ISSN 2278-3091
Volume 10, No.1, January - February 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse631012021.pdf
https://doi.org/10.30534/ijatcse/2021/631012021

Credit Card Fraud Detection Using Naive Bayes and Robust Scaling Techniques

Devendra D. Borse[1], Prof. Dr. Suhas. H. Patil[2], Dr. Sunita Dhotre[3]
[1]M.Tech Student, Department of Computer Engineering,
BharatiVidyapeeth (Deemed To Be University)
College of Engineering Pune, (India)
[2]Professor, Department of Computer Engineering,
BharatiVidyapeeth (Deemed To Be University)
College of Engineering Pune, (India)
[3]Associate Professor, Department of Computer Engineering,
BharatiVidyapeeth (Deemed To Be University)
College of Engineering Pune, (India)

ABSTRACT

The Internet is an important element of our life. Due to the wide use of the internet, the status of online shopping is increasing day by day. The Credit Card is the easiest method for online shopping and paying bills. Therefore, Credit Card becomes popular and appropriate approach for online money transaction and it is growing very quickly. In this paper, machine learning algorithms are utilized for the detection of credit card fraud. Firstly, common type of models is used. After that, hybrid methods which can use to Ada Boost and majority voting methods are activated. Ada Boost method is able to develop the individual results from different algorithms. For finding efficiency so we are used Kaggle dataset. Using we are calculated the result.In this paper, to group the main factors that can manual for unrivaled precision in Visa false exchange location procedure. Moreover, we clarify the presentation of various directed AI calculations that are existed in writing against the great classifier that it executed in this paper. The end-product of this framework have emphatically distinguished that most of casting a ballot technique acquires better quality, precision proportions in getting extortion cases in Mastercard for recognizable proof of genuine Mastercard exchange information.

Key words: Credit Card**,** Fraud detection, supervised machine learning, Naive Bayes, RobustScaler, online shopping, predictive modeling etc.

1. INTRODUCTION

Fraud is an illegal fraud estimated to bring individual gain. Credit card fraud is difficulty by offender use of credit card for property. Credit card fraud is Credit card transactions could be physically otherwise digitally [1]. The real exchange credit card is includes by the exchanges of credit card. In computerized exchanges, this can show up in overabundance of the phone or the web. Credit card misrepresentation is expanding obviously with the improvement of present day information and turned into a basic objective for coerces. Visa extortion has especially pointless freely available datasets. In this paper, the twelve AI determining are making use of for differentiate credit card misstatement. The calculations of positioning from usual neural management to profound studying models. These are sequence making use of both specification and authentic world Visa datasets. What's more, the Ada Boost and greater part casting ballot techniques are tried to create crossover models. To another ascertain the consistency of the models; commotion is joined with a basic informational index. The relation of this paper is the rating of a variety of AI models with a simple credit card dataset for misrepresentation identification [1][2].

1.1 Background

For cheats, the credit card is a basic and recognizable objective in light of the fact that with no danger of fundamental measure of cash is accomplished in a brief period [6]. To accomplish the Visa misrepresentation, fraudsters endeavor to take responsive data like as credit card number, ledger and government backed retirement number [4]. Fraudsters attempt to construct each deceitful exchange substantial which makes misrepresentation location a requesting issue. Expanded Visa exchanges exhibit that for the most part 70% individuals in the US can fall into the hold of these fraudsters [2]. In this paper, all machines learning calculation are ordered utilizing a basic credit card exchange to recognize extortion or non-misrepresentation exchange. The significant reason for this framework to be a legitimate regulated learning technique for the straightforward dataset.

1.2 Motivation

The Credit card is estimated while a "great objective of misrepresentation" in light of the fact that in an exceptionally humble aggressors can acquire heaps of cash with no chance and commonly the extortion is perceived following a couple of days. To accomplish the credit card misrepresentation either disconnected or on the web, fraudsters are looking for responsive information like as Visa number, financial balance, and government managed retirement numbers. In disconnected installment cases to execute the fake exchanges and in online installment an aggressor needs to get the credit card itself, the fraudsters should take costumer's character. On compelling card-based buy, simply card subtleties are determined during on the web or via telephone to produce the installment. In this strategy, the programmer essentially needs to distinguish the card subtleties to allocate misrepresentation [5]. Credit card misrepresentation is a significant issue and has a sensible expense for banks or card guarantor organizations. Thusly, with this impressive issue in an exchange framework, banks get credit card extortion basically and have incredibly muddled security frameworks to notice exchanges and recognize the cheats as quickly as conceivable whenever it is committed. The point of this framework is to accomplish a total examination of various extortion location strategies and picks some creative strategy for conversation [6].

1.3 Objectives and Scope of this Paper

1. In this paper, till analyze the nearly all main variables we can adviser to better accuracy in credit card fake transaction detection technique.
2. For calculate the effecting of unlike supervised machine study algorithms that are presented in literature int opposition to the fine classifier that it performs in this paper.
3. The recognize majority of voting method get good quality, accuracy ratios in grab fraud cases in credit cards for recognize of real credit card transaction data.

2. LITERATURE SURVEY

2.1 Research Gap

As frequency of transactions is increasing, number of fraudulent transactions are also increasing rapidly. In order to reduce fraudulent transactions, machine learning algorithms like Naive Bayes[17], Logistic regression, J48 and AdaBoost etc. are discussed in this paper. The same set of algorithms are implemented and tested using an online dataset [2]. Through comparative analysis it can be concluded that Logistic regression and AdaBoost algorithms perform better in fraud detection [3].

2.2 Review of Literature

In this paper, Authors [1] Kuldeep Randhawa1, Chu Kiong Loo1, ManjeevanSeera, CheePeng Lim, Asoke K. Nandi proposed a productive strategy that is an AI calculations are utilized for identification of credit card misrepresentation. The half breed techniques used AdaBoost and big size selection ballot strategies are attempted. AdaBoost technique can improve the individual outcomes from various calculations. Information under-inspecting was utilized to notice the exhibitions of the calculation, with RF addressing an unrivaled presentation as related with SVM and LOR.They are utilized in blend with the AdaBoost and greater part casting ballot techniques. In this paper, the eventual outcome proposed that the lion's share casting a ballot method has created the best MCC score of 0.942 for 30% clamor consolidated to the dataset. This review sees that the larger part casting a ballot method is secure in execution in the presence of commotion [1].

This author SahilDhankhad, Emad A. Mohammed, Behrouz Far [2] has proposed a technique for charge card extortion is rising broadly among the improvement of present day information and turned into a straightforward item for fakes. Visa misrepresentation is amazingly testing uninhibitedly available datasets. In this paper, we influence a few managed AI calculations for the identification of charge card fake exchanges utilizing dataset. Also, we utilize those calculations to execute a classifier with outfit learning strategies. The neural organization configuration utilized upon an unaided technique utilizing basic exchange access [2].Self-arranging figure of the neural organization by utilizing obvious order it can resolve the issue for each associated with a gathering.

Author [3]presented a powerful strategy by creating charge card exchange installment framework. Additionally one has been improved framework and breaks down Visa extortion that 70% of U.S. clients are to a great extent upset by character misrepresentation. This study considered two strategies for information mining one is SVM and another arbitrary woods. Additionally altogether chipped away at the notable calculated relapse to recognize credit card misrepresentation being essential for an assault. This strategy utilized worldwide charge card measure data set. SVM confronted significant two difficulties of misrepresentation recognition. At first, it is shaky class sizes of legitimate and phony exchanges mark substantial exchanges for over tallying deceitful ones. This study likewise make that a basic procedure like as arbitrary over and under inspecting is usually performed better, and regarded great generally the exhibition of irregular under testing [3].

In this overview, the creators [4] M. Hegazy, A. Madian and M. Ragaie proposed that, to develop a bound together example for each client show ordinary conduct, yet additionally Fraud design that is addressed already and affirmed as extortion exchanges that is rearrange considering fraudsters conduct. The main calculation suggested that, an Apriori calculation utilized in Fraud Miner for regularly Pattern creation and encourage sum up client prior conduct either inside his Legal or Fraud exchanges. Misrepresentation/Legal Pattern creation encourages quick of extortion discovery measure and could be utilized to affirm exchange close to ongoing exchanges. Because of the idea of the tremendous measure of exchanges that should be physically broke down which are restricted that adversely impacts the choice precision, Data digging strategies have produced unexpectedly as the best extortion identification strategy in this overview. This study proposes a Visa extortion discovery model that is dealt with extraordinary dataset and advances information on clients' examples by parting information into lawful and misrepresentation designs.

The Author [5] RanjeetaJha, Abhaya and Vijay Kumar Jhahas introduced the best strategy for Visa misrepresentation is the extortion devoted using someone else's Visa. To keep up safe Visa control a skilled misrepresentation recognition framework is essential. As of now, a few current strategies, basically relies upon Artificial Intelligence, Sequence Alignment, Data Mining, Fuzzy Logic, Machine Learning, Genetic Programming ,and so forth have been gotten for recognizing distinctive charge card fake exchanges. This overview proposes a current strategy utilized in the misrepresentation discovery structure just as delivering a total survey of various methodologies rely upon certain guideline of plan. In this paper, the creator proposes a Neural Network framework dependent on data set digging technique utilized for charge card extortion recognition. This framework has a joined to a scope of monetary information bases alongside a graphical UI..

Author [6] M. Zareapoor and P. Shamsolmoalihave proposed an arrangement of Visa misrepresentation is a significant issue and has an obvious cost for banks and card backer organizations. Accordingly, with this colossal trouble in exchange framework, banks acquire credit card extortion truly, and have extremely muddled security frameworks check exchanges and distinguish the cheats as fast as conceivable one time it is devoted. The point of this overview is to accomplish a general survey of various misrepresentation location techniques and chooses a few inventive strategies [6].

In this paper, Author [7] A. O. Adewumi and A. A. Akinyelu presented the effective strategy of an investigation of improved Visa misrepresentation identification systems. Especially, this review concentrated on current Machine Learning found and Nature Inspired based Visa extortion recognition techniques introduced in the article. This review presents a portrayal of latest things in Visa misrepresentation acknowledgment. Current misrepresentation identification frameworks utilized by shippers and banks are proposed to affirm exchanges through checking plans and execution. Here, two principle techniques used to deal with extortion contain: misrepresentation avoidance and misrepresentation recognition [7].

3. PROPOSED SYSTEM ARCHITECTURE

3.1 Data Set: - Data Set which are used for detection as fraud detection. Robust scaler is used for pre-processing after that we get scaled value for naive bayes algorithm.

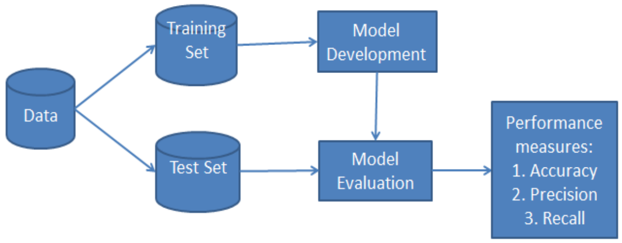


Figure 1: Proposed System Architecture

3.2 Bayes Theorem

Bayes Theorem its used to measure dependent chances. Making a strongest tool in the study of chances, it is used in Machine Learning.

The Formula Bayes Theorem is

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}} = \frac{\mathrm{P(A).P(B|A)}}{\mathrm{P(B)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where,

$\mathrm{P(A)}$ = The Probability of A occurring

$\mathrm{P(B)}$ = The Probability of B occurring

$\mathrm{P(A|B)}$ = The Probability of A given B

$\mathrm{P(B|A)}$ = The Probability of B given A

$\mathrm{P(A \cap B)}$ = The Probability of Both A and B occurringmostly used.

3.3 Naive Bayes Classifier

Naive BayesClassifiers are predicated from Bayes Theorem. There is one suspicion extract is the powerful freedom predications connecting the attribute. The classifiers take it that the value of a specific attribute is not dependent of the value of at all more feature. The supervised study place, Naive Bayes Classifiers are instructed highly skillfully. Naive Bayed classifiers require a little training data for guess the parameters demand for classification. Naive Bayes Classifiers has easy plan and execution and these are put into more actual life condition. Bayes theorem observe probability of the occurrence likely probability of determiner occurrence that have occurred.

3.4 Gaussian Naive Bayes

Once connect with sequence data we are supposition repeatedly get is that the continuous values related with each class are allocate as per to a normal (or Gaussian) distribution. The possibility of the features is supposed to be-

$$\mathrm{P}(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)2}{2\sigma_y^2}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

3.5 Variance

- This is individualistic of Y (i.e., σi),
- Also, individualistic of Xi (i.e., σk),
- also, both (i.e., σ)

Gaussian Naive Bayes assist sequence valued attribute and models all as accepting to a Gaussian (normal) distribution. The process to making an easy model is to suppose the data is detail by a Gaussian distribution with none co-variance (independent dimensions) allying dimensions. The model can be suitable by easily detecting the mean and standard deviation the points inside every flag.

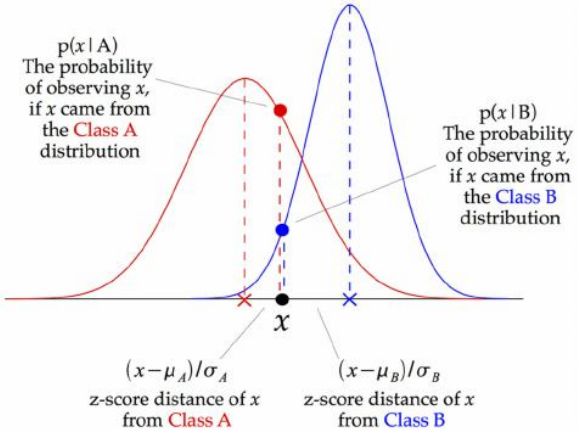

p(x | A)
The probability of observing x, if x came from the Class A distribution
p(x | B)
The probability of observing x, if x came from the Class B distribution
x
$(x-\mu_A)/\sigma_A$
z-score distance of x from Class A
$(x-\mu_B)/\sigma_B$
z-score distance of x from Class B

Figure 2: Gaussian (Normal) Distribution Model

The mention sample showing in what way Gaussian Naive Bayes (GNB) classifier process.The all data point, the z-score spanlinking that point and every class-mean is calculated, specifically the span from the class mean split by the standard deviation this class. Accordingly, we notice the Gaussian Naive Bayes have a somewhat unlike proposal and it can be utilized well.

3.6 Robust Scaler:

Scale highlights utilizing measurements that are powerful to exceptions. This Scalar eliminates the middle and scales the information as indicated by the quintile range (defaults to IQR: Interquartile Range). The IQR is the reach between the first quartile (25th quintile) and the third quartile (75th quintile). Focusing and scaling happen freely on each element by processing the significant measurements on the examples in the preparation set. Middle and interquartile range are then put away to be utilized on later information utilizing the change technique. Normalization of a dataset is a typical prerequisite for some, AI assessors. Normally, this is finished by eliminating the mean and scaling to unit change. Be that as it may, anomalies can regularly impact the example mean/difference in a negative manner. In such cases, the middle and the interquartile range regularly give better outcomes.

4. EXPERIMENTAL SET UP

We collected data from Kaggle [19]. The dataset has a credit card transaction in September 2013 by European cardholders. We imported libraries and then imported the required packages. we downloaded data from a CSV file using pandas. we checked the database. we have 31 different columns as v1 to v28 this occur of a PCA capacity limiting for guard user specification and delicate attribute in our dataset as we do not want to disclose personal ownership and location. Class 1 for fraud transactions, 0 for valid transaction. Amount is in transaction amount. We have 284807 transactions in 31 columns. The script is written in the python language with the IDE used for the PyCharm script. We got the results using the hardware system configuration by intel i5 core, 8 GB RAM, Windows 10 OS.

4.1 Parameters

Accuracy -One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have similar datasets spot values of false positive and false negatives these nearly equal have. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.9778 which means our model is approx. 98% accurate.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 99.79% precision which is pretty good.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the proportion is exactly predicted positive monitoring the all observations in actual class - yes. The question recall answers are: Of all the passengers that truly survived, how many did we label? We have got recall of 97.78%which is good for this model as it's above 0.5.

Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Accuracy works best if false positives and false negatives have nearby value. If the value of false positives and false negatives are very unalike, it's good to see together Precision and Recall. In our case, F1 score is 98.72%.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

AUC Score - AUC stands for Area Under the Curve. ROC can be quantified using AUC. The way it is done is to see how much area has been covered by the ROC curve. In our case, AUC score is 95.74%

4.2 Distribution of Transaction Amount and Time

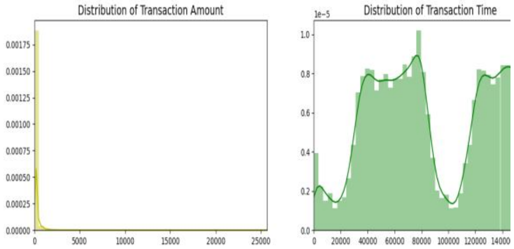


Figure 3: Distribution of Transaction Amount and Time

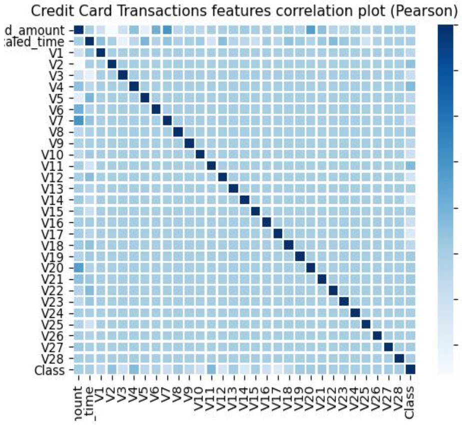


Figure 4: Credit Card Transactions features correlation plot (Pearson)

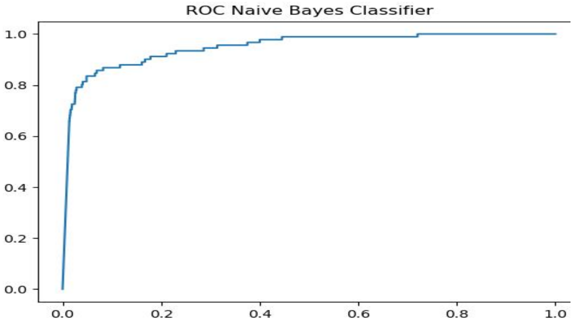
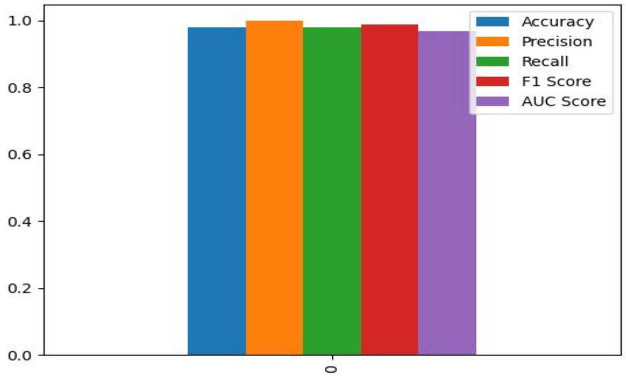


Figure 5: ROC Naive Bayes Classifier.



4.3 Comparision between cureent study vs prevoius study

Table1: The Results from Our Model

Model	Accuracy	Precision	Recall	F1 Score	AUC Score
Naive Bayes using Robust Scaler	0.97779	0.99789	0.97779	0.98711	0.95733

Table 2: The Results from Previous Model

Model	Accuracy	Precision	Recall	F1 Score	AUC Score
Naive Bayes	0.90540	0.91	0.91	0.91	0.85714

5.CONCLUSION

In recent years, fraudulent transactions have become widespread and have become a major problem for banks around the world. In this paper, the Naive Bayes classification of machine learning is used to predict common or fraudulent transactions. The Naive Bayes classifier is based on their accuracy, recall, precision,F1 score, AUC score. The previous Naive Bayes got the Accuracy 90.54%, Precision 91%, Recall 91%, F1 score 91%, AUC score 85.71% compare with our Naive Bayes using robust scaling got the Accuracy 97.78%, Precision 99.79%, Recall 97.78%, F1 score 98.71%, AUC score 95.73%.The result of Naive Bayes models is superior in overall performance. Overall results show Naive Bayes classifier which is used Robustscaleris most promising for predicting fraud transaction in the dataset.

REFERENCES

[1] Y. Sahin, S. Bulkan, and E. Duman, "**A cost-sensitive decision tree approach for fraud detection," Expert Systems with Applications**, *vol. 40, no. 15, pp. 5916–5923, 2013.*

[2] A. O. Adewumi and A. A. Akinyelu, "**A survey of machine-learning and nature-inspired based credit card fraud detection techniques**," *International Journal of System Assurance Engineering and Management, vol. 8, pp. 937–953, 2017.*

[3] A. Srivastava, A. Kundu, S. Sural, A. Majumdar, "**Credit card fraud detection using hidden Markov model**,*" IEEE Transactions onDependable and Secure Computing, vol. 5, no. 1, pp. 37–48, 2008.*

[4] J. T. Quah, and M. Sriganesh, "**Real-time credit card fraud detection using computational intelligence," Expert Systems with Applications**, *vol. 35, no. 4, pp. 1721–1732, 2008.*

[5] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C., "**Data mining for credit card fraud: A comparative study," Decision Support Systems**, *vol. 50, no. 3, pp. 602–613, 2011.*

[6] N. S. Halvaiee and M. K. Akbari, "**A novel model for credit card fraud detection using Artificial Immune Systems," Applied Soft Computing**, *vol. 24, pp. 40–49, 2014.*

[7] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, "**Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning**,*" Information Fusion, vol. 10, no. 4, pp. 354–363, 2009.*

[8] N. Mahmoudi and E. Duman, "**Detecting credit card fraud by modified Fisher discriminant analysis**," *Expert Systems with Applications, vol. 42, no. 5, pp. 2510–2516, 2015.*

[9] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, "**Association rules applied to credit card fraud detection," Expert Systems with Applications**, *vol. 36, no. 2, pp. 3630–3640, 2009.*

[10] SarweenZaza and Mostafa Al-Emran, "**Mining and Exploration of Credit Cards Data in UAE**", *Fifth International Conference on e-Learning, pp 275-79, 2015*

[11] Krishna KeerthiChennam and Lakshmi Mudanna, "**Privacy and Access Control for Security of Credit Card Records in the Cloud using Partial Shuffling**", *IEEE International Conference on Computational Intelligence and Computing Research, 2016*

[12] Rajeshwari U and Dr B SathishBabu, "**Real-time credit card fraud detection using Streaming Analytics**", *2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp 439 – 444, 2016*

[13] John O. Awoyemi, Adebayo O. Adetunmbi and Samuel A. Oluwadare, "**Credit card fraud detection using Machine Learning Techniques**: **A Comparative Analysis",***IEEE, 2017*

[14] Mukesh Kumar Mishra and Rajashree Dash, "**A Comparative Study of Chebyshev Functional Link Artificial Neural Network, Multi-Layer Perceptron and Decision Tree for Credit Card Fraud Detection**", *International Conference on Information Technology, pp 228 -233, 2014*

[15] Pornwatthana Wongchinsri and Werasak Kuratach, "**A Survey - Data Mining Frameworks in Credit Card Processing**", *IEEE, 2016*

[16] Yufeng Kou, .et. al., "**Survey of Fraud Detection Techniques**", *International Conference on Networking, Sensing & Control, pp 749 – 754, 2004*

[17] Kaviani P, Dhotre S. **Short survey on naive bayes algorithm**. *International Journal of Advance Engineering and Research Development. 2017 Nov;4(11):607-11.*

[18] Dandavate PP, Dhotre SS. **Data Leakage Detection using Image and Audio Files**. *International Journal of Computer Applications. 2015 Jan 1;115(8).*

[20] Devendra D. Borse, Prof. Dr. Suhas. H. Patil "**Credit Card Fraud Detection Using Machine Learning**".*International Journal of Innovative Research in Computer and Communication Engineering, Vol. 7, Issue 5, May 2019.*

[21]F. D. Mulla and N. Jayakumar, "**A Review of Data Mining & Machine Learning approaches for identifying Risk Factor contributing to likelihood of Cardiovascular Diseases**," *2018 3rd International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2018, pp. 631-635.*