International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 1244-1250

Credite Card Fraud Detection System

Meiyalakan. 1st, Moulitharan.K 2nd, Nandhakumar.S 3rd , Murali.K 4th, Tamilselvan.J 5th

1st Assistant Professor, 2nd, 3rd, 4th, 5th UG Scholar (B.E), Department of Computer Science and Engineering, Mahendra Institute of Technology, Mahendhirapuri.

Abstract

This work entitled as "CREDIT CARD FRAUD DETECTION". It deals with minimization of problem posed by the old magnetic stripe card technology. For this purpose we are using EMV (Europay -Master card- Visa) chip card design in the credit card business.This EMV chip card technology efficiently deduces the conflicts and problems in the old magnetic stripe card method. There must be detection methods available like fall back in which the technology will fail. WEKA is a data mining tool which is used to classify the transaction tool. In our theses we are using the above for model creation and evaluated were Naïve Bayes,oneR,Logistic Regression, and our proposed J48 and the appropriate algorithm for my study were selected from their accuracy value it would consider a best algorithm for our study. My study value shows that the J48 is more fit in understanding the transaction logs data.

Keywords: *credit card fraud, fraudulent activities, Random Forest, Adaboost, ROC curve*

Date of Submission: 06-06-2022 Date of acceptance: 21-06-2022

I. INTRODUCTION

Credit card fraud is a growing concern in the present world with the growing fraud in the government offices, corporate industries, finance industries, and many other organizations. In the present world, the high dependency on the internet is the reason for an increased rate of credit card fraud transactions but the fraud has increased not only online but also offline transactions. Though the data mining techniques[6] are used the result is not much accurate to detect these credit card frauds. The only way to minimize these losses is the detection of the fraud using efficient algorithms which is a promising way to reduce the credit card frauds. As the use of the internet is increasing[Figure.1], a credit card is issued by the finance company. Having a credit card means that we can borrow the funds. The funds can be used for any of the purposes. When coming to the issuance of the card, the condition involved is that the cardholder will pay back the original amount they borrowed along with the additional charges they agreed to pay.

II. EXISTING METHOD

In this current work the Naïve Bayes machine learning classifier attempts to anticipate a class which is known as result class dependent on probabilities, and furthermore restrictive probabilities of its event from the preparation information. This sort of learning is exceptionally proficient, quick and high in exactness for true situations, and furthermore this learning type is known as directed learning. The usage of Naïve Bayes and oneR calculation on same Visa dataset to compute the exactness of calculations to recognize the deceitful exchanges in the dataset. Test results portray that the two classifiers works contrastingly for the equivalent dataset. The reason for existing is to improve the exactness, precision and increment the adaptability of the calculation . Bayesian system classifiers are exceptionally well known in the territory of machine learning and it goes under the class of regulated order models. Guileless Bayes classifier is likewise a notable Bayesian Network that depends on Bayes hypothesis of restrictive likelihood and thus, is a classifier dependent on likelihood which considers Naïve i.e., solid freedom presumption .It was earlier presented with some other name, into the content recovery network as a standard system for sorting content in light of the fact that there was an issue of choosing in which class the archives do has a place with, with word frequencies as the element. The Naïve Bayes machine learning classifier endeavors to anticipate a class which is known as result class dependent on probabilities, and furthermore contingent probabilities of how frequently it happened from the preparation information. This sort of learning is extremely productive, quick and high in exactness for true situations, and is known as directed learning. Likewise, this is exceptionally effective on the grounds that it evaluates the parameters by utilizing little preparing information which is utilized for characterization and depends on word autonomy. In spite of the fact that Naïve Bayes is very easy to execute and comprehend and utilizes solid suspicions. It gives entirely precise outcomes and furthermore it has been demonstrated again and again the time that Naïve Bayes works viably in different territories identified with machine learning

IV. PROPOSED SYSTEM

The main aim of this paper is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithms like that the Random Forest and the Adaboost algorithms. Then these two algorithms are compared to choose the algorithm that best detects the credit card fraud transactions. The process flow for the credit fraud detection problem [Figure.3.]includes the splitting of the data, model training, model deployment, and the evaluation criteria. Figure.3 Process Flow The detailed architecture diagram for the credit card fraud detection system [Figure. 4.] includes many steps from gathering dataset to deploying model and performing analysis based on results. In this model we take the Kaggle credit card fraud dataset and pre-processing is to be done for the dataset. Now to prepare the model we have to split the data into the training data and the testing data. We use the training data to prepare the Random Forest and the Adaboost models. Then we develop both the models. Finally, the accuracy, precision, recall, and F1-score is calculated for bot the models. Finally the comparison of the credit card fraud transactions more accurately

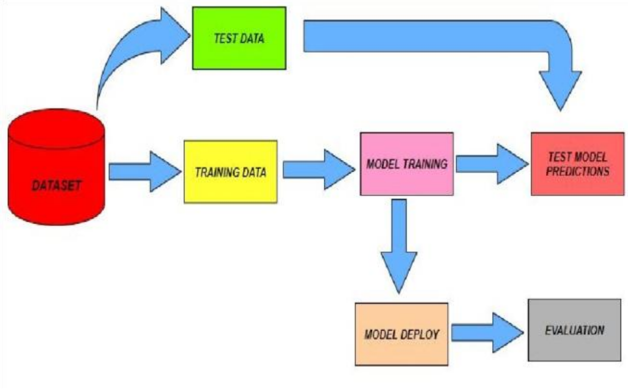


Figure 1. Process flow diagram

The detailed architecture diagram for the credit card fraud detection system [Figure. 4.] includes many steps from gathering dataset to deploying model and performing analysis based on results. In this model we take the Kaggle credit card fraud dataset and pre-processing is to be done for the dataset. Now to prepare the model we have to split the data into the training data and the testing data. We use the training data to prepare the Random Forest and the Adaboost models. Then we develop both the models. Finally, the accuracy, precision, recall, and F1-score is calculated for bot the models. Finally the comparison of the credit card fraud transactions more accurately.

4.1. Random forest algorithm:

The Random Forest algorithm [Figure. 5]is one of the widely used supervised learning algorithms. This can be used for both regression and classification purposes. But, this algorithm is mainly used for classification problems. Generally, a forest is made up of trees and similarly, the Random Forest algorithm creates the decision trees on the sample data and gets the prediction from each of the sample data. Then Random Forest algorithm is an ensemble method. This algorithm is better than the single decision trees because it reduces the over-fitting by averaging the result.

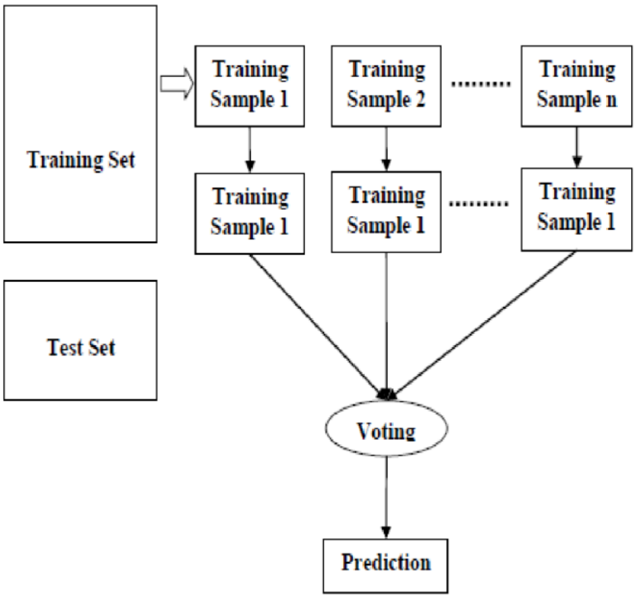


Figure.2.Random forest algorithm

1. Take the Kaggle credit card fraud dataset that is trained and randomly select some of the sample data.
2. Using the randomly created sample data now creates the Decision Trees that are used to classify the cases into the fraud and non-fraud cases.
3. The Decision Trees are formed by splitting the nodes, the nodes which have the highest Information gain make it as the root node and classify the fraud and non-fraud cases.
4. Now the majority vote is performed and the decision Trees may result in 0 as output which includes that these are the non-fraud cases.
5. Finally, we find the accuracy, precision, recall, and F1 -score for both the fraud and non-fraud cases.

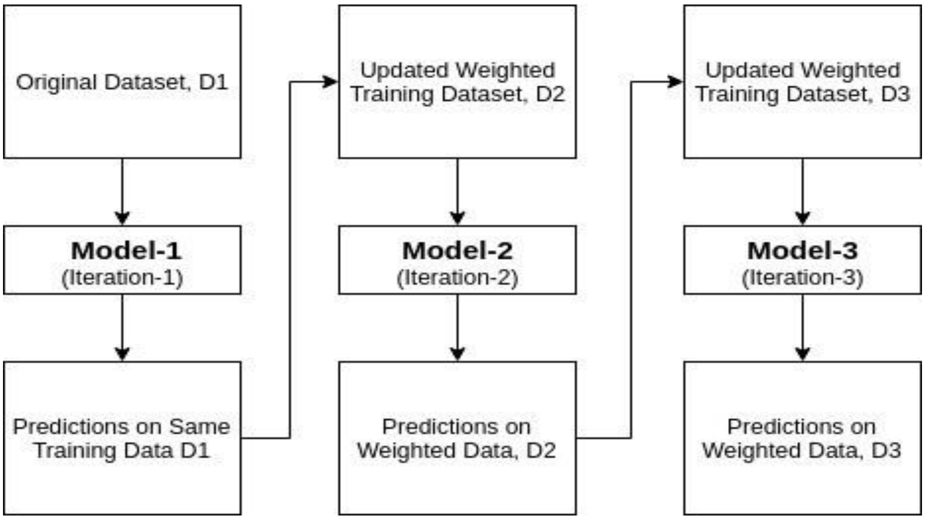
Random Forest algorithm

```
 Algorithm Random Forest : To generate  c classifiers:
              For i=1 to c do
              Randomly select the training data D with
              replacement to produce Di
               Create a root node N containing Di and cell
Build Tree(N) End for
Majority Vote

Build Tree(N)
             Randomly select x% of all the possible            splitting
features in N
            Select the features F that has the highest Information
             A gain for further splitting
            Gain (T,X)=Entropy (T)-Entropy(T,X)
             Now to calculate the entropy we use,
             $E(S) = \sum_{i=1}^{c}(-Pi \log Pi)$
             Create f child nodes
 For i=1 to f do   Set contents f N to Di
                         Call Build Tree(Ni)
             End for
 End
```

4.2.Adaboost algorithm:

Boosting is one of the ensemble techniques. This algorithm is used to build strong classifiers from weaker classifiers. This can be done by building a strong model by using a weak model in the series. Initially, a model is built from the training data. Then the second model is built from the first model by correcting the errors that represent in the model that is created before. This is a repetitive process and is continued until either the maximum number of models is added or the complete training dataset is predicted correctly. Adboost was one of the most successful boosting algorithms that were developed for the binary classification.

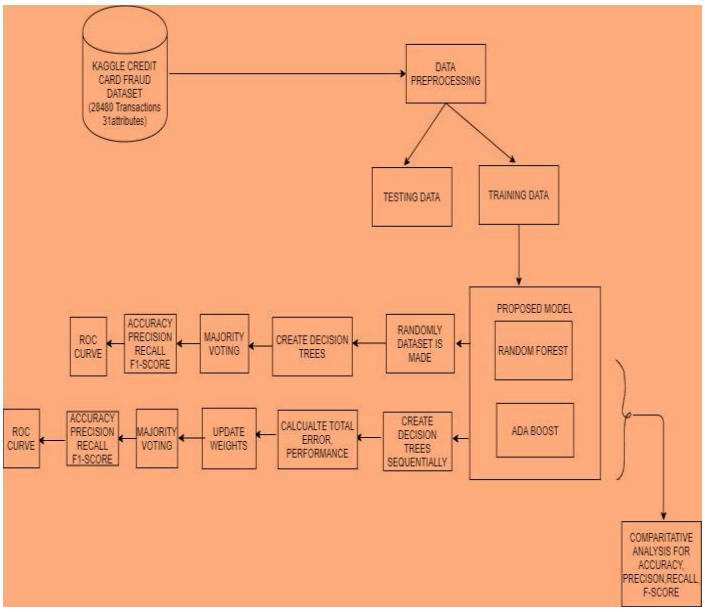

Original Dataset, D1
Updated Weighted Training Dataset, D2
Updated Weighted Training Dataset, D3
Model-1 (Iteration-1)
Model-2 (Iteration-2)
Model-3 (Iteration-3)
Predictions on Same Training Data D1
Predictions on Weighted Data, D2
Predictions on Weighted Data, D3

The short name for Adaboost is adaptive boosting. It is best used with weak learners. This Adaboost boosting technique [Figure. 6]combines the multiple weak classifiers into a strong classifier. Adaboost algorithm can be used with short decision trees. The way the Adaboost is created is such that initially at first the nodes are created and the tree is made, then the performance of the tree on each of the instances is checked. Also, a weight is assigned. The training data that is hard to predict is the one that gives more weight. The Adaboost algorithm is a powerful classifier that works well on both the basic and complex problems. The disadvantage of this algorithm is that this algorithm is mostly sensitive to noisy data. This algorithm is also sensitive to outliers.

Steps for Adaboost Algorithm

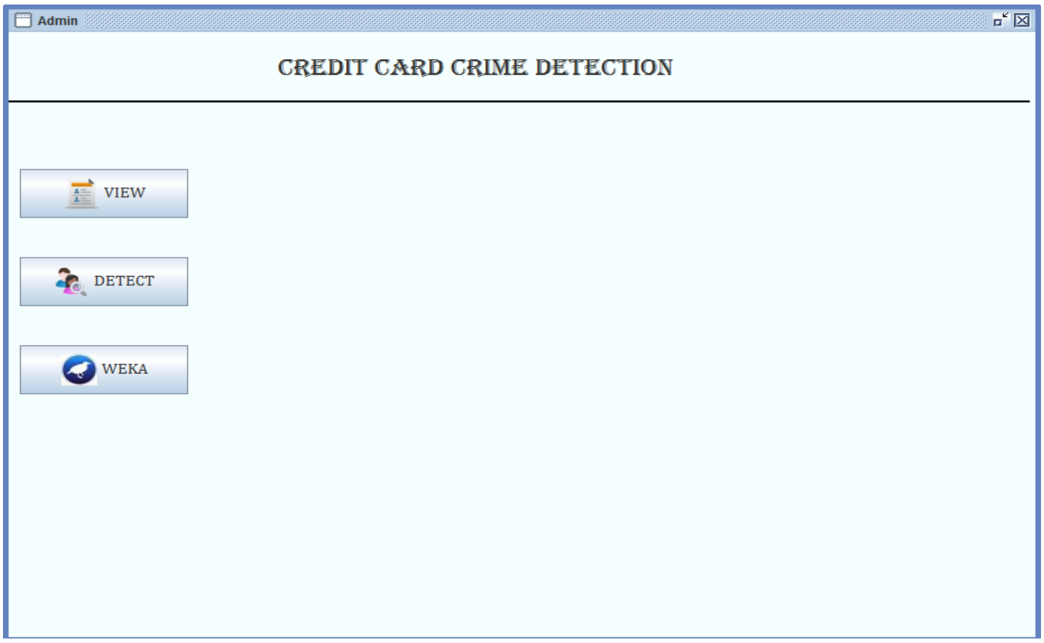
1. The Kaggle credit card fraud dataset is taken and is trained. Randomly select some of the sample data.
2. Using the randomly created sample data now creates the decision trees sequentially for classifying the fraud and non-fraud cases.
3. The decision trees are formed initially. This can be done by splitting the node based on which has the highest information gain, make it as the root node, and classify the fraud and non-fraud cases.
4. Now calculate the error rate, performance, and update the weights of the fraud and non-fraud transactions that are incorrectly classified.
5. Now majority vote is performed and the decision trees may result as output which indicates the nonfraud cases.
6. The decision trees may output 1 which indicates that it is a fraud case.
7. Finally, we find the accuracy, precision, recall, and F1-score for both the fraud and non-fraud cases.

4.4 Architecture Diagram


KAGGLE CREDIT CARD FRAUD DATASET (28480 Transactions 31attributes)
DATA PREPROCESSING
TESTING DATA
TRAINING DATA
PROPOSED MODEL
RANDOM FOREST
ADA BOOST
RANDOMLY DATASET IS MADE
CREATE DECISION TREES
MAJORITY VOTING
ACCURACY PRECISION RECALL F1-SCORE
ROC CURVE
CREATE DECISION TREES SEQUENTIALLY
CALCUALTE TOTAL ERROR, PERFORMANCE
UPDATE WEIGHTS
MAJORITY VOTING
ACCURACY PRECISION RECALL F1-SCORE
ROC CURVE
COMPARITATIVE ANALYSIS FOR ACCURACY, PRECISON,RECALL, F-SCORE

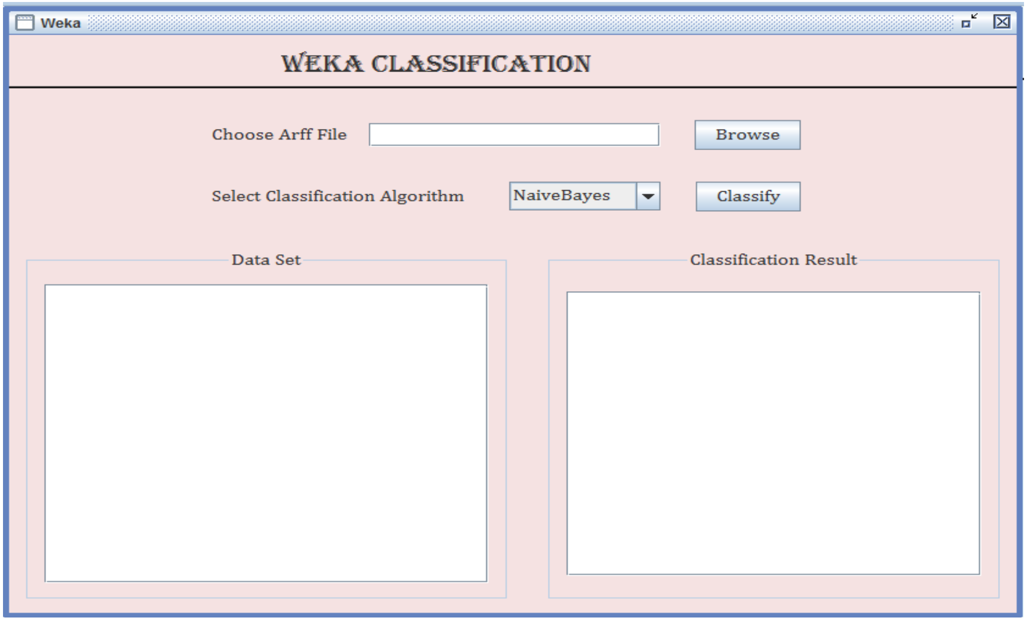
V.KEY RESULTS

5.1 Main widow frame:



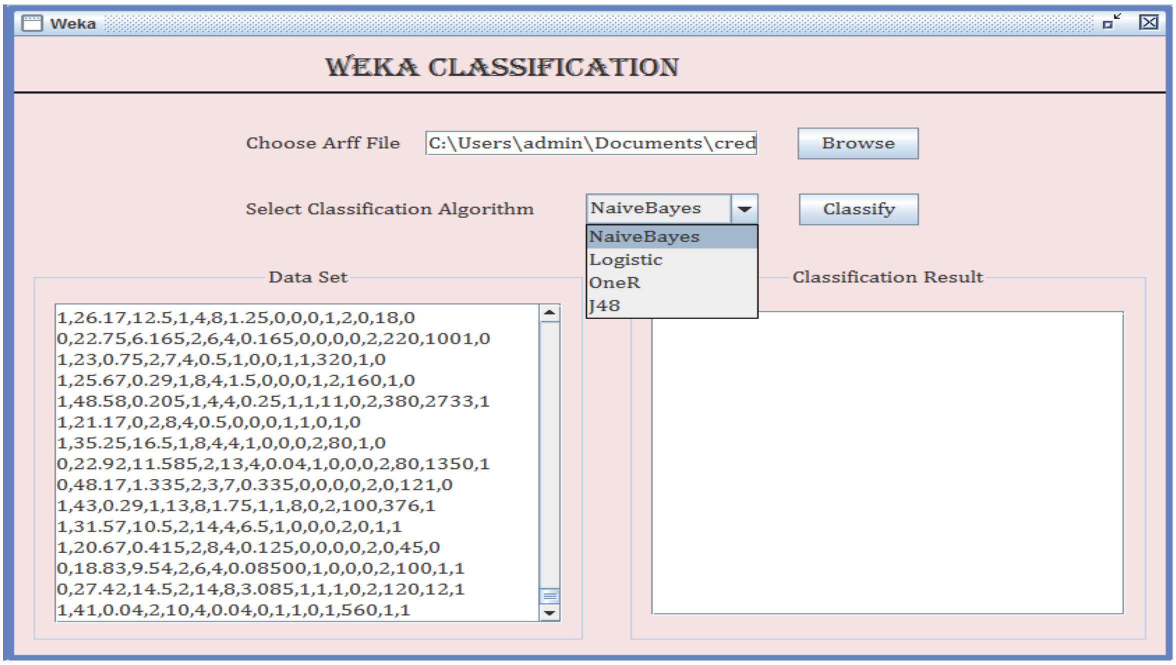
5.2 Classification And Detection Frame:

This form helps to load credit card dataset as Attribute related file format and to classify using list of algorithm with our proposed system.



5.3. Detectin using proposed work:

This form helps to select and classify credit card dataset using our proposed work.



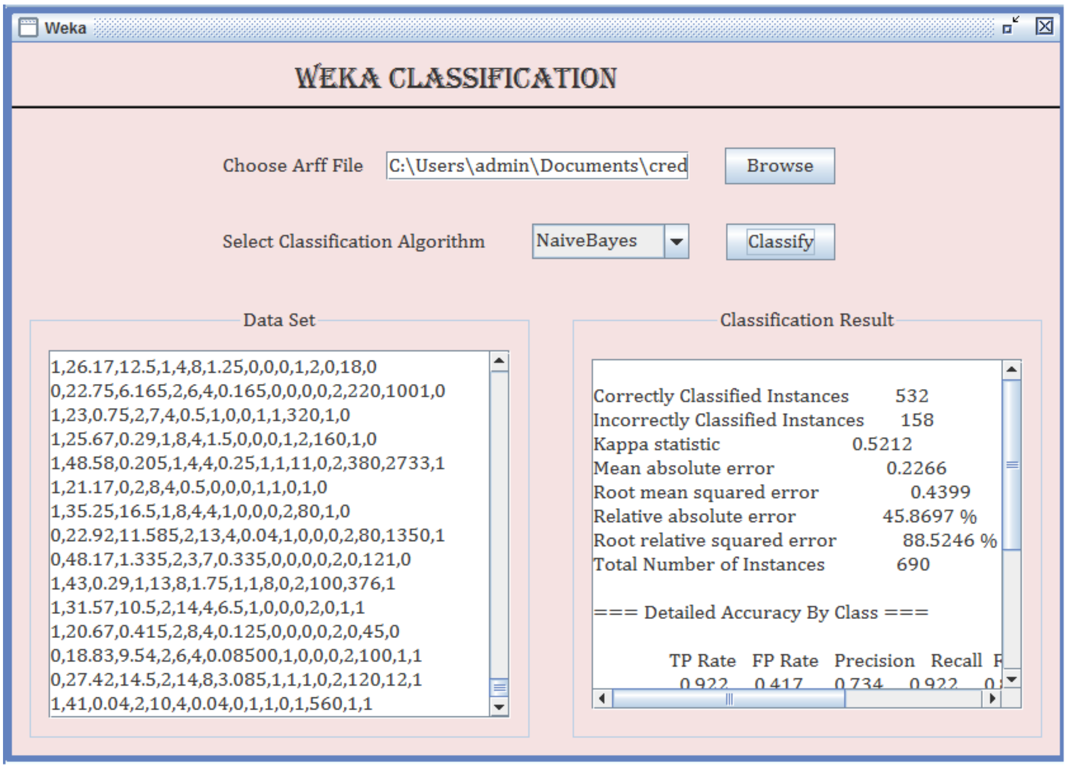
Choose Arff File: C:\Users\admin\Documents\cred

Select Classification Algorithm: NaiveBayes

```
1,26.17,12.5,1,4,8,1.25,0,0,0,1,2,0,18,0
0,22.75,6.165,2,6,4,0.165,0,0,0,0,2,220,1001,0
1,23,0.75,2,7,4,0.5,1,0,0,1,1,320,1,0
1,25.67,0.29,1,8,4,1.5,0,0,0,1,2,160,1,0
1,48.58,0.205,1,4,4,0.25,1,1,11,0,2,380,2733,1
1,21.17,0,2,8,4,0.5,0,0,0,1,1,0,1,0
1,35.25,16.5,1,8,4,4,1,0,0,0,2,80,1,0
0,22.92,11.585,2,13,4,0.04,1,0,0,0,2,80,1350,1
0,48.17,1.335,2,3,7,0.335,0,0,0,0,2,0,121,0
1,43,0.29,1,13,8,1.75,1,1,8,0,2,100,376,1
1,31.57,10.5,2,14,4,6.5,1,0,0,0,2,0,1,1
1,20.67,0.415,2,8,4,0.125,0,0,0,0,2,0,45,0
0,18.83,9.54,2,6,4,0.08500,1,0,0,0,2,100,1,1
0,27.42,14.5,2,14,8,3.085,1,1,1,0,2,120,12,1
1,41,0.04,2,10,4,0.04,0,1,1,0,1,560,1,1
```

5.4. Result Analysis Form:

In this form that helps to predict the result of classification using our proposed work.



Choose Arff File: C:\Users\admin\Documents\cred

Select Classification Algorithm: NaiveBayes

```
1,26.17,12.5,1,4,8,1.25,0,0,0,1,2,0,18,0
0,22.75,6.165,2,6,4,0.165,0,0,0,0,2,220,1001,0
1,23,0.75,2,7,4,0.5,1,0,0,1,1,320,1,0
1,25.67,0.29,1,8,4,1.5,0,0,0,1,2,160,1,0
1,48.58,0.205,1,4,4,0.25,1,1,11,0,2,380,2733,1
1,21.17,0,2,8,4,0.5,0,0,0,1,1,0,1,0
1,35.25,16.5,1,8,4,4,1,0,0,0,2,80,1,0
0,22.92,11.585,2,13,4,0.04,1,0,0,0,2,80,1350,1
0,48.17,1.335,2,3,7,0.335,0,0,0,0,2,0,121,0
1,43,0.29,1,13,8,1.75,1,1,8,0,2,100,376,1
1,31.57,10.5,2,14,4,6.5,1,0,0,0,2,0,1,1
1,20.67,0.415,2,8,4,0.125,0,0,0,0,2,0,45,0
0,18.83,9.54,2,6,4,0.08500,1,0,0,0,2,100,1,1
0,27.42,14.5,2,14,8,3.085,1,1,1,0,2,120,12,1
1,41,0.04,2,10,4,0.04,0,1,1,0,1,560,1,1
```

```
Correctly Classified Instances       532
Incorrectly Classified Instances      158
Kappa statistic                     0.5212
Mean absolute error                     0.2266
Root mean squared error                 0.4399
Relative absolute error               45.8697 %
Root relative squared error             88.5246 %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

          TP Rate   FP Rate   Precision   Recall  F
           0.922     0.417     0.734     0.922    0.
```

VI. FUTURE ENHANCEMENT

From the above analysis, it is clear that many machine learning algorithms are used to detect the fraud but we can observe that the results are not satisfactory. So, we would like to implement deep learning algorithms to detect credit card fraud accurately.

VII. CONCLUSION

Even though there are many fraud detection techniques we can't say that this particular algorithm detects the fraud completely. From our analysis, we can conclude that the accuracy is the same for both the Random Forest and the Adaboost algorithms. When we consider the precision, recall, and the F1-score the Random Forest algorithm has the highest value than the Adaboost algorithm. Hence we conclude that the Random Forest Algorithm works best than the Adaboost algorithm to detect credit card fraud.

.REFERENCES

[1]. Aleskerov.E, Freisleben.B, And Rao.B, "CARDWATCH: A Neural Network Based Database Mining System For Credit Card Fraud Detection", Proc. IEEE/IAFE: Computational Intelligence For Financial Engineering, Pp. 220-226, 1997.

[2]. Brause.R, Langsdorf.T, And Hepp.M, "Neural Data Mining For Credit Card Fraud Detection," Proc. IEEE Int'l Conf. Tools With Artificial Intelligence, Pp. 103-106,

[3]. Ghosh.S, And D.L.Reilly, "Credit Card Fraud Detection With A Neural-Network", Proc. 27thhawaiiinternational Conference On System Sciences: Information Systems: Decision Support And Knowledge-Based Systems, Vol. 3, Pp. 621-630, 1994.

[4]. Phua.C, Alahakoon.D, And Lee.V, "Minority Report In Fraud Detection: Classification Of Skewed Data," ACM SIGKDD Explorations Newsletter, Vol. 6, No. 1, Pp. 50-59, 2004.

[5]. Phua.C, Lee.V, Smith.K, And Gayler.R, "A Comprehensive Survey Of Data Mining-Based Fraud Detection

[6]. Statistic Brain Research Institute (2014, July 12). Credit Card Fraud Statistics (2014).

[7]. Steven J. Murdoch, Saar Drimer, Ross Anderson, And Mike Bond, "Chip And PIN Is Broken" In IEEE Symposium On Security And Privacy, 2010

[8]. Stolfo.L And A.L. Prodromidis, "Agent-Based Distributed Learning Applied To Fraud Detection," Technical Report CUCS-014-99, Columbia Univ., 1999.

[9]. Tej Paul Bhatla, Vikramprabhu, & Amit Dua, "Understanding Credit Card Frauds".Tata Consultancy Services.

[10]. Varun Chandola, Arindam Banerjee, & Vipin Kumar, "Anomaly Detection: A Survey" In ACM Computing Surveys, 2009 ©, Vol. 41, No. 3, Article 15, Pp 15:1 - 15:58

[11]. Wen-Fang Yu & Na Wang, "Research On Credit Card Fraud Detection Model Based On Distance Sum" In International Joint Conference On Artificial Intelligence,

[12]. Philip K Chan, Wei Fan, Andreas Prodomidis, Salvatore Stolfo, "Distributed Data Mining In Credit Card Fraud Detection". Copyright 1999.