit card transactions, and strategies to create models based on artificial intelligence, data analysis, and machine learning have been held in researchers' minds. The identification of credit card fraud is a complex but common issue to tackle. Machine learning was used to construct the credit card fraud identification in the proposed scheme. Machine learning approaches are becoming more advanced. Machine learning has been described as a useful tool for detecting fraud. During online purchase operations, a vast volume of data is exchanged, resulting in a binary result: true or fake [11]. Features are built inside the study fraud datasets. There are data points such as the customer account's age and value and the credit card's source. There are hundreds of functions, each of which adds to the likelihood of fraud to differing degrees [14]. The artificial intelligence of the computer, which is powered by the training collection, determines the degree to which each function contributes to the fraud ranking, although it is not calculated by a fraud analyst. So, in the case of card theft, where the use of cards to perform fraud is known to be high, the fraud weighting of a credit card transaction would be equivalent. Nevertheless, if this were to diminish, the amount of contribution would also decrease. In Simplest Terms, these models self-learn without the need for specific programming or manual analysis. Machine learning is used to diagnose credit card theft

by deploying classification and regression algorithms. For classifying the credit card dataset in a planned scheme, supervised learning algorithms like the Decision Tree algorithm, Artificial Neural Networks, Logistic Regression, K-Nearest Neighbors, and Gaussian Naive Bayes Model are used [14]. Decision Tree is a member of the nursing algorithmic classification and regression software. One of the best types of learning algorithms based on different learning methods is decision trees. They improve the precision, readability, and stability of predictive models. Since they can address data-based problems including regression and classifications, the methods are also useful for fitting non-linear relationships.

Decision trees have the advantage of being easy to understand and interpret. Visualizing trees is simple and requires no data processing. Other strategies also necessitate data normalization, the development of dummy variables, and the removal of null values. A decision tree and its closely related effect diagram are used as a visual and empirical decision support method in decision analysis to quantify the predicted values of competing alternatives [4]. The contents of the leaf node form the result of the tree, which can be linearized into decision law, with the constraints along the path forming a conjunction in the if clause. A random subset of the training set is sampled to train each tree, and then a decision tree is developed, with each node separating on a feature selected from a random subset of the full feature set. Training in this classifier is incredibly fast, even for massive data sets with many features and data instances, since each tree is trained independently of the others [10][6].

A Supervised Machine Learning classification algorithm called logistic regression is used to estimate the likelihood of a categorical dependent variable [7]. The objective or categorical dependent variable is dichotomous in existence, implying that there are only two possible groups. The dependent variable in logistic regression is a binary variable that includes data coded as 1 (yes, achievement, etc.) or 0 (no) (no, failure, etc.). In other words, as a function of X, the logistic regression model predicts P(Y=1). The likelihood of the default class is modeled using logistic regression. While logistic regression is a linear procedure, the logistic function is used to transform the predictions [10][7].

KNN is a classification method in which the function is only approximated locally, and all computation is suspended until after the function has been evaluated [10]. Since this algorithm depends on distance for classification, normalizing the training data will significantly increase its accuracy if the features match different physical units or come in dramatically different sizes [12][13].

Naive Bayes is a concise method for building classifiers, which are models that assign class labels to problem instances represented as vectors of feature values, with the class labels drawn from a finite set. These classifiers are a subset of basic "probabilistic classifiers" based on Bayes' theorem and strict function independence assumptions [4]. They are one of the most basic Bayesian network models, but they can reach higher levels of precision when combined with kernel density estimation [10][12].

An artificial neural network (ANN) is a component of a computer system that simulates how the human brain analyses and processes data. Artificial intelligence (AI) is built on this basis, and it solves problems that would be impractical or difficult to solve by human or statistical criteria. It has self-learning skills, allowing them to improve their performance as more data becomes available. These are made up of processing units, which have inputs and outputs [13]. The ANN learns from the inputs to generate the desired product. Backpropagation is a compilation of learning principles that artificial neural networks use to direct them [11].

II. Related Work

The prevention of financial crime is an emerging domain where the suspects are to be held upside down. There are several aspects of intelligent fraud prevention that have not been studied. According to a fraud identification survey, there are various categories of frauds and various computational tools for identifying financial fraud. Different computational methods for detecting fraud have been stated by computing various parameters for each type of algorithm and representing the computing time with a graphical view. In the current method, fraud detection is accomplished using ID3 and support vector machine algorithms, as well as a survey. Implying the amount of fraud that occurred, identifying, and comparing various criteria for the use of algorithms [7]. The modern finance industry relies heavily on fraud identification. The scheme that was suggested is a supervised learning algorithm such as Artificial Neural Networks, Decision Tree, Logistic Regression, K-nearest neighbor, and Gaussian Naive Bayes, contrasting the accuracy obtained by all these five learning algorithms [12]. Despite differences in effectiveness, each strategy was found to be relatively capable of detecting different types of financial fraud. The ability of statistical approaches such as decision trees to learn and adjust to new strategies becomes extraordinarily successful against fraudsters emerging tactics. With the available dataset, to identify the transaction as valid or fraudulent, and from these values, it can measure and graph the sensitivity and efficiency [10][6][9].

III. Literature survey

A literature review involves all the studies conducted on a certain subject by different scholars. It incorporates both existing and unfinished works from all secondary materials into the researcher's perspective. Its primary goal is to increase awareness in this region. Sahil Dhankhad et al. [1] have applied Supervised ML techniques to recognize fraudulent transactions. They implemented super classifiers by using various ensemble learning methods. They also recognized the variables that prompted good accuracy. Furthermore, the comparison of various algorithms classifiers executed in this paper is done. K. Chaudhary et al. [3] implemented a model that performed online learning and recorded higher percentage accuracy of 91%, hence misclassification cases were reduced to the barest minimum, and the efficiency of fraud detection was increased. Siddhartha Bhattacharyya et al. [5] presented a logistic regression algorithm that was applied to an international credit card operation. This approach maintains similar performance with varying proportions of fraud in the training data.

O.S. Yee et al. [7] used a set of attributes of the dataset at the root of the tree and divides the training sets into subsets, where these subsets made a degree of the way that each contain data of the same value for an attribute, and this is repeated till it reaches the leaf node. Logistic regression was also used where the regression analysis was conducted when the dependent variable is binary. Neha Sethi et al. [8] presented modern and new techniques which are based on ANN. The accuracy obtained was medium. The paper also calculated the speed of deduction and cost which was fast and expensive. It is used to model complex relationships between inputs and

outputs and to find patterns in data. This paper focuses on creating a hybrid approach for developing some effective algorithms that can perform well with minimum costs and higher accuracy for future work. Samidha Khatri et al. [10] presented an analogy between different supervised learning algorithms for the detection of counterfeit transactions. The models used in the paper are as follows: Decision Tree, KNN, Random Forest, Naive Bayes, and Logistic regression on the imbalanced dataset. Precision, Time, and sensitivity are used to predict the best model in detecting fraud transactions. Aayushi Agarwal et al. [11] Different classification techniques have been applied for detecting the frauds that occur in credit card transactions. Most used were Artificial neural networks and k nearest neighbors. They provide a basis and present different advantages and disadvantages by the usage of those algorithms.

Sunil Bhatia et al. [12] used comparative analysis and calculated the true positive rate, false-positive rate, and accuracy generated by the systems. They also compared the ROC curve, the error rate for different classification techniques to measure the performance for good analysis. It was required to find out the perfect solution to overcome the drawbacks by creating hybrids of different techniques. Sometimes depending upon the environment and the applications, it might give higher accuracy. So, to increase the accuracy, they planned to try unsupervised learning in the future to achieve better performance. F.N. Ogwueleka et al. [13] presented a methodology of a four-stage credit card fraud detection model where they focus more on the ROC curve. This eliminated the classification of legal transactions as fraudulent and guaranteed a precise and consistent outcome. The paper confirms the reliability of ANNs intelligent identification methodologies to be used in an effective fraud detection scheme.

F. N. Ogwueleka et al.[13] presented a methodology of four-stage credit card fraud detection model where they focus more on the ROC curve which summarizes the possible performances of a detector and used both traditional data mining and neural network approaches to achieve a synergy that better manages the Nigerian credit card fraud situation using four clusters instead of the two-stage model/clusters usually used in fraud detection algorithms This eliminated the classification of legal transactions as illegitimate and guaranteed a precise and consistent outcome. The paper confirms the validity and reliability of ANNs as a testing instrument and lays the foundations for intelligent identification methodologies to be used in an effective fraud detection scheme. SamanehSorournejad et al.[14] the paper has investigated the difficulties of credit card fraud detection and could bring up the advantages and disadvantages of fraud detection methods are enumerated and compared. They presented supervised and unsupervised techniques. All types of datasets were used to get an effective outcome.

IV. OVERVIEW OF THE SYSTEM

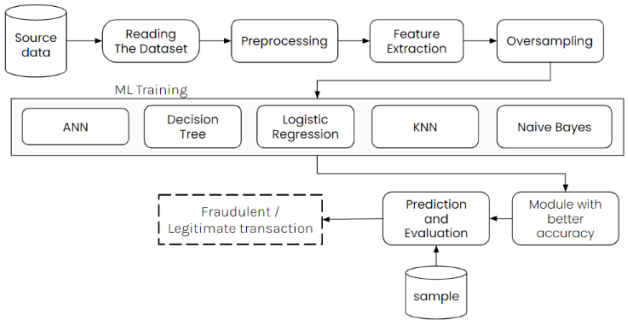

Fig 4.1: System Architecture

4.1 Disadvantages of Existing System

- Data mining is being used in the current system. This method of detecting fraudulent transactions was time-consuming and complex.
- Since millions of transactions are made per day, the data is heavily distorted. It necessitates the use of highly effective techniques to scale down all data while still attempting to distinguish fraudulent rather than legal transactions.

4.2 Proposed Scheme

The functionality of the proposed fraud detection method is depicted in the diagram above. The source credit card data collection is read and loaded into the Colaboratory notebook, where numerous Data cleaning and pre-processing operations are performed to become familiar with the data. The entire dataset is separated into two categories: dependent and independent functions. Since the data collection used is highly unbalanced, it is oversampled to get the best results. By using Machine Learning methods such as Artificial Neural Networks, Gaussian Naive Bayes Model, Logistic Regression, Decision Tree, and K-nearest Neighbours. The different parameters of these algorithms, such as accuracy, precision, and recall, are compared to find the best classifier for potential estimation of unknown credit card transactions. The chosen model should identify the given singular unknown transaction sample as either valid or fraudulent during the testing process.

4.3 Advantages of Proposed System

- Decision Trees may be used to rate the importance of variables in a regression or classification problem in an accurate way.
- This system is capable of handling large datasets.

4.4 Process Logic

4.4.1 Data Collection

The reports of European cardholders who used their credit cards in September 2013 are used as the source data to identify credit card fraud. This dataset contains records of transactions made for two days, with a total of 284,807 transactions, 492 of which were fraudulent, making the dataset somewhat imbalanced and more geared toward the positive class, i.e., fraud transactions account for 0.172 percent of total transactions. Fig 5.2 represents this imbalance dataset. The data is stored in CSV format and contains 31 features in total. It only has numeric input variables that have undergone a PCA transformation. The inclusion of the original functionality is not possible due to security concerns. The principal components obtained with PCA are features V1, V2... V28; the

only features not transformed with PCA are 'Time' and 'Amount.' The response variable is called 'Class,' and it has a value of 1 when there is fraud and 0 when there isn't. Fig 5.1 shows the summary of all the features present in the dataset [5].

4.4.2 Data Pre-processing

Pre-processing entails the following essential and standard steps:

1. Formatting: This is the way of getting data into a reasonable format that can be worked with. The data files should be formatted by the requirements. Most used are the CSV files.
2. Cleaning: Data cleaning is a critical step in the data science process since it accounts for most of the work. It covers things like deleting missed details and complexities, among other things. There may be instances of missing data and do not include the information needed to solve the problem.
3. Sampling: The information in the dataset is uneven. When using learning algorithms to train an unbalanced dataset, it's possible that the minority class would be misclassified. As a result, using the SMOTE oversampling process introduced in the imbalanced-learn kit to optimize for the imbalance. It is an oversampling technique in which synthetic samples for the minority class are produced. This algorithm aids in overcoming the issue of overfitting caused by random oversampling. It concentrates on the feature space to create new instances by interpolating between positive instances that are close together. Fig 5.2 depicts the count of class 0 (legitimate) and class 1 (fraudulent) and also shows that all the categories are balanced after the implementation of SMOTE.
4. Scaling: It is a data pre-processing phase that is implemented to independent variables or data features. It essentially assists in the normalization of data within a specific range. It may also aid in the speeding up of calculations in an algorithm.

4.4.3 Data Visualization

Data visualization is a compilation of data points and statistics that are graphically displayed to make it simple and easy to interpret for users. This description is effective, has a specific purpose, and is incredibly simple to understand without the use of context. By using visual effects or items such as a table, graphs, and diagrams, data visualization tools make it easier to see and understand trends, outliers, and patterns in data. To determine the number of frauds and legitimate transactions, a Count plot is created. Histograms and Scatter plots are used to figure out how variables are distributed. Fig 5.3 shows the illustrations of the plots used to describe the dataset.

4.4.4 Feature Selection

The method of analysing the actions and pattern of the analysed data and creating features for further research and training is known as feature extraction. The gathered labelled dataset is utilized. The aim attribute is the Feature Class, which is stored in a separate variable from the rest of the attributes. To differentiate pre-processed data, some machine learning algorithms were used.

4.4.5 Training the Model

The dataset is partitioned into a training set and a test set when developing machine learning models. The trained model should do well on new, undiscovered results [5]. The usable data set is divided into two parts, often referred to as the train-test split, to simulate the latest, unknown data. The first subset is often a larger data subset used as the training set (for example, accounting for 70% of the initial data. The second is typically a smaller subset used as the research set (the remaining 30 percent of the data). After this, the various Classifier algorithms are used to train the models. The models will be examined using the rest of the labelled information [5].

4.4.6 Model Evaluation

Model assessment is an important step in the implementation of a model. It aids in the selection of the best model to reflect the data and the prediction of how well the chosen model will do in the future. To quantify the true positive, true negative, false positive, and false negative produced by a system or an algorithm and use these in quantitative calculations to test and compare the output of different systems to compare different techniques. The below are the different measurement metrics:

1. Accuracy is the percentage of transactions accurately classified. It is one of the most widely used and efficient evaluation metrics [10][5].

$$\text{Accuracy} = \frac{(TN + TP)}{(TP + FP + FN + TN)} \qquad (1)$$

2. Precision: It is also known as identification rate, which is the number of transactions accurately identified that were legitimate or fake [10][5].

$$\text{Precision} = \frac{TP}{(TP + FP)} \qquad (2)$$

3. Recall: The number of accurate predictions divided by the number of outcomes that could have been correctly predicted is known as recall [10][5].

$$\text{Recall} = \frac{TP}{(TP + FN)} \qquad (3)$$

A brief performance study of all classifiers is showcased in table 1.1. A ROC curve (Receiver Operating Characteristic curve) is another graphical visualization measure that displays a classification model's efficiency overall classification thresholds. The best model is chosen based on the model's efficiency on the testing set [13]. Fig 5.5 describes the ROC plot of all techniques used in this paper.

4.4.7 Testing Phase

Model testing for our fraud detection system involves two main test cases, which completely is dependent upon our target variable: 1) When a singular known Legitimate transactional record is given to the entire system, 2) When a single familiar Fraudulent transaction sample is given, for classification purposes, to determine whether it is predicting the given tuple correctly. Fig 5.6 demonstrates the testing cases conducted.

4.4.8 Fraud Detection

Decision Tree outperforms other algorithms. As a result, this classifier is used to determine whether an unidentified record is fraudulent or not. Fig 5.7 shows the fraud detection performed by The Decision Tree Classifier.

V. RESULTS

The results of this study are seen in the following figures. The three major metrics used to test a classification model, such as accuracy, precision, and recall of all the classifiers applied, are taken from the confusion matrix. A *confusion matrix* is a table that is often used to represent the utility of a classification model on a set of test data on which the true values are known.

There are four possible consequences when making classification predictions: False Positive (FP), False Negative (FN), True Positive (TP), and True Negative.

- TP: It stands for the true positive rate, which is the percentage of fraudulent transactions that are accurately classified as such.
- TN: the true negative rating is the percentage of regular transactions that are properly labelled as such.
- FP: It is the percentage of non-fraudulent transactions that are incorrectly reported as fraudulent.
- FN: The false negative rate refers to the percentage of fraudulent transactions that are mistakenly classified as regular. Both algorithms and techniques have the goal of lowering FP and FN rates, while increasing TP and TN rates while maintaining a high detection rate.

Fig 5.4 explains the confusion matrices of all the machine learning approaches implemented in this paper.

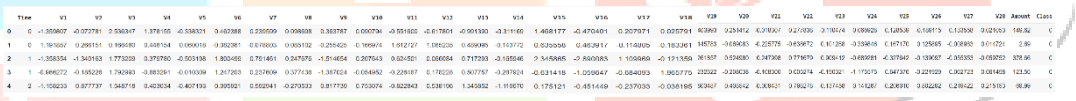

Fig 5.1. Overview of first five records.

Fig 5.2. Analysis of Valid and Fraud Transactions before and after implementation of SMOTE

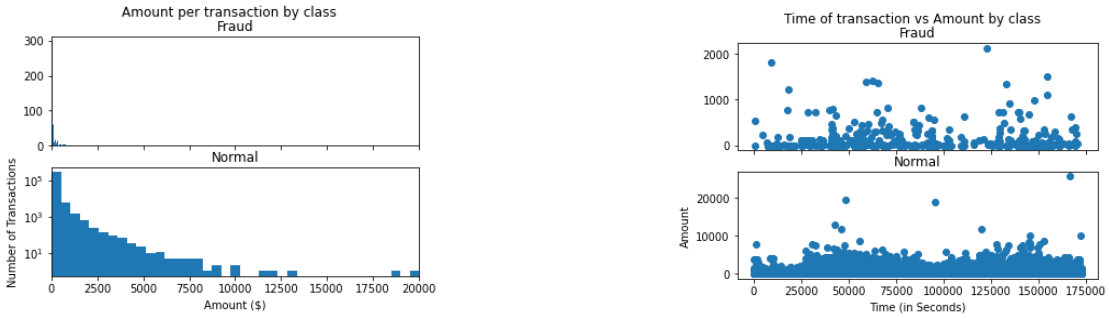

Fig 5.3. Visualization of various classes in the data set

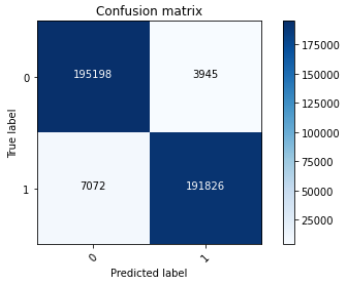

Fig 5.4 (a)

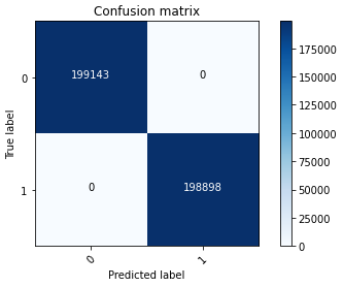

Fig 5.4 (b)

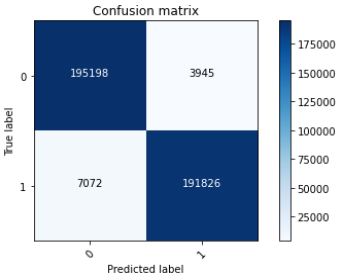

Fig 5.4 (c)

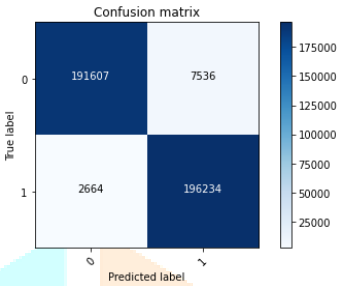

Fig 5.4 (d)

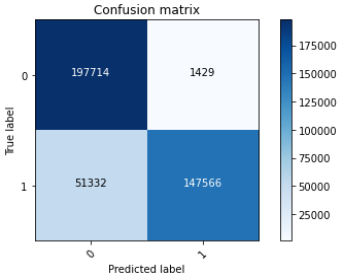

Fig 5.4 (e)

Fig 5.4. Model Evaluation of a) Artificial Neural Networks b) Decision Tree c) Logistic Regression d) K-Nearest Neighbours e) Naive Bayes.

Table 1.1: Performance analysis of different Supervised Machine Learning Algorithms.

Classifiers	Classes	Accuracy	Recall	Precision
Artificial Neural Networks	0	97.034%	0.99	0.95
	1		0.95	0.99
Decision Tree	0	99.861%	1.00	1.00
	1		1.00	1.00
Logistic Regression	0	97.176%	0.98	0.96
	1		0.96	0.979
K-Nearest Neighbours	0	95.933%	0.94	0.97
	1		0.98	0.95
Naive Bayes	0	86.721%	0.99	0.79
	1		0.742	0.99

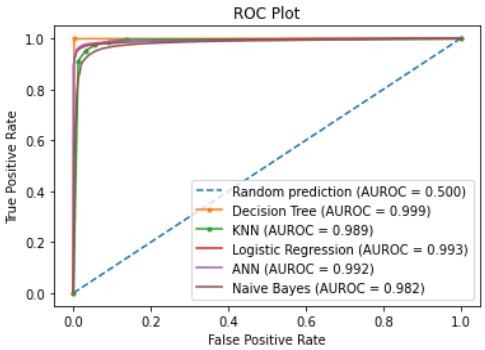

Fig 5.5. ROC Curves of all classifiers implemented.

```
# making a single prediction using logistic Regression
sample = [[4465,-2.303349568,1.75924746,-0.359744743,2.330243051,-0.821628328,-0.075787571,0.562319782,
          -0.399146578,-0.238253368,-1.525411627,2.032912158,-6.560124295,0.022937323,-1.470101536,
          -0.698826069,-2.282193829,-4.781830856,-2.615664945,-1.334441067,-0.430021867,-0.294166318,
          -0.932391057,0.172726296,-0.087329538,-0.156114265,-0.542627889,0.039565989,-0.153028797,239.93]]
pre = logistic.predict(sample)
print('Predicted Class: %d' % pre)
if pre==1:
  print("The given Transaction is Fraudulent")
else:
  print("The given Transaction is Valid")

Predicted Class: 1
The given Transaction is Fraudulent
```

```
# making a single prediction using Artificial Neural Network
sample = [[4465,-2.303349568,1.75924746,-0.359744743,2.330243051,-0.821628328,-0.075787571,0.562319782,
          -0.399146578,-0.238253368,-1.525411627,2.032912158,-6.560124295,0.022937323,-1.470101536,
          -0.698826069,-2.282193829,-4.781830856,-2.615664945,-1.334441067,-0.430021867,-0.294166318,
          -0.932391057,0.172726296,-0.087329538,-0.156114265,-0.542627889,0.039565989,-0.153028797,239.93]]
p = model1.predict(sample)
print('Predicted Class: %d' % p)
if p==1:
  print("The given Transaction is Fraudulent")
else:
  print("The given Transaction is Valid")

Predicted Class: 0
The given Transaction is Valid
```

```
# making a single prediction using KNN
sample = [[4465,-2.303349568,1.75924746,-0.359744743,2.330243051,-0.821628328,-0.075787571,0.562319782,
          -0.399146578,-0.238253368,-1.525411627,2.032912158,-6.560124295,0.022937323,-1.470101536,
          -0.698826069,-2.282193829,-4.781830856,-2.615664945,-1.334441067,-0.430021867,-0.294166318,
          -0.932391057,0.172726296,-0.087329538,-0.156114265,-0.542627889,0.039565989,-0.153028797,239.93]]
yp = knn.predict(sample)
print('Predicted Class: %d' % yp)
if yp==1:
  print("The given Transaction is Fraudulent")
else:
  print("The given Transaction is Valid")

Predicted Class: 1
The given Transaction is Fraudulent
```

```
# making a single prediction using Naive Bayes
sample = [[4465,-2.303349568,1.75924746,-0.359744743,2.330243051,-0.821628328,-0.075787571,0.562319782,
          -0.399146578,-0.238253368,-1.525411627,2.032912158,-6.560124295,0.022937323,-1.470101536,
          -0.698826069,-2.282193829,-4.781830856,-2.615664945,-1.334441067,-0.430021867,-0.294166318,
          -0.932391057,0.172726296,-0.087329538,-0.156114265,-0.542627889,0.039565989,-0.153028797,239.93]]
yhat = NBC.predict(sample)
print('Predicted Class: %d' % yhat)
if yhat==1:
  print("The given Transaction is Fraudulent")
else:
  print("The given Transaction is Valid")

Predicted Class: 0
The given Transaction is Valid
```

Fig 5.6. Testing Of Fraud Detection using ANN, Logistic Regression, KNN, Naive Bayes algorithms.

```
# making a single prediction using Decision Tree
sample = [[4465,-2.303349568,1.75924746,-0.359744743,2.330243051,-0.821628328,-0.075787571,0.562319782,
          -0.399146578,-0.238253368,-1.525411627,2.032912158,-6.560124295,0.022937323,-1.470101536,
          -0.698826069,-2.282193829,-4.781830856,-2.615664945,-1.334441067,-0.430021867,-0.294166318,
          -0.932391057,0.172726296,-0.087329538,-0.156114265,-0.542627889,0.039565989,-0.153028797,239.93]]
pr = decc.predict(sample)
print('Predicted Class: %d' % pr)
if pr==1:
  print("The given Transaction is Fraudulent")
else:
  print("The given Transaction is Valid")

Predicted Class: 1
The given Transaction is Fraudulent
```

Fig 5.7. Testing Of Fraud Detection using Decision Tree algorithm.

VI. CONCLUSIONS

In this project, the Supervised Classification algorithms are used to detect the credit card frauds of the real datasets. The dataset is highly imbalanced. Hence, the SMOTE oversampling method is utilized. This dataset is used to train and test the mentioned classification models. To evaluate the performance of each model, metrics like Accuracy, Confusion Matrix, Precision, and Recall are calculated. This paper has acquired the result of an accurate value of fraud detection around 99.8% using a Decision Tree algorithm with new enhancements. Hence, The Decision Tree algorithm will provide better performance. This system also handles large datasets.

REFERENCES

[1] S. Dhankhad, E. Mohammed and B. Far,' Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," IEEE International Conference on Information Reuse and Integration, pp. 122-125, 2018.
[2] Priyanka Sharma and Santhoshi Pote, "Credit Card Fraud Detection using Machine Learning models", IJCRT Volume 8, pp. 3-5, April 2020.
[3] K. Chaudhary, J. Yadav, and B. Mallick, "A review of Fraud Detection Techniques: Credit Card," Int. J. Comput. Appl., vol. 45, pp. 975-8887, April 2020.
[4] S. Venkata Suryanarayana, G. N. Balaji, and G. Venkateswara Rao, "Machine Learning approaches for Credit Card Fraud Detection", International Journal of Engineering & Technology, pp. 917-920, 2018.
[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decis. Support Syst., vol. 50, pp. 602-613, 2011.
[6] Linda Delamare, Hussein Abdou, John Pointon, "Credit card fraud and detection techniques: a review," Banks and Bank Systems, pp. 57-68, 2009.
[7] O. S. Yee, S. Sagadevan, N. Hashimah, and A. Hassain, "Credit Card Fraud Detection Using Machine Learning As Data Mining Technique," vol. 10, no. 1, pp. 23-27.
[8] N. Sethi and A. Gera, "A Revived Survey of Various Credit Card Fraud Detection Techniques", International Journal of ComputerScience and Mobile Computing, vol. 3, no. 4, pp. 780-791, 2009.
[9] T. P. Bhatla, V. Prabhu, and A. Dua, "Understanding Credit Card Frauds," Cards Bus. Rev., vol. 1, no. 6, pp. 1–15, 2009, doi: 10.1.1.431.7770.

[10] Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison", IEEE, 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp.680-683, 2020.
[11] Aayushi Agarwal, Md Iqbal, Baldivya Mitra, "Survey of Various Techniques used for Credit Card Fraud Detection", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 8 Issue VII, pp. 1642-1646, July 2020.
[12] Sunil Bhatia, Rashmi Bajaj, Santosh Hazari. "Analysis of Credit Card Fraud Detection Techniques", International Journal of Science and Research, Volume 5 Issue 3, pp. 1302-1307, March 2016
[13] F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System," vol. 6, no. 3, pp. 311-322,2011.
[14] SamanehSorournejad, Z. Zojaji, R. E. Atani, and A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective," no. November 2016, [Online]. Available: http://arxiv.org/abs/1611.06439.