A MINOR PROJECT REPORT ON

**“Credit Card Fraud Detection”**

**Submitted**

*In the partial fulfillment of the requirements**for*

*the award of the degree of*

**BACHELOR OF TECHNOLOGY**

In

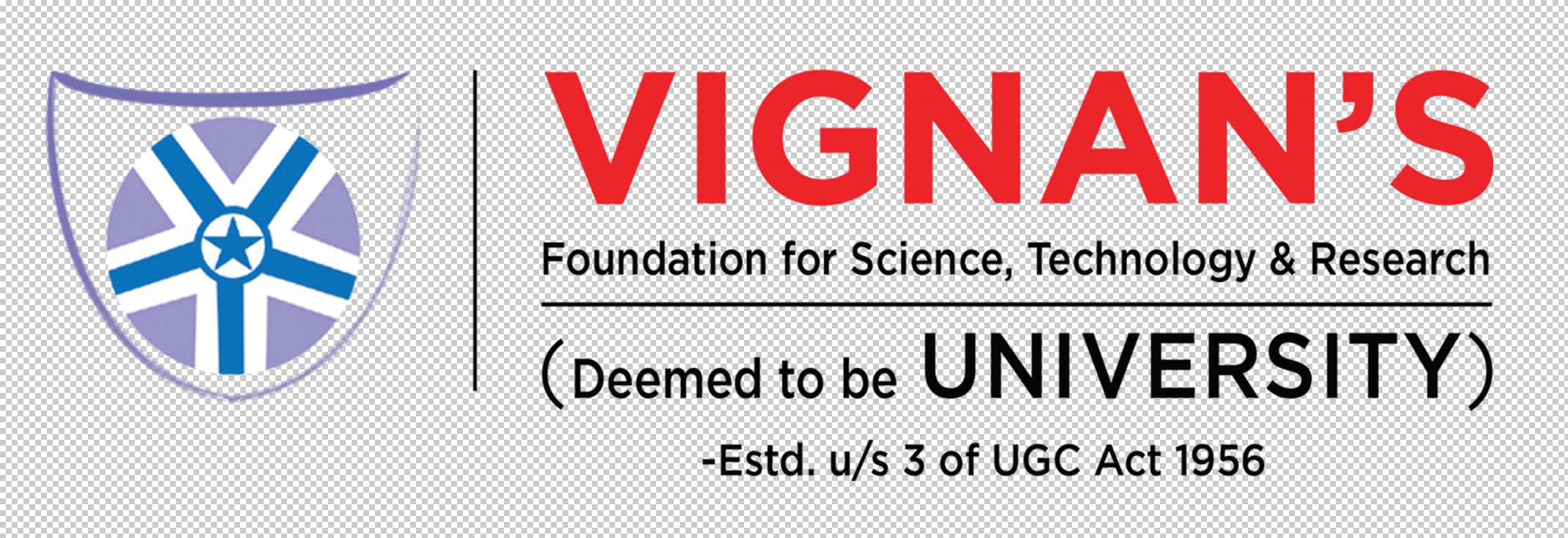
**COMPUTER SCIENCE & ENGINEERING**

By

|  |  |
| --- | --- |
| M. Manoj Kumar Reddy | 171FA04511 |
| Z. Yeshwanth | 171FA04512 |
| K. Akashdeep | 171FA04517 |

Under the guidance of

#### Dr. SV Phani Kumar, Associate Professor



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH UNIVERSITY**

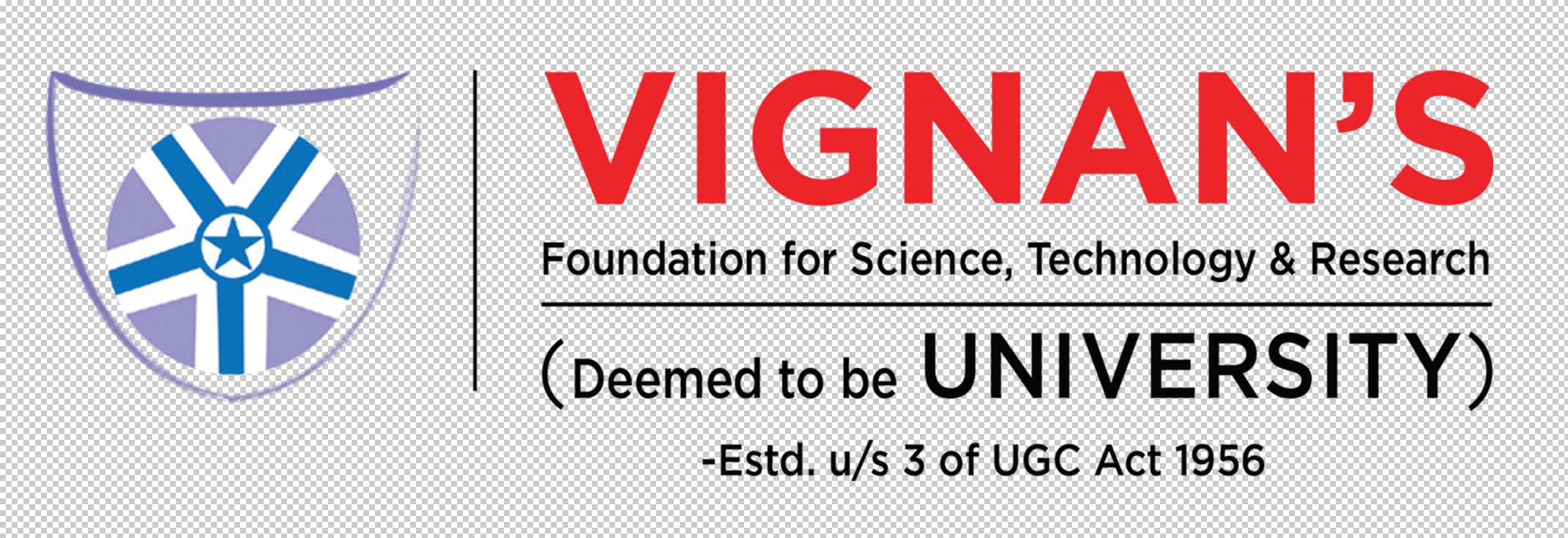
(**Accredited by NAAC “A” grade**)

**Vadlamudi, Guntur.**

**VIGNAN’S FOUNDATION FOR SCIENCE TECHNOLOGY AND RESEARCH UNIVERSITY**

**(Accredited by NAAC “A” grade)**

VADLAMUDI, GUNTUR DIST, ANDHRA PRADESH, INDIA, PIN-522 213



**CERTIFICATE**

This is to certify that the Minor project Report entitled **“Credit Card Fraud Detection”** that is being submitted by **M.Manoj kumar reddy (171FA04511), Z.Yeshwanth (171FA04512), K.Akashdeep (171FA04517)** in partial fulfillment for the award of B.Tech degree in Computer Science and Engineering to the Vignan’s Foundation for Science, Technology and Research, Deemed to be University, is a record of bonafide work carried out by them in the CSE Department.

Dr.SV Phani Kumar External Examiner Dr.D.Venkatesulu

Associate Professor HOD, CSE

**DECLARATION**

I hereby declare that the Minor project entitled “**Credit Card Fraud Detection**” submitted for the **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**. This dissertation is our original work and the project has not formed the basis for the award of any degree, associate-ship and fellowship or any other similar titles and no part of it has been published or sent for publication at the time of submission.

By

M.Manoj kumar reddy(171FA04511)

Z.Yeshwanth(171FA04512)

K.Akashdeep(171FA04517)

Date: 26/07/2020

**ACKNOWLEDGEMENT**

We are very grateful to our beloved Chairman **Dr. Lavu. Rathaiah**, and Vice- Chairman

**Mr. Lavu. Krishna Devarayalu,** for their love and care.

It is our pleasure to extend our sincere thanks to Vice-Chancellor **Dr. M. Y. S. Prasad** and Dean Engineering & Management, **Dr. V. Madhusudan Rao**, for providing an opportunity to do my academics in VFSTR.

It is a great pleasure for me to express my sincere thanks to **Dr. D. Venkatesulu, HOD,CSE** of VFSTR, for providing me an opportunity to do my Mini Project.

We feel it our responsibility to thank  **Dr.SV Phani Kumar**  under whose valuable guidance that the project came out successfully after each stage.

We extend our whole hearted gratitude to all our faculty members of Department of Computer Science and Engineering who helped us in our academics throughout course.

Finally we wish to express thanks to our family members for the love and affection overseas and forbearance and cheerful depositions, which are vital for sustaining effort, required for Completing this work.

With Sincere regards,

M.Manoj kumar reddy

Z.Yeshwanth

K.Akashdeep

**Index**

|  |  |  |
| --- | --- | --- |
| SNO | TOPICS | PAGE NO |
| 1 | Abstract | 6 |
| 2 | Introduction | 7 |
| 3 | Related Work | 8 |
| 4 | Proposed Methodology | 10 |
| 5 | Experimentation and Results | 15 |
| 6 | Conclusion | 16 |
| 7 | References | 16 |
|  |  |  |
|  |  |  |
|  |  |  |

1. **ABSTRACT**

With the advent improvement of technology most of the people are using online transactions to transfer the money from one person to another person. Credit Card Fraud is the rising problem in the financial department. Most of the people are doing research in this area and we are using some classification algorithms to predict the transaction is fraud or normal. The Credit Card Fraud Detection project is developed to spot whether a new transaction is fraudulent or not by checking the past credit card transactions with the knowledge of the previous data like fraud transactions which have happened earlier. We will use various predictive models to ascertain how accurate they are in predicting whether a transaction is a normal or abnormal. Techniques like Decision Tree, Logistic Regression, SVC and Naïve Bayes are the classification algorithms to detect the non-fraud and fraud transactions. Unfortunately, in modern conditions, classification techniques do not perform well when it involves huge numbers of differences in data distribution and also changing data.

Credit card frauds happens frequently and ends in a huge financial loss [5]. The number of online transactions has grown in bulk these days and online credit card transactions holds an enormous share of these transactions. More numbers of people are using the credit card for shopping, e-commerce and even for education purposes also. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can be of different categories. They may be through Online or Offline. This paper deals with these two categories

**Keywords -** Accuracy, Classification, Distribution, Logistic Regression, Decision Tree, SVC, Naïve Bayes.

## Introduction

Credit Card Fraud is an illegal activity which the fraudsters do in order to gain profit in a less amount of time and this will be known to the users after a few days when the fraud has been happened and they will respond later and they will register a complaint regarding the fraud which they have undergone .Basically the fraudsters may use online payment or offline by using credit card and through online the fraudsters need to have complete details of the user who is going to undergone in the loss of amount from there account .The fraudsters then use there username and password to login to their accounts to gain amount from the others this really requires a hacker to hack the accounts and to capture the details of the user and bank the hackers now a days are hacking the bank servers to gain a huge amount of money form the accounts of the bank customers.

The fraudsters by using credit card need to have their credit card and there pin to get money from there account in case they want to do offline frauds, but to get the credit card they used to stole from the users or the user may lost the card at some place and this can be taken by the hacker who can hack the details completely and unlock the card to take the amount from their accounts.

Credit card fraud detection is a difficult task as the money may differ from one account to other and so there is no particular pattern for identifying the fraud .Now a days all the business is through online and net banking so there will be more number of frauds and this will have a huge data set and this will be very confidential and not released to the public without prior permission from the head of the department (finances) .Even the customer with a minimum balance is using their net banking to buy things online and they will do shopping, e-commerce etc. through online .Data Mining is mostly used to detect this kind of frauds using several algorithms. In this project we are using logistic regression, linear regression, SVC, Naive Bayes, Decision Tree, KNN, K-Means and DBSCAN algorithms in order to figure out the best suited algorithm for this kind of problems.

The Transaction will have an id to track whether the transaction is fraud or not. Basically, this kind of problems will have two classes legit or fraud. The fraud transactions lead to illegal activities. Credit card datasets are rarely available as it is confidential as it is related to the finance department, and are highly imbalanced. The data set which we are using is having a complete data from a customer for two days by calculating the PCA and named as v1-v26. The data sets available are imbalanced and having a smaller number of frauds. so, this is difficult to identify the patterns in them, so these algorithms are best suited to fit the data and predict the data.

The dataset which we have used in this project is a collection of different transactions in a particular ample of two days and the attributes include ‘Time’, ‘V1-V28’, ‘Amount’ and ‘Class’. The ‘V1-V28’ are the results of PCA Transformation. The attribute Time and Amount are the time between the transaction and the first transaction and the attribute Amount is the transacted amount. The attribute Class is the one which have the value as 1 if the case is fraud and 0 otherwise. In this data set we have 492 fraud cases out of 2,84,807 transactions.

The fraudulent transactions can be of any type like online fraud or offline but the loss may be a huge one for the users these kinds of illegal activities reduce the reputation of the banks too. While online transactions happen through technical gadgets and virtual money, offline transactions happen in banks through physical cash.

These days the people are engaged in using social media and online transaction Applications which made them easy to transfer money from their account to other users and vendors for their business. So, most of the business people are using online transactions and there may be a greater number of chances to have fraudulent transactions in these cases.

1. **Related Work**

In this Project we have performed the data normalization before performing the cluster analysis and classification. The purpose of data normalization is to bring the data into a single and scalable format. Normalization may involve various methods such as min-max, decimal scaling, Z-score etc. Credit Card Fraud Detection has a most immense research to find out the fraud cases by using Machine Learning, Neural Networks and Data Mining as major fields.

Fraud detection involves continuous monitoring of the natural behavior of users and discriminating them from the unusual users. Although the frauds in a very large database are minute or negligible, they create a huge impact and leave many users vulnerable in the process. There are basically two approaches which are supervised and unsupervised approaches and the recently popularized semi-supervised approach.

Firstly, in a supervised approach the data is pre-labelled or well classified. The machine is then provided with a new set of examples and produces an outcome using the labelled data. It is much more simplified and highly accurate on datasets that do not change continuously.

One of the most popular supervised approaches in detecting fraud is logistic regression. It works by classifying the users into two classes fraudulent and non-fraudulent based on a sigmoid function and this is to be monitored regularly.

However, in the modern times where data changes in minutes or seconds and class labels much more difficult to predict and assign, the supervised methods may not be that accurate. This created the need for unsupervised learning strategy. The clustering algorithms have become much more accurate tools in the field of fraud detection.

Our project ran using different unsupervised algorithms such as DBSCAN, K-Means etc. While DBSCAN showed a mere accuracy of 0, logistic regression turned out to be highly accurate with accuracy of 0.9. In our case supervised learning algorithms turned out to be lot more fruitful than the unsupervised learning algorithms. There were many classification algorithms that we used such as decision tree, KNN, Naïve Bayes etc. There is no one perfect algorithm. Hence, we followed the approach of combining multiple algorithms for better results, which is commonly referred as ensemble method.

Being implemented in python offers much more flexibility compared to other programming environments as it already has all algorithms implemented in packages such as Scikitlearn, TensorFlow, Pytorch etc. It has many functions to ensure that we can implement any supervised or unsupervised algorithm by knowing the function name and parameters. And hence, we can also predict the accuracy measures of the algorithm using some of those functions as well.

This paper presents a case study involving credit card fraud detection. We demonstrate how a seemingly perfect transactional database may contain a few unnoticeable frauds. It defines the types and sub-types of fraud, the nature of data, performance metrics and other methodologies. It draws its inspiration from many other related fields such as text mining, game theory, firewall breach, intrusion detection etc. These fields are all based on fraud detection in different approaches.

Credit card fraud detection has drawn a lot of research and numerous techniques with special emphasis on neural networks, data mining was suggested. For example, we can consider the neural network proposed by Ghosh and Reilly which is trained on a large sample of labelled credit card transactions. These transactions contain a variety of fraud cases such as lost cards, stolen cards, application fraud, e-mail fraud etc. Training on a variety of data certainly makes the model invulnerable to almost any kind of fraud. Hence, the quality or variety of data matters more than the quantity or bulk.

There were many other approaches in the past and there are going to be many in the future and there is still a lot of scope for this field as the frauds continue to be inevitable. Some of the approaches in the past were meta-classifier-based fraud detection, fraud detection based on historical data, distributed data mining approach, neural classification, web services-based detection, Hidden Markov Model and so on. Especially with voluminous data generated these days almost any algorithm could shine and the frauds of course shine accordingly. And hence this cycle of new frauds - solution appears to be never ending.

1. **Proposed Methodology**
   1. Logistic Regression

Logistic Regression is a popular means of supervised learning which is used to estimate outcomes such as win/loss, positive/negative etc. It makes use of a sigmoid function whose value lies in between 0 and 1. The basic logistic regression model is as follows,

As shown above it makes use of linear regression but it might not be restricted to one variable. It models the probability of example. It is not a classifier although it can be used to make a classifier such as classifying examples with values greater than threshold into one class and the rest into the other.

* 1. Linear Regression

Linear Regression is one of the many supervised approaches to fraud detection. It is used for modelling the relationship between a dependent variable and one or more independent variables. The variables are referred as attributes in data mining context and as features in context of Artificial Intelligence.

Y=a0+a1x

Above equation represents simple linear regression in mathematical form.

However, the general form of linear regression is as follows,

Y=a0+a1x1+a2x2+-------an-1xn-1

Y=A’X

The above representation is the vector representation of linear regression.

* 1. SVM Model (Support Vector Machine)

Support Vector Machines are models based on supervised learning which are used for multiple purposes such as regression, classification and even fraud detection. The given training examples are classified into two classes mainly positive and negative. It is a hyperplane which separates positive class and negative class. The general hyperplane is as follows,

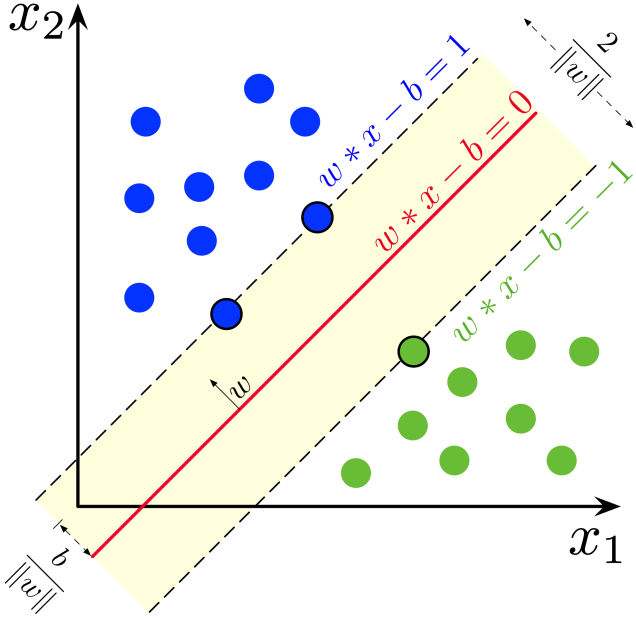


Fig 4.1: Hyperplanes in SVM

The class of interest is our choice. If we choose positive class, we can continue the same positive class, negative class classification on the selected class. This iterative process is continued based on our requirement.

* 1. Decision Tree

Decision Tree is a tree structured algorithm that uses the decisions to predict the class and to conclude the final decisions. It is having three different algorithms like ID3, CART and C4.5. It has an attribute selection measures like Information Gain, Gini Index and Gain Ratio.

The ID3 Algorithms uses Information Gain. whereas the C4.5 use Gain Ratio as it was the extension of Information Gain, and Finally the Gini Index is used in CART.

By using the any of the above algorithms the tree has been generated and then based on the condition the class attribute has been classified. The path from top to bottom of the tree indicates the association rules.

Decision tree helps to determine the worst, best and expected values for different scenarios, simplified to understand and interpret and allows addition of new possible scenarios.

* 1. Random Forest

Random Forest is an ensemble method which relies on averaging a lot of decision trees and is used for classification and regression. Unlike decision trees this method is less prone to overfitting. Random forests are a way of averaging multiple decision trees trained on different parts of the training examples. Its goal is to reduce variance. Although there is a small increase in bias and some loss of interpretability the overall performance is boosted.



Fig 4.2: Overview of Random Forest

* 1. K-NN classifier

K-Nearest Neighbors is an algorithm for classification and regression. A data object is classified using the majority vote of its nearest neighbors. For example, if K=2 then the object is assigned to class of the two nearest neighbors. In order to find out how near a data object is to a neighbor and finally assign it a class, there are several distance measures.

* 1. Naïve Bayes

This Naïve Bayes classifier is based on simplest Bayesian network models. This classifier is highly scalable requiring a number of parameters in a problem. It is based on Bayes theorem on conditional probability and the attributes however independent with each other. The formulation is as shown below,

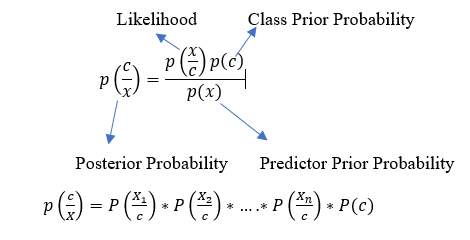


Fig 4.3: Naïve Bayes Terminologies

* 1. K-Means clustering

K-Means is a flavor of DBSCAN that use the measure mean to classify data objects. However, the measure may be median, mode etc. In this case it is mean. The parameter that is required is the value of K. If that is specified then the clustering is performed with each cluster having K items. We categories each data object to its closest mean and then the mean’s coordinates are updated iteratively.

The process is repeated for a given number of iterations. The proximity measures used are many. Some of the popular ones are Cosine distance, Minkowiski distance, Euclidean distance etc.

It is highly accurate in case of large data and is computationally efficient. Its easy implementation and adaptability to new examples makes it much better than man other hierarchical clustering methods.

* 1. DBSCAN

Density Based Spatial Clustering of Applications with Noise is a widely used clustering algorithm in detecting fraud. The key idea is that for each point of a cluster the neighborhood in a given radius has to contain at least a minimum number of data points. It requires two parameters

1. The radius or minimum distance threshold. Every data point must have distance less than or equal to radius.
2. Minpoints or the minimum number of points required inside the radius.

It is one of the commonly used algorithms in case of noisy data and data of any arbitrary shape.

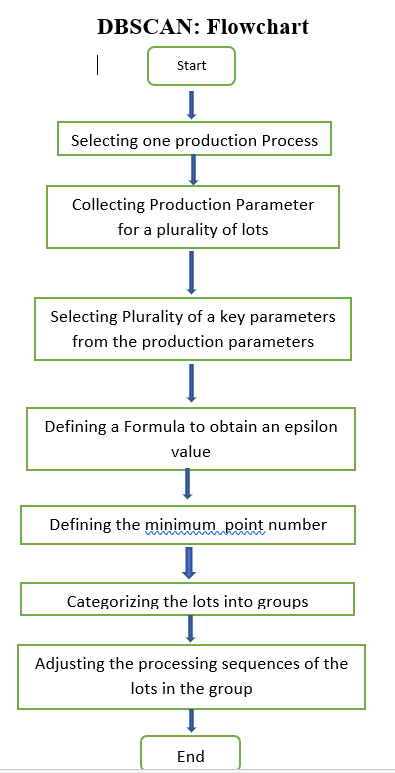


Fig 4.4: Flow chart of DBSCAN Algorithm

1. **Experimentation and Results**

We have implemented different classification and clustering algorithms in python and WEKA tool the results which we obtain are as follows.

Table 5.1: Accuracy of different algorithms

|  |  |  |
| --- | --- | --- |
| S. No | Algorithm | Accuracy |
| 1 | Random Forest | 0.99958 |
| 2 | Logistic Regression | 0.99927 |
| 3 | Naïve Bayes | 0.9778 |
| 4 | Decision Tree | 0.99915 |
| 5 | SVC | 0.99935 |
| 6 | DBSCAN | 0.00093 |
| 7 | K-Means | 0.45828 |

From the Table 4.1 it has been identified that almost all algorithms have high accuracy as the data is unbalanced and it will be having high accuracy in case of classification algorithms and in case of clustering algorithms it is having accuracy value less than 0.5, so the data should be balanced and to balance the imbalanced data we have used *sklearn.imblearn* module to oversample the data so that data will be balanced.

In the next table we are having some other metrics called as Precision, Recall and F1-Score as follows.

Table 5.2: Results of Different Algorithms on different metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S. No | Algorithm | Precision | Recall | F1-Score |
| 1 | Random Forest | 0.9583 | 0.7718 | 0.8550 |
| 2 | Naïve Bayes | 0.0573 | 0.8389 | 0.1073 |
| 3 | Logistic Regression | 0.8785 | 0.6308 | 0.7343 |
| 4 | Decision Tree | 0.6966 | 0.83221 | 0.7584 |

We have performed some algorithms using WEKA tool and the results are as follows.

Table 5.3: Results of Different Algorithms on Weka Explorer

|  |  |  |
| --- | --- | --- |
| S. No | Algorithm | Accuracy |
| 1 | Decision Tree (J48) | 99.921% |
| 2 | Naïve Bayes | 99.6254% |
| 3 | Random Forest | 99.9463% |

1. **Conclusion**

Credit Card Fraud has now become more rampant these days [6]. In order to find these cases and improve the risk management level in the scientific way is an efficient and easy to handle the transactions. In this study, several algorithms were used to analyses the credit card fraud from the past history. So, the users can now use this prototype for the security of these accounts and they can have safe transactions. We present the work which we did in this project and in our case the supervised algorithms perform much better than unsupervised algorithms in terms of anomaly detection.

1. **References**
2. Navanshu Khare, Saad Yunus Sait, Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models, International Journal of Pure and Applied Mathematics (IJPAM), Volume 118 No.20 2018, ISSN:1314-3395, www.ijpam.eu.
3. Dushyant Singh, Saubhagya Vardhan, Dr. Neha Agrawal, Credit Card Fraud Detection Analysis, International Research Journal of Engineering and Technology (IRJET), Volume 5 Issue:11, Nov 2018, ISSN:2395-0056, www.irjet.net.
4. Devi Meenakshi.B, Janani.B, Gayatri.S, Indira.N, Credit Card Fraud Detection Using Random Forest, International Research Journal of Engineering and Technology (IRJET), Volume :6 Issue:3, March2019, ISSN:2395-0056, www.irjet.net.
5. Priyanka Yadav, Pavan Wangade, Manish Thakur, Mohammed Fakih, Gayatri Hegde, Proposed Distributed Data Mining in Credit Card Fraud Detection, International Research Journal of Engineering and Technology (IRJET), Volume :3 Issue: 4, April 2016, ISSN:2395-0056, www.irjet.net.
6. Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi, Real Time Credit Card Fraud Detection Using Machine Learning, IEEE, July 2019. **INSPEC Accession Number:**18868933.
7. Dinesh L.Talekar, K.P.Adhiya, Credit Card Fraud Detection System, International Journal of Modern Engineering Research (IJMER),Volume:4 Issue:9,Sept 2014,ISSN:2249-6645, [www.ijmer.com](http://www.ijmer.com).
8. Franscisca Nonyelum Ogwueleka, Data Mining Application in Credit Card Fraud Detection System, Journal of Engineering Science and Technology, Volume :6 Issue:3, June 2011.
9. Suman, Nutan, Review Paper On Credit Card Fraud Detection, International Journal of Computer Trends and Technology (IJCTT), Volume:4 Issue:7,July 2013, ISSN:2231-2803,[www.ijcttjournal.org](http://www.ijcttjournal.org).
10. Sushmito Gosh, Douglas L.Reilly, Credit Card Fraud Detection with a Neural Network, International Conference on System Sciences,1994.
11. Aihua Shen, Rencheng Tong, Yaochen Deng, Application of Classification Models on Credit Card Fraud Detection, IEEE 2007.
12. Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar, Credit Card Fraud Detection Using Hidden Markov Model, IEEE Transactions on Dependable and Secure Computing, Volume:5 Issue:1, Jan 2008.
13. S. Benson Edwin Raj, A. Annie Portia, Analysis on Credit Card Fraud Detection Methods, International Conference on Computer, Communication and Electrical Technology (ICCCET),18th March 2011.
14. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, Asoke K. Nandi, Credit Card Fraud Detection Using AdaBoost and Majority Voting, IEEE 2018.
15. Shaiesh S. Dhok, Credit Card Fraud Detection using Hidden Markov Model, International Journal of Soft Computing and Engineering (IJSCE),Volume: 2 Issue:1,March 2012,ISSN:2231-2307.