Credit Card Fraud Detection using Machine Learning

[1]Aditya Kumar Pandey, [2]Aman Srivastava, [3]Aman Kumar Singh, [4]Abhishek Kumar Singh,[5] Dr. Sasidhar Babu Suvanam

School of Computer Science and Engineering, REVA University, Karnataka

[1]R18CS021@cit.reva.edu.in, [2]R18CS033@cit.reva.edu.in,[3] R18CS032@cit.reva.edu.in,
[4]R18CS013@cit.reva.edu.in, [5]sasidharbabu.suvanam@reva.edu.in

Abstract.

Mastercard fraud is growing at a rapid pace, owing to the advancement of current technology and the global interstates of correspondence. Each year, Mastercard fraud costs buyers and the financial institution billions of dollars, and cyber attackers are constantly on the lookout for new guidelines and strategies to commit illegal activities. As a result, extortion identification frameworks have become critical for banks and the monetary establishment in order to limit their losses. Despite this, there is a dearth of distributed writing on Visa's misrepresentation identification procedures, owing to analysts' inability to access the credit card exchanges dataset. Classification methods can be constructed using a variety of Algorithms. In this article, we will evaluate several algorithms and then construct the model with the best performing calculation.

Keywords. Machine Learning, Credit Card, classification methods

1. INTRODUCTION

A customer's account is deemed to be in default if he or she fails to repay the money after a particular notification period specified by the bank. This money is later retrieved from clients with the assistance of collection agencies. Too many defaults will almost certainly cause financial institutions to lose money. It'll be expected to evaluate the probability of a record defaulting and to terminate its processes.

The majority of defaulted accounts occur when financial institutions issue credit cards to customers in order to improve their market value. Certain customers may lack financial literacy and wind-up spending excessively or purposefully do not plan to repay. Financial institutions may be unable to collect funds from some consumers who are unable to repay or who have incorrect credentials. Financial institutions suffer significant financial losses because of these circumstances. Thus, it is critical to assess the trend of delinquent accounts and take appropriate steps to prevent more fraudulent activities. This will assist in limiting future misfortunes and the number of defaulted records. Machine learning methodologies help in developing systems that significantly improve the likelihood of detecting false representation. A model is proposed and compared to previous models for detecting default payment credit card frauds.

2. LITERATURE SURVEY

The authors have tried the approach of Majority Voting and Adaboosting upon multiple classification models to get a precise prediction of result [2]. In the benchmark dataset the adaboost method gave accuracy for fraud detection up to 82.317% using SVM and gave lowest 42.683% for Random Forrest whereas majority voting gave accuracy of 78.862% for combination of Naïve Bayes and Neural Network algorithm and gave lowest 23.780% for combination of Random Forrest and Gradient Boost. After selecting only relevant features and applying the models on a real time dataset they got 96.078% accuracy with Deep Learning and 98.039% with the combination of Naïve Bayes and Gradient Boost.

The authors executed multiple deep learning models for fraud recognition and assessed their precision [3]. The models implemented were Artificial NN, Recurrent NN, LSTM, Gated Recurrent Unit. They implemented feature engineering to derive important features and used undersampling to balance the uneven dataset. The GRU gave the highest accuracy of 91.6%, LSTM gave accuracy of 91.2%, RNN gave accuracy of 90.433% and ANN gave the lowest accuracy of 88.9%.

The author has focused in on the tree-based ensemble learning [9] to counter the insecurity of conventional Decision Tree. Dataset resembles what has been utilized in this paper. To compare the performances, both weighted and unweighted methods were utilised. Random Forest possesses the most precise testing data. with an accuracy of 82.12%. Adaboost achieved 68.9%, while Logistic Regression achieved 64.64%.

3. ALGORITHMS FOR DEFAULT DETECTION

Allowing frameworks to learn on their own is the goal of Machine Learning. As a result, rather than being tailored to perform a specific action, the framework learns on its own, makes do, and adjusts. Software engineers who are able to access and modify the given information in accordance with the needs of the client are the real focus of this field. There are three types of artificial intelligence: those that are directed, those that are solo, and those that support learning. A type of Machine learning known as "directed learning" is the most widely accepted one. The purpose of this paper is to conduct experiments and assess the results of various Machine learning models.

A. Logistic Regression

Essentially, it is a factual model that utilizes a strategic capacity to envision a paired ward variable. It is frequently used in situations where the possibility of a binary classification exists. It performs well on classes that are easily isolated.

B. K-Nearest Neighbor

It is a characterization and regression technique that is used in a wide assortment of uses. It is based on the resemblance of features. It is alluded to as a non-parametric procedure as it generates no inferences concerning data being used.The calculation determines a point's k-closest neighbors and assigns the point the name with the greatest number of k neighbors. This is a straightforward calculation, which makes it easier to comprehend. KNN is resource intensive due to the way it should register the k-closest neighbors of all preparation data, as well as the fact that all preparation data should be saved as well as calculated.

C. Decision Tree

It is a parallel tree that maps elective outcomes to a collection of issue-related inquiries. The tree begins at the root and gradually grows in quantity as it develops. The leaf nodes replicate the marks, whereas the within nodes address a few critical inquiries and branch left or right as necessary. This procedure is repeated till we reach a leaf node. These inquiries are created utilizing characteristics such as the Gini index. The index causes the information to be as lopsided as possible, creating ambiguity about where the branch should go. While it is a straightforward method that works with both class and mathematical data, it proves to be more perplexing when dealing with highly vulnerable data.

D. Naïve Bayes

It is a predictive classification method, which means that this may generate prediction for multiple classes concurrently. The Naive Bayes corroborates this. Probability - based Classification methods are those that enable the prediction of multiple class characterizations. Conditional probability is used to make the determination. Instead of a single algorithm, this paradigm employs a collection of algorithms that share a common idea. Each characteristic is assumed to contribute equally and independently to the outcome in this model. This model has a number of benefits above other models, including the fact that it requires very little preparation information.

E. Support Vector Machine

It is a type of Controlled Learning strategy used to address Regression and Classification problems. In any case, it is frequently used in Computational challenges involving classification. The SVM calculation's objective is to locate the optimal line or choice limit that effectively classifies n-layered space, allowing us to order new pieces of information later. A hyperplane is a term that refers to this ideal choice limit. The hyperplane's outlandish focuses/vectors are chosen by SVM.

F. Random Forest

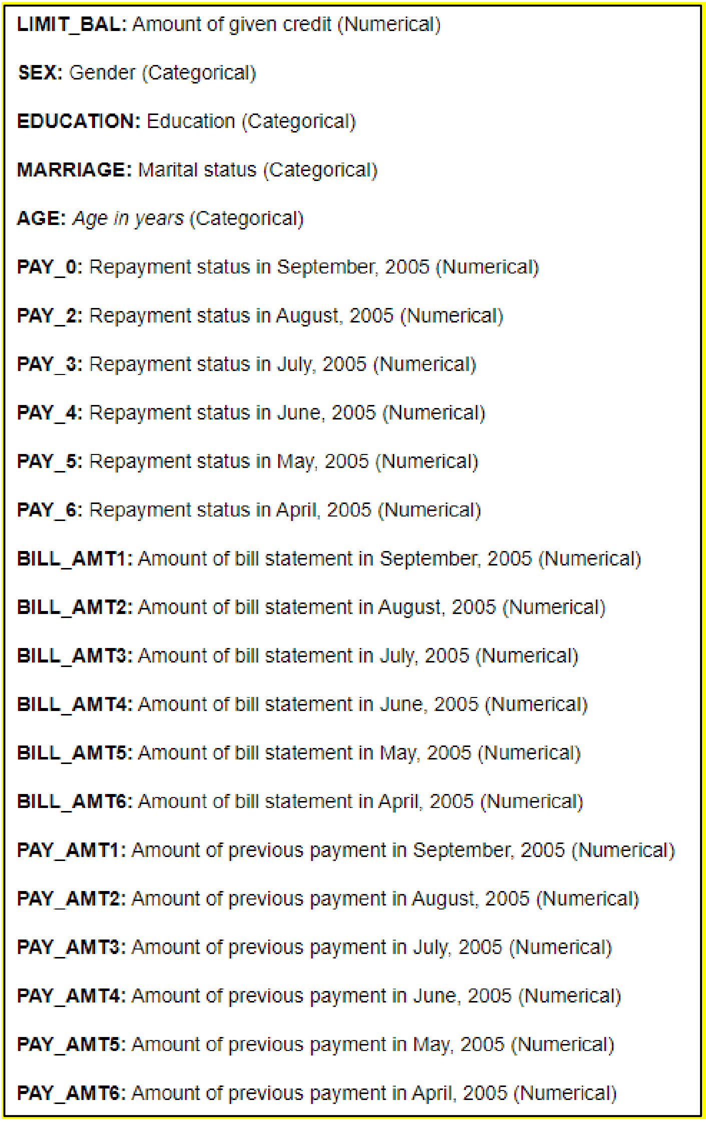
It is a classification technique based on ensemble learning. Random Forest is a flexible, easy-to-use AI calculation that consistently produces fair results, even in the absence of hyper-boundary changes. Additionally, it is a prominent approach due to its simplicity and suitability for regression and classification applications. Random Decision Forests are used to address the issue of overfitting in the training set in decision trees. Forests develop naturally without the need for excessive preparation, as forest splits along angled hyper planes, resulting in increased precision.

G. XGBoost

XGBoost is a choice tree-based gradient boosting technique. Choice tree-based calculations are great for small to medium-sized organized/even data at the moment. This computation generates a series of choice trees. In XGBoost, loads are critical. All free variables are assigned weights and accounted for in the choice tree that predicts the results. The weight of incorrectly anticipated factors is increased and dispatched to the following choice tree. After that, individual classifiers/indicators are combined to form a more accurate model. This device can be used to address a predicting issue that is defined by the user.

4. IMPLEMENTATION

Business Understanding was required as the initial step in the construction of the paper, followed by data collection. This dataset was obtained from UCI Machine Learning (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+customers). It contains 300,000 records and 23 attributes. The depiction of attributes is shown in the image below.



LIMIT_BAL: Amount of given credit (Numerical)

SEX: Gender (Categorical)

EDUCATION: Education (Categorical)

MARRIAGE: Marital status (Categorical)

AGE: *Age in years* (Categorical)

PAY_0: Repayment status in September, 2005 (Numerical)

PAY_2: Repayment status in August, 2005 (Numerical)

PAY_3: Repayment status in July, 2005 (Numerical)

PAY_4: Repayment status in June, 2005 (Numerical)

PAY_5: Repayment status in May, 2005 (Numerical)

PAY_6: Repayment status in April, 2005 (Numerical)

BILL_AMT1: Amount of bill statement in September, 2005 (Numerical)

BILL_AMT2: Amount of bill statement in August, 2005 (Numerical)

BILL_AMT3: Amount of bill statement in July, 2005 (Numerical)

BILL_AMT4: Amount of bill statement in June, 2005 (Numerical)

BILL_AMT5: Amount of bill statement in May, 2005 (Numerical)

BILL_AMT6: Amount of bill statement in April, 2005 (Numerical)

PAY_AMT1: Amount of previous payment in September, 2005 (Numerical)

PAY_AMT2: Amount of previous payment in August, 2005 (Numerical)

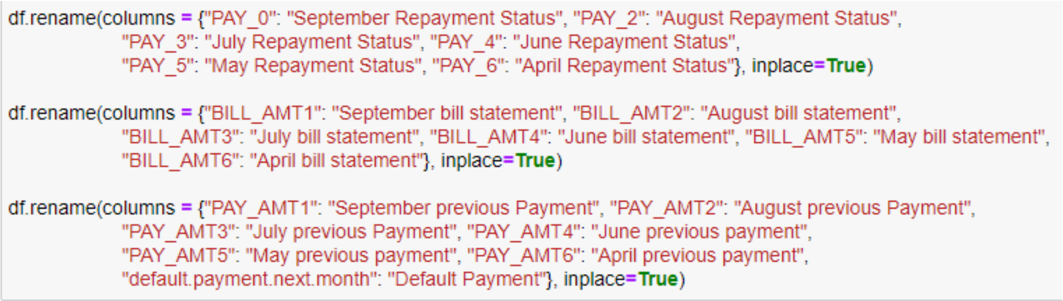
PAY_AMT3: Amount of previous payment in July, 2005 (Numerical)

PAY_AMT4: Amount of previous payment in June, 2005 (Numerical)

PAY_AMT5: Amount of previous payment in May, 2005 (Numerical)

PAY_AMT6: Amount of previous payment in April, 2005 (Numerical)

Next was the data preprocessing or data cleaning stage. It entails removing invalid attributes, if present, omitting irrelevant sections, examining the information type for features, etc. We named the features according to our own understanding.



```
df.rename(columns = {"PAY_0": "September Repayment Status", "PAY_2": "August Repayment Status",
           "PAY_3": "July Repayment Status", "PAY_4": "June Repayment Status",
           "PAY_5": "May Repayment Status", "PAY_6": "April Repayment Status"}, inplace=True)

df.rename(columns = {"BILL_AMT1": "September bill statement", "BILL_AMT2": "August bill statement",
           "BILL_AMT3": "July bill statement", "BILL_AMT4": "June bill statement", "BILL_AMT5": "May bill statement",
           "BILL_AMT6": "April bill statement"}, inplace=True)

df.rename(columns = {"PAY_AMT1": "September previous Payment", "PAY_AMT2": "August previous Payment",
           "PAY_AMT3": "July previous Payment", "PAY_AMT4": "June previous payment",
           "PAY_AMT5": "May previous payment", "PAY_AMT6": "April previous payment",
           "default.payment.next.month": "Default Payment"}, inplace=True)
```

Fig. 1: Image showing code for renaming of columns

There was an additional one class in section “0” of the marriage code, so it was eliminated. Education column also had one extra category “6”, this was also removed.

The dataset was represented as graphs to obtain additional information. This aided external comprehension of the information. Using heatmap, we additionally examined for correlated factors.

Labels		
Gender	**Defaulting status**	**Education**
1 – Male 2 – Female	0 – Non defaulted 1 – Defaulted	1 – Graduate school 2 – University 3 – High School 4 – Others 5 – Unknown

Table 1. Labels used for plotting the graph

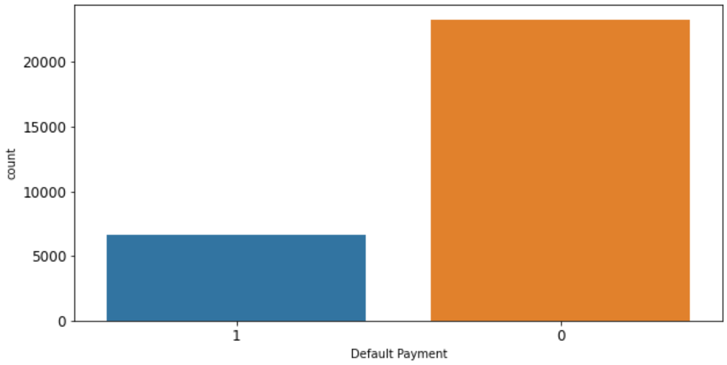


Fig. 1. Graph showing defaulted and non-defaulted accounts

From figure 1 we can infer that there are about 6500 accounts which have been defaulted.

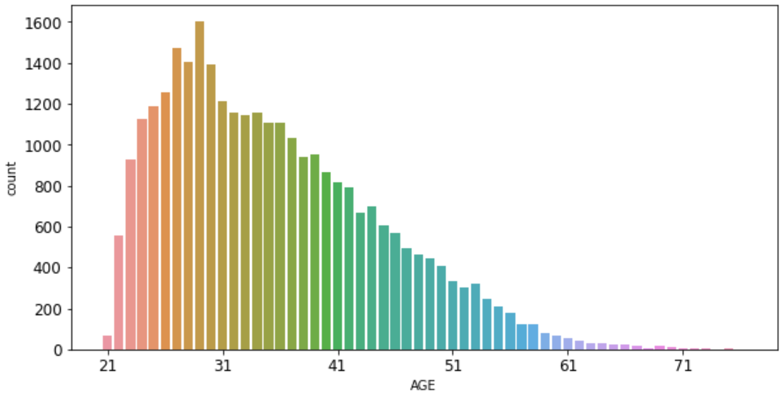


Fig. 2. Graph showing credit card users based on age

In figure 2 we can see that there is a large group of young population using credit cards. Most of the people who come under the influence of financial freedom are young people and it is easy to manipulate them to get any scheme. [26]

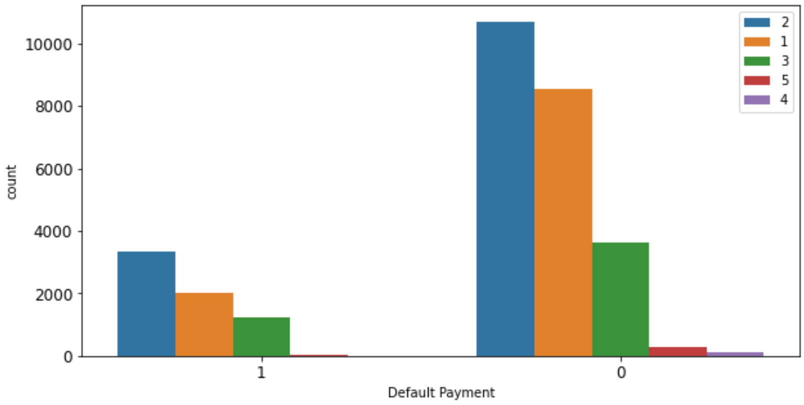


Fig 3. Graph showing defaulted and non-defaulted accounts based on education level

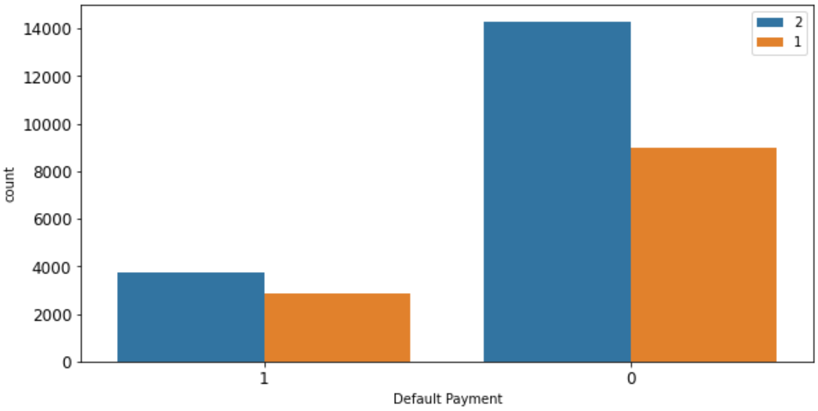


Fig. 3. Graph showing defaulted and non-defaulted accounts based on gender

Even though there are significantly more males than females using the credit card, we can see that gender does not matter much in the case of defaulted accounts.

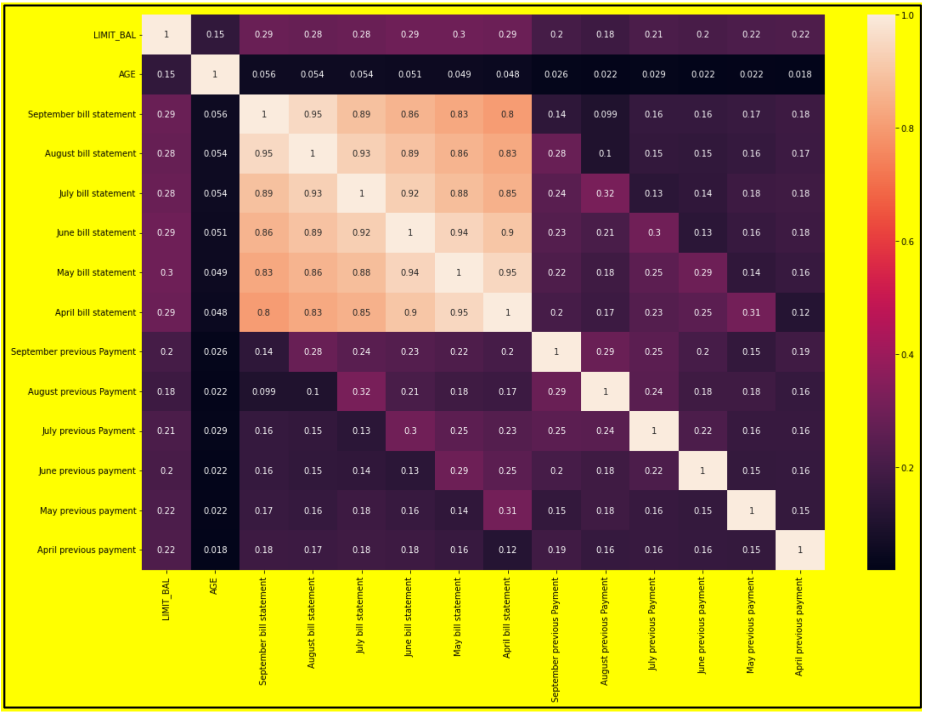


Fig. 4. Heatmap showing correlation between attributes

According to the illustrations, the majority of the segments were inclined. To counter this, the element section was scaled using a standard scaler. To modify the data, we utilized the SMOTE method. It was then divided into test and training sets, 20% and 80% respectively. We constructed models using the algorithmsmentioned earlier and reviewed the performance.

5. RESULTS

We one by one trained the model and evaluated the performance on the basis of accuracy. Different ML algorithms was used for determining which is giving the best result and Naïve bayes gave the best result.

Algorithm	Accuracy (%)	Precision (%)
Logistic Regression	80.84	79
K-Nearest Neighbor	75.55	75
Decision Tree	71.24	73
Naïve Bayes	80.87	79
SVC	80.84	79
Random Forest	72.54	73
XG Boost	80.14	78

Table 2. Table showing experimental results of each model used for prediction

From table 2 we can infer that LR, NB and SVC have similar accuracy and precision but, on the basis of numbers NB is the best performing model.

6. CONCLUSION

We utilized our data to assess the usefulness of various AI models for predicting its probability of a payment defaulting. To arrive at a specific result, we used precision as the deciding factor. We looked at Logistic Regression, KNN, Decision Tree, Naive Bayes, SVM, Ensemble technique, and XGBoost in this review. We concluded further that Naive Bayes model is the most appropriate model for forecasting the likelihood of a payment default.

Future research works may perform resampling on specific datasets. This contributes to decreasing the lopsidedness proportion of datasets, resulting in improved classification results. A model with a reasonable degree of precision and accuracy can be delivered as a web application or integrated into frameworks for the purpose of monitoring transactions.

7. REFERENCES

[1]S. Khatri, A. Arora and A. P. Agrawal, "Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 680-683, doi: 10.1109/Confluence47617.2020.9057851.

[2]K. Randhawa, C. K. Loo, M. Seera, C. P. Lim and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," in IEEE Access, vol. 6, pp. 14277-14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[3]A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," 2018 Systems and Information Engineering Design Symposium (SIEDS), 2018, pp. 129-134, doi: 10.1109/SIEDS.2018.8374722.

[4]S. Dhankhad, E. Mohammed and B. Far, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018, pp. 122-125, doi: 10.1109/IRI.2018.00025.

[5]S. Mittal and S. Tyagi, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp. 320-324, doi: 10.1109/CONFLUENCE.2019.8776925.

[6]D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic and A. Anderla, "Credit Card Fraud Detection - Machine Learning methods," 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), 2019, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717766.

[7]S. S. H. Padmanabhuni, A. S. Kandukuri, D. Prusti and S. K. Rath, "Detecting Default Payment Fraud in Credit Cards," 2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT), 2019, pp. 15-153, doi: 10.1109/ICISGT44072.2019.00018.

[8]T. M. Alam et al., "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," in IEEE Access, vol. 8, pp. 201173-201198, 2020, doi: 10.1109/ACCESS.2020.3033784.

[9]Y. Yu, "The Application of Machine Learning Algorithms in Credit Card Default Prediction," 2020 International Conference on Computing and Data Science (CDS), 2020, pp. 212-218, doi: 10.1109/CDS49703.2020.00050.

[10]S. N. Kalid, K. -H. Ng, G. -K. Tong and K. -C. Khor, "A Multiple Classifiers System for Anomaly Detection in Credit Card Data With Unbalanced and Overlapped Classes," in IEEE Access, vol. 8, pp. 28210-28221, 2020, doi: 10.1109/ACCESS.2020.2972009.

[11]Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802.

[12]F. N. Khan, A. H. Khan and L. Israt, "Credit Card Fraud Prediction and Classification using Deep Neural Network and Ensemble Learning," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 114-119, doi: 10.1109/TENSYMP50017.2020.9231001.

[13]A. Bačová and F. Babič, "Predictive Analytics for Default of Credit Card Clients," 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2021, pp. 000329-000334, doi: 10.1109/SAMI50585.2021.9378671.

[14] Hridya V Devaraj, Anju Chandran, Dr. Suvanam Sasidhar Babu “MANET Protocols: Extended ECDSR Protocol for Solving Stale Route Problem and Overhearing” IEEE proceedings of the 2016 InternationalConference on Data Mining and Advanced Computing (SAPIENCE), 2016. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7684168 DOI: 10.1109/SAPIENCE.2016.7684168

[15] Divya. D, Dr. Suvanam Sasidhar Babu “Methods to detect different types of outliers” IEEE proceedings of the 2016, International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016 http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7684114 DOI: 10.1109/SAPIENCE.2016.7684114

[16] Sajay K.R, Dr. Suvanam Sasidhar Babu“A Study of Cloud Computing Environments for High Performance Applications”, IEEE proceedings of the 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016. http://ieeexplore.ieee.org/document/7684127/ DOI: 10.1109/SAPIENCE.2016.7684127, Electronic ISBN: 978-1-4673-8594-7 Print on Demand (PoD) ISBN: 978-1-4673-8595-4

[17] Teena Jose, Vijayalakshmi, Yellepeddi, Dr. Sasidhar Babu Suvanam and Dr. Mani Megalai“Cyber Crimes in India: A Study”, IEEE proceedings of the SCOPES 2016.Date Added to IEEE Xplore: 26 June 2017 DOI: 10.1109/SCOPES.2016.7955584 http://ieeexplore.ieee.org/document/7955584/ Pages: 960 - 965

[18] Advances in Computational Sciences and Technology, “Cyber Crimes in Kerala: A Study” May 2017.Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, 5 (2017) pp. 1153-1159http://www.ugc.ac.in/pdfnews/8919877_Journals-1.pdf/1279 http://www.ripublication.com/acst17/acstv10n5_45.pdf

[19] Advances in Computational Sciences and Technology, “Blue brain – A massive storage Space” May 2017. (http://www.ugc.ac.in/pdfnews/8919877_Journals-1.pdf/ 1279)

[20] Bindhia K.F, Yellepeddi Vijayalakshmi, Dr.P. Manimegalai& Suvanam Sasidhar Babu, Classification using Decision Tree Approach towards Information Retrieval

Keywords Techniques and A Data Mining Implementation using WEKA data set, International Journal of Pure and Applied Mathematics,ISSN: 13118080 (printed version), Volume 116 No. 22 2017, 19-29, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) http://acadpubl.eu/jsi/2017-116-13-22/issue22.html https://www.scopus.com/sourceid/19700182690

[21] Sreedhar, K. S., Ahmed, S. T., & Sreejesh, G. (2022, June). An Improved Technique to Identify Fake News on Social Media Network using Supervised Machine Learning Concepts. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 652-658). IEEE.

[22] Y. Vijayalakshmi, P. Manimegalai, GKD Prasanna Venkatesan, Dr. S. Sasidhar Babu "Contextual Information Retrievalin Digital library and Research Over Current Search Engines" Jour of Adv Research in Dynamical & Control Systems, Vol.11, 01- ISSN 1943-023X, 2019, pp 790-793.

[23] Sajay KR, Suvanam Sasidhar Babu, Vijayalakshmi Yellepeddi, Enhancing the Security of Cloud Data Using Hybrid Encryptionalgorithm, "Journal of Ambient Intelligence and Humanized Computing (SPRINGER)" 2019 https://doi.org/10.1007/s12652- 019-01403-1, July 2019, Impact Factor: 7.104 (2020)

[24] Bindhia K Francis, Suvanam Sasidhar Babu, predicting academic performance of students using a hybrid data miningapproach, "Journal of Medical Systems (SPRINGER)" 43:162, 30th April 2019, Impact Factor: 5.23 (2020)) – Q1 Rated Journal. (2020), https://doi.org/10.1007/s10916-019-1295-4

[25] Vijayalakshmi Y, Manimegalai, Suvanam Sasidhar Babu, Accurate Approach towards Efficiency of Searching Agents in Digital Libraries using Keywords" "Journal of Medical Systems (SPRINGER)" 43:164. https://doi.org/10.1007/s10916-019-1294-5,1st May 2019, Impact Factor: 5.23 (2020)

[26] Yellepeddi Vijayalakshmi, Neethu Natarajan, Dr.P. Manimegalai and Dr. Suvanam Sasidhar Babu, "Study on Emerging Trends In Malware Variants", for publication in IJPAM International Journal of Pure and Applied Mathematics (SCOPUS), ISSN 1314-3395.Volume 116 No. 22 2017pages 479-489, ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version) http://acadpubl.eu/jsi/2017-116-13-22/issue22.html https://www.scopus.com/sourceid/19700182690