AN EFFICIENT CREDIT CARD FRAUD CLASSIFIER OF THE FOUR DATA MINING CLASSIFICATION ALGORITHMS-A COMPARATIVE ANALYSIS

[1] M.V.Jisha , [2]Dr. D .Vimal Kumar
[1] PhD Full Time Research Scholar, [2] Professor, Department of computer science.
[1]Department of computer science.
[1]Nehru arts and Science College, Coimbatore, Tamil Nadu, India.

Abstract: Credit card payment has now become popular .Credit card is an easiest way to pay directly through your bank account. In every mode of transaction method ,there is pros and cons, there exists many credit cards fraud committed online , via web ,phone shopping or card holder-not-present. However ,there is very limited information available to distinguish dynamic fraud from genuine customer behavior in such an extremely sparse and imbalanced data environment, which makes the instant and effective detection become more and more important and challenging. In this paper , a comparative analysis is done using four data mining classification algorithms, namely, Random Forest, Logistic Regression ,K-nearest neighbor and Support Vector Machine, to detect the efficient algorithm which is having high accuracy in detecting the credit card fraud, as the transaction increases day by day.

Index Terms: ***Random Forest, Logistic Regression , K-nearest neighbor and Support Vector Machine, confusion matrix.***

1. INTRODUCTION

Real time credit card is most suitable mode of payment for both online as well as daily purchases. Credit card frauds are mainly of two types: first is Offline frauds. This type detects the stolen credit card at storefront or call center and uses the personal details. Second, is the Online fraud . This fraud detected via internet, mobile phone ,e-mails ,web, shopping .Credit card fraud uses credit card details like cardholder names ,credit card pin ,credit card cvv [20]etc. Only the card's details are need, and a manual signature and card imprint are not required at the time of purchase.

Mainly there are different classification algorithms in data mining to detect the fraud in Credit card transactions. Fraud detection detects the data streams of transactions and learns the fraud's patterns. A fraud shows a small fraction of the daily transactions. There are many Statistical as well as machine learning techniques for detecting the frauds which classes ,clusters or associate the patterns in the transactions. These techniques study the unusual behavior of the patterns in the imbalanced data. It may be done under supervised learning or un supervised. In this paper, four classification algorithms are tested for the dataset , to detect the frauds, i.e. Random Forest ,Logistic Regression ,support Vector Machine and K- nearest neighbor[20].

Rest of this paper is organized as follows. Section II presents an overview of the algorithms used in the work. Section III provides the description of the data set used. Section IV gives the information about the tool used in the work. Section V describes the methodology .The last section VI gives the Conclusion and the future work of my study.

II . PROPOSED WORK

2.1 Random Forest

The random forest is an ensemble approach that can also be thought of as a form of nearest neighbor predictor. Ensembles are a divide and conquer approach used to improve performance. The main principle behind the ensemble methods is that a group of "weak learners" can come together to form a "strong learner". Each classifier , individually , is a weak learner, while all the classifier taken together are a strong learner. Random Forest runs quite fast and they are able to deal with unbalanced and missing data. Its weakness are that when used for regression they cannot predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy. Of course, the best test of any algorithm is how well it works upon the data set.

Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

2.2 Support Vector Machine

SVM has attracted a great deal of attention in the last decade. It also applied to various domains applications. SVMs are used for learning classification, regression or ranking function. SVM is based on statistical learning theory and structural risk minimization principle. And have the aim of determining the location of decision boundaries. It is also known as a hyper plane . That produces the optimal separation of classes. There by creating the largest possible distance between the separating hyper plane. Further, the instances on either side of it have been proven. That is to reduce an upper bound on the expected generalization error. it decides the target variable value for future predictions. We should decide upon a hyper plane that maximizes the margin. The advantage of this is that they can make use of certain kernels to transform the problem. Such that we can apply linear classification techniques to non-linear data.

Applying the kernel equations. That arranges the data instances in a way within the multi-dimensional space. That there is a hyper plane that separates data instances of one kind from those of another. The kernel equations may be any function. That transforms the non-separable data in one domain into another domain. In which the instances become separable. Kernel equations may be linear, quadratic, Gaussian, or anything else. That achieves this particular purpose.

2.3 Logistic Regression

Logistic Regression is one of the machine learning algorithm after linear regression. In a lot of ways, linear regression and logistic regression are similar. But, the biggest difference lies in what they are used for. Linear regression algorithms are used to predict/forecast values but logistic regression is used for classification tasks. There are many classification tasks done routinely by people. For example, classifying whether an email is a spam or not, classifying whether a tumour is malignant or benign, classifying whether a website is fraudulent or not, etc. These are typical examples where machine learning algorithms can make our lives a lot easier. A really simple, rudimental and useful algorithm for classification is the logistic regression algorithm. Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be class variable, i.e. 0-no, 1-yes. Therefore, we are squashing the output of the linear equation into a range of [0,1]. To squash the predicted value between 0 and 1, we use the sigmoid function.

Linear Equation and Sigmoid Function

$$Z = \Theta_0 + \Theta_1 .x_1 + \Theta_2 .x_2 + \ldots \quad (1)$$

$$g(x) = \frac{1}{1+e^{-x}} \quad . \quad (2)$$

Squashed output-h

$$h = g(z) = \frac{1}{1+e^{-z}} \quad . \quad (3)$$

We take the output (z) of the linear equation and give to the function g(x) which returns a squashed value h, the value h will lie in the range of 0 to 1.

2.4 K- nearest Neighbour

K-NN is the classification of object performed on the basis of closeness of training data available within the feature based is called KNN algorithm. It is lazy learning algorithm known as instance based learning which utilize in order to perform regression. With the given labelled location of training data, the space is divided into regions. If there is most frequent class available amongst the KNN to which class the points in space is assigned. If there is numerical values given the Euclidean distance is used

to estimate distance metric[9]. KNN classifier are rely upon learning by relationship, it implies by contrasting a given test information and preparing test which is like it. KNN consists of two processes: distance ranking and distance computing.

The various phases of KNN are training phase ,testing phase, classification phase. In training phase, the algorithm only stores the feature vectors and corresponding class labels. In testing phase, the algorithm make the decisions on the basis of the training set. In classification phase, a single number is given to k, which decides how many neighbours influence the classification. The value of k can be large or small. If k=1, considered as the nearest neighbour. If k is large then it reduces the effect of noise on classification.

III. DATA SET DESCRIPTION

The data set of credit card fraud detection from Kaggle.com repository is used in this work. It contains 3.9 lacs of data. The total numbers of attributes are 31 in this data set. There are 463 frauds in this dataset.All the attributes are used as input to the respective algorithms. The default payment next attribute is class identifier with the value "0" indicates no fraud and value "1" indicates presence of fraud.
Predictable attributes:
Label
Value: 0= No Fraud

Value: 1= Having Fraud

Input attributes:

1. Amount: It includes the amount of given credit.

2. Time: First and current transaction timing elapse.

3. Class: Response Variable (1-Fraud,0-No Fraud)

3. V_1 TO V_{28} Variables: Principal Components through PCA(Principal Component Analysis (Dimensionality –reduction technique)

The required libraries are added. SMOTE Function is used to make the imbalanced data to balanced data.

Few libraries mentioned below:

library (CARET) for classification and regression.

library(ggplot)for plotting the graphs.

library(ROSE) for random over sampling.

library(CaTools) for classification tools and statistical analysis.

IV. TOOL USED

R studio software is used in this work for the analysis of various algorithms. In R studio it is very easy to install required packages because of its user friendly behaviour. It is an open source integrated development environment (IDE) for R programming language. R language provides a wide variety of statistical and modern graph techniques. It is very easy to understand and

implanting a code with this tool. At backend NoSql is used for storing and processing the database. NoSql provides highly reliable, flexible and available data management services and play an important role in database world[1] [4][20].

V. METHODOLOGY USED

In this work, I used four different classification algorithms for the fraud detection. They are Random Forest, Logistic Regression , K-nearest neighbour and Support Vector Machine and then analysed the results. Each algorithms used the above mentioned attributes ,four metrics are measured and compared ,their accuracy, sensitivity, specificity , precision and AUC (Area under curve),(in percentage).For detecting ,the whole transaction is to detect 25 % fraud, 50 % fraud, 75 % fraud and 100% fraud respectively using each algorithm.[1]

Name of the algorithm	Accuracy (100% fraud)
Random Forest	97.15
Logistic Regression	95.09
Support Vector Machine	96.12
K nearest neighbour	68.62

5.1. Confusion Matrix:

The Confusion matrix is one of the most intuitive and easiest (unless of course, you are not confused)metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.

Before dividing into what the confusion matrix is all about and what it conveys, Let's say we are solving a classification problem where we are predicting whether a transaction is fraud or not. The value 1 represent the fraud and value 0 represents the no fraud ,specified in the attribute class.The confusion matrix, is a table with two dimensions ("Actual" and "Predicted"), and sets of "classes" in both dimensions. Our Actual classifications are columns and Predicted ones are Rows.

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Terms associated with Confusion matrix:

1.True Positives (TP): True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True).
2. True Negatives (TN): True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False).
3. False Positives (FP): False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. (1).
4. False Negatives (FN): False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0)
Sensitivity compares the amount of items correctly as fraud to the amount incorrectly listed as fraud, also known as the ratio of true positives to false positives. Specificity refers to the same with legitimate transactions, or the comparisons of true negatives to false negatives.

Formulas for detecting the above mentioned factors are:

1. Accuracy: $\frac{TP+TN}{TP+FP+FN+TN}$

2. Sensitivity : $\frac{TP}{TP+FN}$

3. Specificity : $\frac{TN}{TN+FP}$

4. Precision : $\frac{TP}{TP+FP}$

5. AUC : The true positive rates are plotted against false positive rates.

5.2 Table representing the statistical analysis

Table 1: Analysis of detecting 25% and 50% fraud

Measures	*25% of fraud*				*50% of fraud*			
	RF	*LR*	*SVM*	*KNN*	*RF*	*LR*	*SVM*	*KNN*
Accuracy	*0.9585*	*0.9521*	*0.9617*	*0.6262*	*0.9827*	*0.9686*	*0.9686*	*0.6311*
Sensitivity	*0.9452*	*0.9810*	*0.9467*	*0.5959*	*0.9702*	*0.9539*	*0.9700*	*0.5894*
Specificity	*0.9701*	*0.9226*	*0.9755*	*0.6527*	*0.9940*	*0.9862*	*0.9674*	*0.6687*
Precision	*0.9650*	*0.9281*	*0.9726*	*0.6000*	*0.9932*	*0.9881*	*0.9636*	*0.6159*
AUC	*0.9590*	*0.9520*	*0.9610*	*0.6240*	*0.9830*	*0.9700*	*0.9670*	*0.6300*

Table 2: Analysis of detecting 75% and 100% fraud

Measures	*75% of fraud*				*100% of fraud*			
	RF	*LR*	*SVM*	*KNN*	*RF*	*LR*	*SVM*	*KNN*
Accuracy	*0.9755*	*0.9370*	*0.9658*	*0.6564*	*0.9715*	*0.9509*	*0.9612*	*0.6862*
Sensitivity	*0.9511*	*0.9180*	*0.9816*	*0.6222*	*0.9498*	*0.9315*	*0.9713*	*0.6349*
Specificity	*0.9979*	*0.9100*	*0.9523*	*0.6879*	*0.9924*	*0.9740*	*0.9522*	*0.7339*
Precision	*0.9977*	*0.9651*	*0.9467*	*0.6481*	*0.9914*	*0.9771*	*0.9474*	*0.6893*
AUC	*0.9770*	*0.9390*	*0.9670*	*0.6560*	*0.9730*	*0.9530*	*0.9620*	*0.6870*

5.2 Screen shots of various algorithms:

KNN Algorithm:

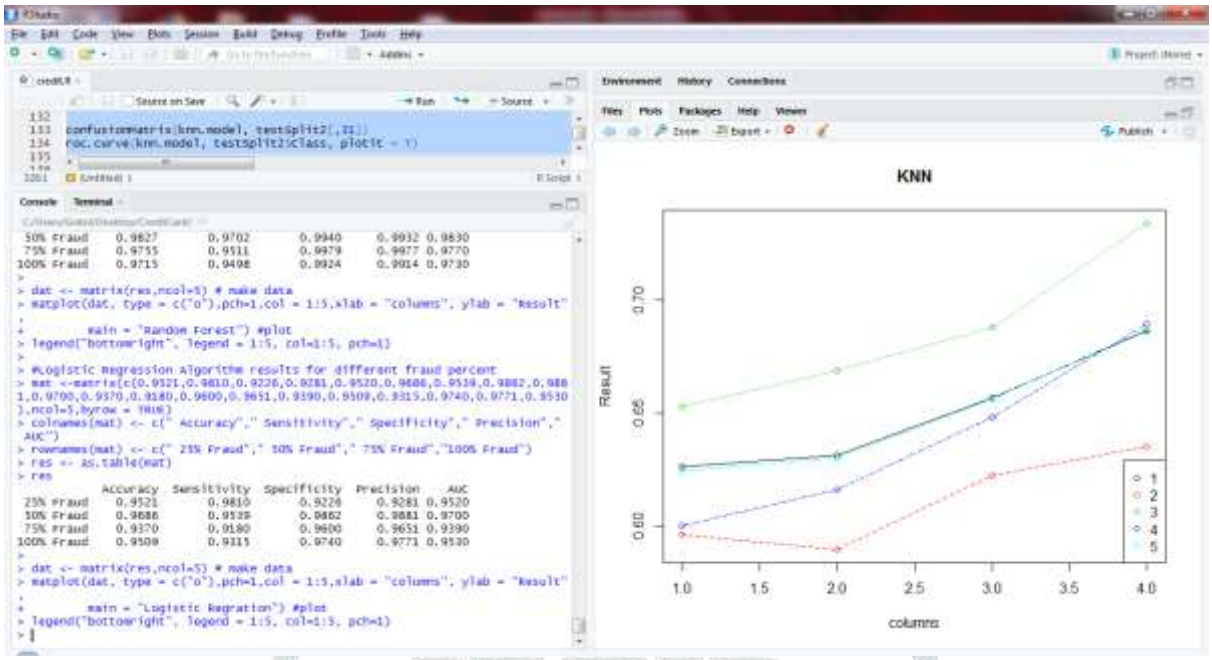

Fig 1: Representation of KNN

SVM Algorithm:

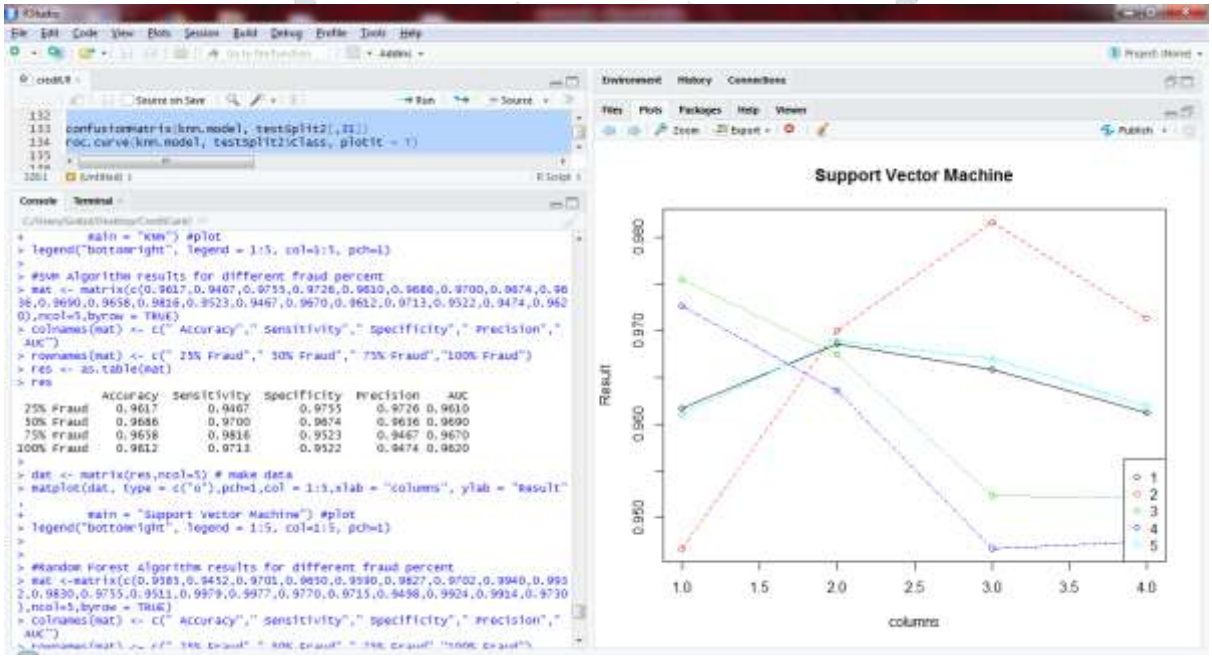

Fig 2: Representation of SVM

Random Forest Algorithm:

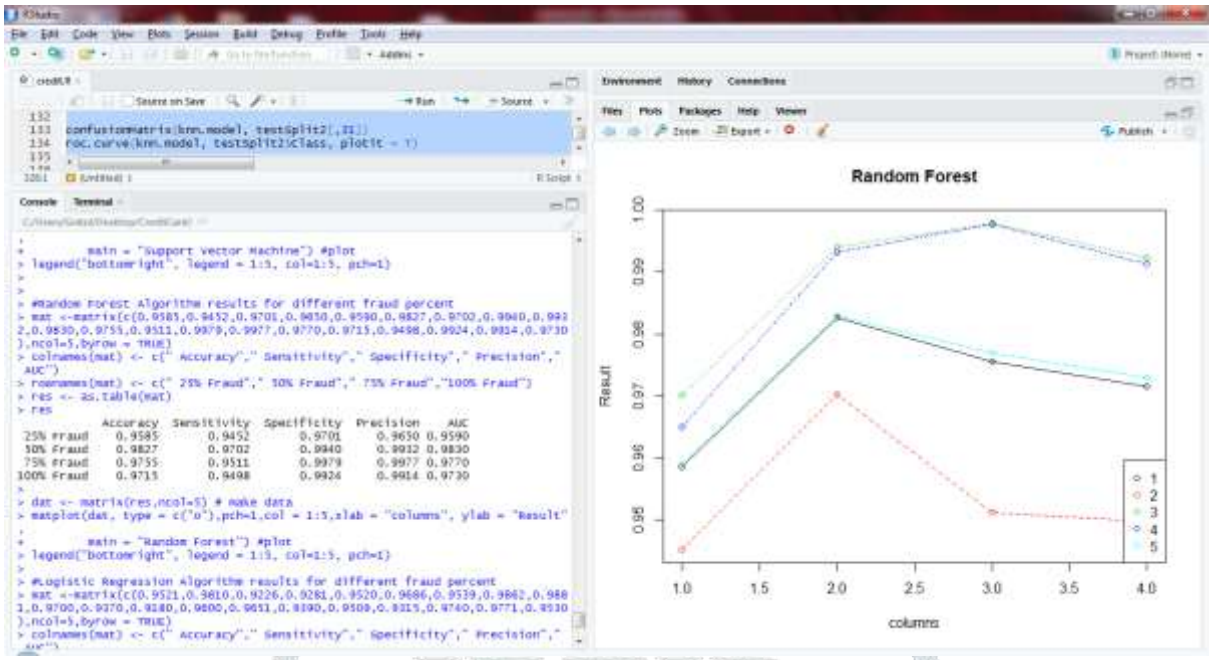

Fig.3: Representation of Random Forest

Logistic Regression Algorithm:

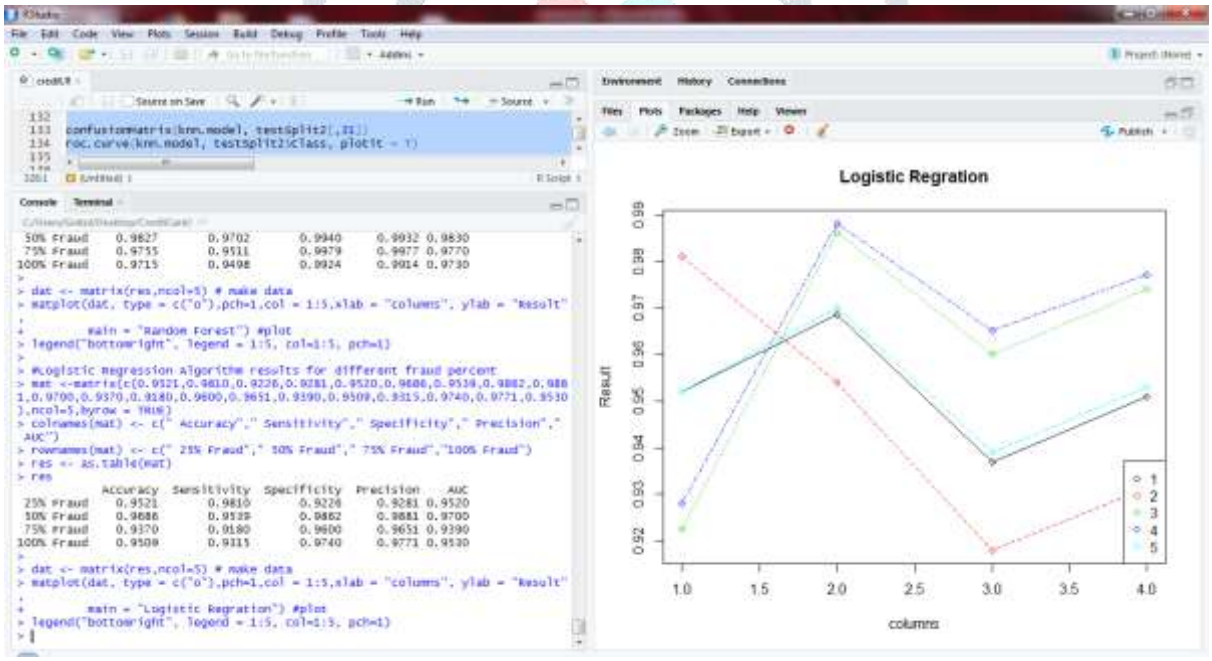

Fig 4: Representation of Logistic Regression

VI . CONCLUSION

The objective of this work is to detect the presence of fraud in credit card database with the help of four classification algorithms and then compared the result of all the four algorithms. In this work first used Logistic Regression and got the accuracy of 95.09%. Which is very good result. Then we used Support Vector Machine for the classification of fraud. It gave the accuracy of 96.12%. This is also good. Then used Random Forest got a very good accuracy of 97.15%.At last we used KNN and got the accuracy of 68.62%.So the results of comparisons specified that Random Forest give better results for the classification of fraud, for large transactions. Therefore ,an efficient credit card fraud detector is Random Forest Classifier among the four other classification algorithms. The next proposed work, in progress is to enhance the random forest algorithm to detect the fraud.

VII. ACKNOWLEDGEMENT

I would like to show my deepest gratitude to my guide for giving support and encouragement in doing my work. I would also like to expand my gratitude to all those who directly and indirectly guided me in completing my work. Finally, I would thank the publisher to help me press my paper.

VIII. REFERENCES

[1] Srivastava, Aman, et al."Credit card fraud detection at merchant side using neural networks.' Computing for sustainable Global Development (INDIACom),2016 3rd International Conference on.IEEE,2016.
[2] Mishra,Mukesh Kumar, and Rajashree Dash."A Comparative Study of Chebyshev Functional Link Artificial Neural Network, Multi-layer Perceptron and Decision Tree for Credit Card Fraud Detection"Information Technology (ICIT),2014 International conference on,IEEE,2014.
[3] Khan, Azeem Ush Shan,Nadeem Akhtar ,and Mohammad Naved Qureshi."Real-time credit card fraud detection using Artificial Neural Network tuned by simulated annealing algorithm" Proceedings of International Conference on Recent Trends in Information,Telecommunications and computing ,ITC,2014
[4] Duman, Ekrem, Ayse Buyukkaya and Ilker Elikucuk."A novel and successful credit card fraud detection system implemented in a turkish bank." Data mining workshops,2013 IEEE 13th International Conference ,IEEE,2013.
[5] P. Pahwa, M. Papreja, and R. Miglani, "Performance Analysis of Classification Algorithms," vol. 3, no. 4, pp. 50–58, 2014.
[6] I. Charalampopoulos and I. Anagnostopoulos, "A comparable study employing weka clustering/classification algorithms for web page classification," *Proc. - 2011 Panhellenic Conf. Informatics, PCI 2011*, pp. 235–239, 2011.
[7] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," *Int. J. Comput. Sci. Manag. Res.*, vol. 1, no. 4, pp. 686–690, 2012.
[8] I. M. M. Mitkees, A. Ibrahim, and B. Elseddawy, "Customer Churn Prediction Model using Data Mining techniques," pp. 262–268, 2017.
[9] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the cognitive skill of students in education environment," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res.*, pp. 1–4, 2014.
[10] M. N. Amin and A. Habib, "Comparison of Different Classification Techniques Using WEKA for Hematological Data," *Am. J. Eng. Res.*, no. 43, pp. 2320–847, 2015.
[11] I. Technology, "International Journal of Research in Computer & Information Technology (IJRCIT) Vol . 1 , Special Issue 1 , 2016 ISSN : 2455-3743 „ A SYSTEMATIC OVERVIEW ON DATA MINING : CONCEPTS AND TECHNIQUES " International Journal of Research in Computer & Informat," vol. 1, no. 1, pp. 136–139, 2016.
[12] . Rosset, S., Murad, U., Neumann, E., Idan, Y., Pinkas, G.: Discovery of fraud rules for telecommunications challenges and solutions. In: Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 409–413. NY, USA (1999).
[13].Taniguchi,M.,Haft,M.,Hollmén, J., Tresp,V.: Fraud detection in communication
networks using neural and probabilistic methods. In: Proc. of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, WA, USA, pp. 1241–1244 (1998).
[14]. Cox, K.C., Eick, S.G., Wills, G.J., Brachman, R.J.: Brief application description; visual data mining: recognizing telephone calling fraud. Data Mining and Knowledge Discovery **1**(2), 225–231 (1997)
[15] . Taniguchi, M., Haft, M., Hollmén, J., Tresp, V.: Fraud detection in communication networks using neural and probabilistic methods. In: Proc. of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, WA, USA, pp. 1241–1244 (1998)
[16]. Phua , C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research. Arxiv preprint (2010). Accessed 5 Jan 2012
[17]. Kovach, S., Ruggiero, W.V.: Online banking fraud detection based on local and global behavior.In: Proc. of the Fifth International Conference on Digital Society, Guadeloupe, France, pp. 166–171 (2011)
[18] Sahin.Y, Bulkan.S and Duman.E, "A cost-sensitive decision tree approach for fraud detection", Science Direct, Expert System with Applications 40 pp-5916-5923, 2013.
[19] .Li.J, Wei.W, Yuming.O and Chen.J, "Effective detection of sophisticated online banking fraud on extremely imbalanced data", Springer World Wide Web, pp 449-475, 2012.
[20]. Ankur Rohilla "Comparative analysis of various classification algorithms in the case of fraud detection." www.ijert.org, vol.6 Issue 09,September -2017.
[21].Srivastava,Aman, et al."Credit card fraud detection at merchant side using neural networks."Computing for sustainable Global Development(INDIACom),2016 3rd International Conference on.IEEE,2016.