CREDIT CARD DEFAULT PREDICTIVE MODEL USING MACHINE LEARNING CLASSIFICATION TECHNIQUE

Marilou O. Espina
Bukidnon State University
Malaybalay City, Philippines
marilouespina@buksu.edu.ph

ABSTRACT *- Credit cards have become a vital element in the banking industry. It gives a significant value to the banks. Managing the risks in credit cards becomes one of the crucial tasks of the banks. In this study, a credit card default model was created based on the cardholder's demographic variables, spending records, and debt repayment status over the last six months using Naive Bayes, Decision Table, DTNB, ADtree, And LADtree classifying algorithm. Based on the classifier's evaluation metrics the ADTree got the best results. The creation of the Credit card default model was based on the optimal classifier ADTree.*

Keywords - Credit card default, Naïve Bayes, Decision Table, DTNB, ADTree

I. INTRODUCTION

The credit card is one of the most popular financial instruments in modern society. It becomes a replacement for the use of cash and leads to reduced circulation of the currency and printing of money [1], [2]. And it also speeds up the mobility of cash flows. Cardholders, merchants, and Banks all benefit from the use of credit cards. Credit cards are considered an important product for the banking industry, and managing credit card risks is one of the banks' crucial tasks [3]. Credit card default is one of the risks; it means clients failed to pay a required minimum payment before the bill was due. Creating a model to predict the behavior of credit card clients who will eventually default is one of the important methods to reduce the risks. Knowing the determinants of credit card default will guide the financial institution in crafting measures and policies to reduce the occurrence of such. There are studies that created predictive model like in the study of [4] that utilized the weighted SVM algorithm and in the study of [5] using Recurrent Neural Network (RNN). However, it is good to compare the result of different classification algorithm. The main objective of this paper is to create a credit card default model based on cardholder's demographic variables, spending records, and debt repayment status over the last six months which totaled 23 predictors, cardholders were classified into two categories, clients who default (yes) and who do not default (no) using Naive Bayes, Decision Table, DTNB, ADtree And LADtree classifying algorithm and selecting the optimal classifier.

II. REVIEW OF RELATED LITERATURE

A. Classification Algorithms

NAIVE BAYES CLASSIFIER: Naive Bayes is based on Bayes' theorem, in which it is assumed that all attributes are independent given the value of the class variable [6, 7].

DECISION TABLE: A classification model employed for predictive analysis, featuring a tabulated structure representing each attribute and its possible range of values, alongside the corresponding predicted output based on attribute value combinations. The model may be visualized as a hierarchical table, with higher-level tables breaking down into lower-level tables based on the values of additional attribute pairs, resulting in enhanced prediction accuracy[8].

DTNB: A combination of decision table/naive Bayes classifier. The set of attributes is divided into two groups, the class probabilities assigned in one group are based on naive Bayes, and the other group's class probabilities are based on a decision table, and the combination of the resulting probability estimates is utilized [9], [10].

ADTree (Alternating decision tree): machine learning technique for classification that uses a boosting approach to generalize decision trees. It is designed with a series of alternating decision and prediction nodes, with the former indicating predicate conditions and the latter containing single-value data. In the classification process, ADTrees perform a complete traversal of all decision paths that evaluate as true, then sum any prediction nodes that were visited to generate a final classification result. This methodology allows for highly accurate and nuanced classifications that can be used in a variety of applications.[11]

LADTree (Logical Analysis of Data): The analysis focuses on a subset of variable combinations that are significantly associated with either positive or negative observations. To optimize for reliability and efficiency, the methodology extracts only essential models, constructed from a limited set of combinatorial patterns[12].

B. Related Works

There are several studies about credit card default modeling. Various Data Mining Techniques and modeling tools were utilized to create credit card default models. [13]'s paper provided a comprehensive literature survey related to applied data mining techniques in the credit scoring model. [14] built credit scoring models by using different data mining technologies to predict whether a customer will default or not and the authors come up with the conclusion that C5.0 decision tree model is the best model use for credit card applicant classification .Another study by [2] was conducted to prevent defaulting risk, it focused on clustering and classification techniques in coming up with application scoring and behavior scoring. On the other hand, a two-stage model for cardholder behavior scoring was developed by [15], utilizing Chi-square automatic interaction detector (CHAID) and artificial neural network (ANN) approaches to construct the initial classification models in the first stage. To optimize the model's accuracy, important variables from

the classification models were selected as input and output variables in a second-stage analysis using data envelopment analysis (DEA). This approach enabled the creation of a robust behavioral scoring model that delivers improved performance in credit risk assessment. Furthermore, [16] compared the Decision tree, Logistic Regression, and ANN techniques in credit scoring classification in which ANN performed the best. A different study by [17] used artificial neural networks (ANN) and decision tree were used to develop a model to classify and predict the behavior of cooperative members' behavior in paying their obligations. Moreover, a study [18] tried to improve a credit card score model by using text analysis on the application form. However, none of the mentioned studies explored feature selection with the Naive Bayes, Decision Table, DTNB, ADtree and LADtree classification algorithm.

C. Classification Metrics

In classification training, the evaluation metric plays a vital role in achieving the optimal classifier.[19] said that evaluation metric tasks are to decide the best classifier among different types of trained classifiers which focus on the future performance when used with test data and it serves as a discriminator to select the optimal solution among all generated solutions during the classification training. Shown in (1), (2), (3), and (4) are the different evaluation metrics. Equation (1) Precision is used to measure the percentage of how many are correctly predicted from the total predicted patterns in a positive class.

$$Precision = TP/(TP + FP) \quad (1)$$

Equation (2) Recall is the percentage of correctly labeled(predicted) from the actual observations.

$$Recall = (TP/(TP + FN) \quad (2)$$

Equation (3) F measure combines precision and recall as a measure of the effectiveness of classification in terms of the ratio of weighted importance on either recall or precision as determined by the β coefficient.

$$F\ Measure = \frac{\left((1+\beta)^2 \times \text{Recall} \times \text{Precision}\right)}{\left(\beta^2 \times \text{Recall} + \text{Precision}\right)} \quad (3)$$

Equation (4) Specificity metric is used to measure the percentage of correctly classified From the negative patterns.

$$Specifity = TN/(TN + FP) \quad (4)$$

ROC curve analysis is another metric. ROC curve displays a relation between sensitivity and specificity for a given classifier. That is an area of 1 represents a perfect test; an area of .5 represents a worthless test. For imbalance data, [20] stated that sensitivity, specificity, analysis of ROC curve is the best measure compared to accuracy.

III. OPERATIONAL FRAMEWORK

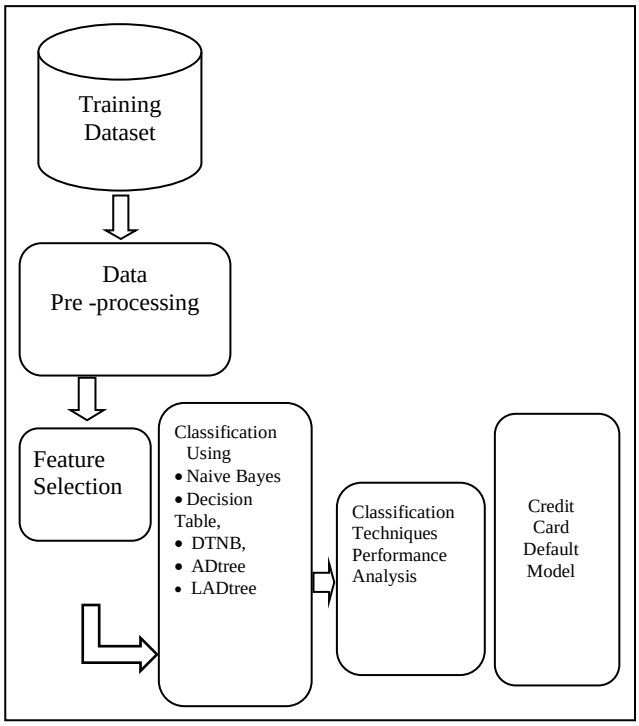

Figure 1- Credit Card Default Model Operational Framework

Shown in Figure 1 is the process of how the Credit Card Default Model was created. The dataset was pre–processed so that it will be compatible with the classification algorithms. A feature selection was done to determine which of the attributes is most relevant to the model. The Naive Bayes, Decision Table, DTNB, ADtree, and LADtree classifying algorithms were used to train the dataset, and classifier evaluation metrics were analyzed to determine which classifying technique is the best.

IV. RESEARCH METHODOLOGY

This study was based on the University of California, Irvine (UCI) Machine Learning Repository using the default credit card clients dataset[21] with 30000 samples.

The description of the 24 attributes is shown in Table 1.

TABLE 1-DESCRIPTION OF VARIABLES

Variable Name	Role	Variable Type	Description	
Default Payment	Target	Binary	Payment status 1: Default Payer 0:not Default Payer	
Limit Amount		input	Numerical	Amount of the given credit
Gender		input	Categorical	Male female
Education		Input	Categorical	Grad school, university, high school others
Marital status		Input	Categorical	Married, single, others
Age		Input		Age in Years
History of past payment measurement scale for the repayment status Pay_0 Pay_2 Pay_3 Pay_4 Pay_5 Pay_6		Input	categorical	past monthly payment September 2005 August 2005 July 2005 June 2005 May 2005 April 2005
Amount of bill statement Bill_Amt1 Bill_Amt2 Bill_Amt3 Bill_Amt4 Bill_Amt5 Bill_Amt6		Input	numerical	= amount of bill statement in September 2005 August 2005 July 2005 June 2005 May 2005 April 2005
Amount of previous payment Pay_Amt1 Pay_Amt2 Pay_Amt3 Pay_Amt4 Pay_Amt5 Pay_Amt6		Input	numerical	amount paid in September 2005 August 2005 July 2005 June 2005 May 2005 April 2005

A. *Preprocessing of Data*

The credit card default dataset was provided in .xls format. It was converted to .csv format and then converted to .arff format. It was then uploaded onto the data mining tool called WEKA.

B. *Attribute Selection*

Selection of relevant feature is necessary to reduce the dimensionality of the feature space and reduce the classification error [22]. Chi-squared Ranking Filter Feature selection method is Computationally cost effective[23] and robust with respect to the distribution of the data. Moreover, it Work well when there is large data representation [24].

Feature selection was done using Chi-squared Ranking Filter. An attribute was selected based on how it affects the predictive capability of the models. The chi-squared statistic of each attribute with respect to the class was computed.

Attributes were ranked and a specific number of the feature set are included for the creation of the model [25], [26] The result is shown in Figure2. from the result, only the following attributes were included in the experiment PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6,PAY_AMT1, LIMIT_BAL, PAY_AMT2.

t

C. *Classification Experiment*

The dataset was randomly partitioned with 70% of the data extracted for model training and 30% for model testing. NaiveBayes, DecisionTable, DTNB, ADTree and LADTree Classification Algorithm were used in the dataset.

V. RESULTS AND DISCUSSION

```
Attribute Evaluator (supervised, Class (nominal): 24 default_payment_next_month):
        Chi-squared Ranking Filter

Ranked attributes:
5349.346    6 PAY_0
3461.521    7 PAY_2
2600.781    8 PAY_3
2304.966    9 PAY_4
2165.211   10 PAY_5
1875.992   11 PAY_6
 955.348   18 PAY_AMT1
 938.335    1 LIMIT_BAL
 775.628   19 PAY_AMT2
 640.519   20 PAY_AMT3
 564.103   21 PAY_AMT4
 456.358   23 PAY_AMT6
 441.366   22 PAY_AMT5
 160.41     3 EDUCATION
 112.469    5 AGE
  47.905    2 SEX
  30.449   12 BILL_AMT1
  28.13     4 MARRIAGE
   0       13 BILL_AMT2
   0       14 BILL_AMT3
   0       15 BILL_AMT4
   0       16 BILL_AMT5
   0       17 BILL_AMT6

Selected attributes: 6,7,8,9,10,11,18,1,19,20,21,23,22,3,5,2,12,4,13,14,15,16,17 : 23
```

Figure 2- Feature Selection Result.

TABLE 2- CLASSIFIER EVALUATION METRICS RESULTS

Classifier	TP Rate Class	FP Rate	Precision	Recall	F-Measure	ROC Area
Naïve Bayes	0.629	0.303	0.765	0.629	0.661	0.726
Decision Table	0.819	0.513	0.802	0.819	0.797	0.74
DTNB	0.808	0.481	0.79	0.808	0.794	0.713
ADTree	0.836	0.503	0.82	0.836	0.817	0.768
LADTree	0.816	0.501	0.798	0.816	0.797	0.763

Shown in Table 2 were the results of the evaluation metrics of the different classification algorithms. ADT Tree classifier have the highest Precision value while Naïve Bayes classifier got the lowest value. On the other hand, in terms of Recall and F-Measure, ADT tree classifier have the best result closely followed by the Decision Table classifier and LAD Tree classifier respectively. ADT tree got the highest True positive classification result however; it is the Naïve Bayes got the lowest false positive classification result. ADtree had a Recall of .836 which means 83.6% of the actual observations were labeled (predicted) correctly. An ROC curve displays a relation between sensitivity and specificity for a given classifier. Value of 0.768 ROC area which was

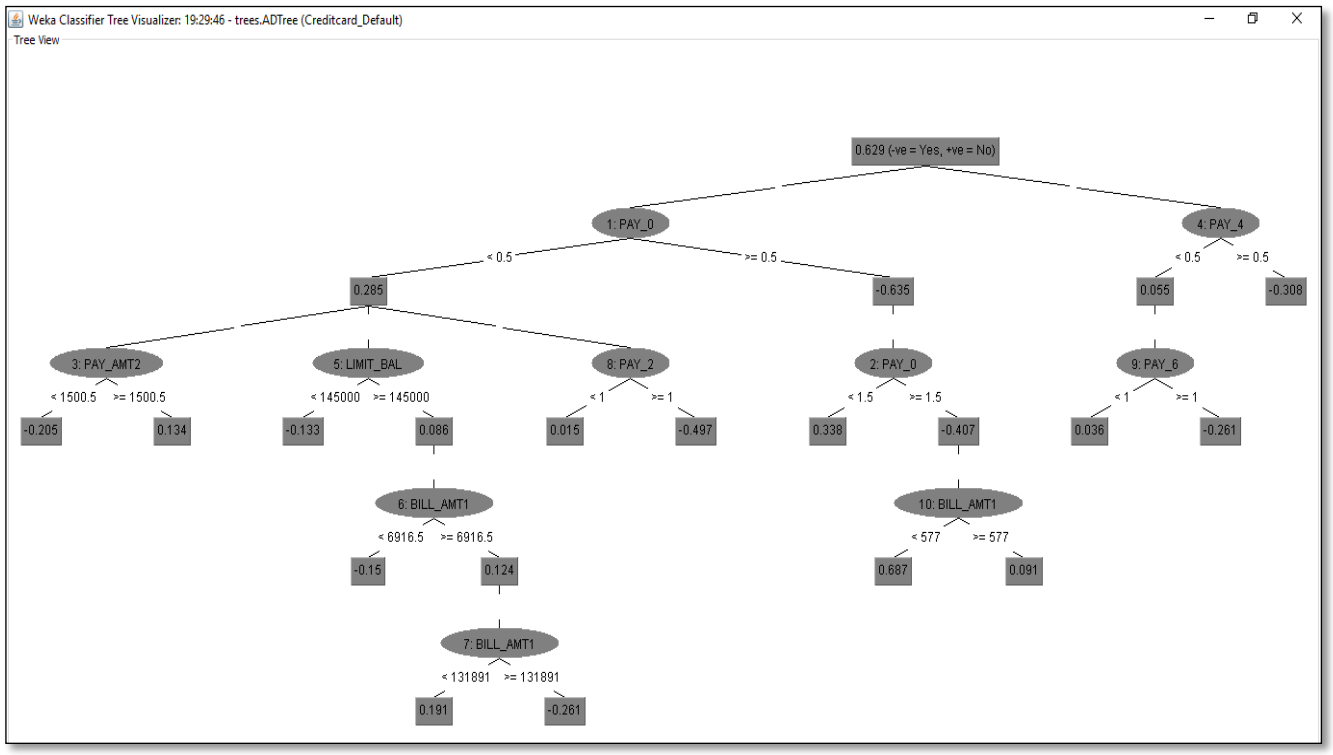

Figure 3- ADTree

Credit Card Default Model

The credit card default model was created using the optimal classifier ADTree (Alternating Decision Tree).

The alternating decision tree consists of decision nodes that specify a predicate condition and prediction nodes that contain a single value. Figure 3 is the graphical image of the ADTree while Figure 4 is in the form of rules.

Using the client's data (Table 3) in testing the model, an instance was classified by an ADTree by following all paths for which all decision nodes were true and the value of the prediction nodes was added. From the Legend: -ve = Yes, +ve = No: If the sum will have a negative value it means that the particular client is a credit card defaulter and if the value is positive it implies that the client is not a credit card defaulter.

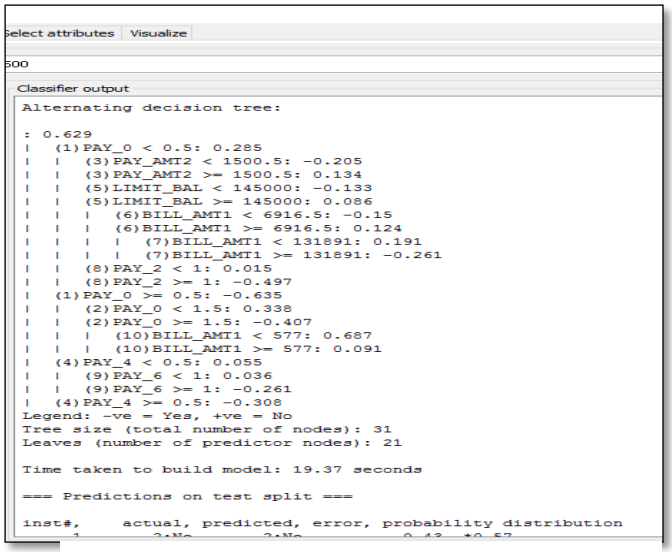

Figure 4- ADTree Equivalent Rules

TABLE 3- SAMPLE TEST DATA

Attributes	Data1 from Dataset	Data2 from Dataset
LIMIT_BAL	30000	200000
SEX	male	female
EDUCATION	university_graduate	Highschool_graduate
MARRIAGE	Single	Single
AGE	30	34
PAY_0	2	0
PAY_2	2	0
PAY_3	2	2
PAY_4	2	0
PAY_5	2	0
PAY_6	2	-1
BILL_AMT1	20732	11073
BILL_AMT2	21451	9787
BILL_AMT3	20808	5535
BILL_AMT4	21761	2513
BILL_AMT5	22762	1828
BILL_AMT6	23139	3731
PAY_AMT1	1347	2306
PAY_AMT2	0	12
PAY_AMT3	1300	50
PAY_AMT4	1500	300
PAY_AMT5	900	3738
PAY_AMT6	0	66
default payment next month	**Yes**	**No**

From the sample data taken from the dataset, a simulation was done using the ADTree Credit Card Default Model. And the result is shown in Table 4. It gives the correct prediction as compared to the actual data.

TABLE 4- SIMULATION USING SAMPLE DATA

Sample Data 1			
iteration	**Attribute**	**Data**	**Prediction Node Value**
0		.629	
1	Pay_0	2	-.635
2	Pay_0	2	-.407
3	Bill_Amt1	20732	.091
4	Pay_4	2	-.308
Total			-0.63
Prediction		Yes	Yes
Sample Data2			
iteration	**Attribute**	**Data**	**Prediction Node Value**
0			.629
1	Pay_0	0	.285
2	Pay_ Amt2	12	-.0.205
3	Limit_ Bal	200000	0.086
4	Bill_ Amt1	11073	0.124
5	Bill_ Amt1	11073	0.191
Total			1.11
Prediction		No	No

VI. CONCLUSION AND RECOMMENDATION

In this paper, a credit card defaulter model was created based on cardholder's demographic variables, spending records, and debt repayment status over the last six months using NAIVE BAYES DECISION TABLE, DTNB, ADTree, and LADTree classifying algorithm. Based on the classifier evaluation metrics the ADTree got the best results. The creation of the Credit card default model was based on the optimal classifier ADTree. This credit card default model can be utilized for early detection of possible credit card defaulter for intervention and it will be a good base for the financial institutions in making future policies to avoid credit card default.

For future research works, it is recommended that additional attributes can be explored and tested in the feature selection which might affect the model. Another recommendation is to try several splitting of the dataset to check which will give a better result for the model. It is also recommended that a dataset from a certain country/place can be used in the training set to check the suitability of the model in that particular country/place.

REFERENCES

[1] M. Khaled Yaseen, M. Raheem, and V. Sivakumar, "Credit Card Business in Malaysia: A Data Analytics Approach," 2020. [Online]. Available: www.ijacsa.thesai.org

[2] A. L. A. Li, W. L. W. Li, and Y. S. Y. Shi, "Study on the Application of Data Mining Algorithms in Credit Card Management," *2009 International Conference on E-Business and Information System Security*, pp. 0–4, 2009, doi: 10.1109/EBISS.2009.5138096.

[3] S. Arora, S. Bindra, S. Singh, and V. Kumar Nassa, "Prediction of credit card defaults through data analysis and machine learning techniques," in *Materials Today: Proceedings*, Elsevier Ltd, 2021, pp. 110–117. doi: 10.1016/j.matpr.2021.04.588.

[4] J. Cai, X. Liu, and Y. Wu, "SVM Learning for Default Prediction of Credit Card under Differential Privacy," in *PPMLP 2020 - Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, Association for Computing Machinery, Inc, Nov. 2020, pp. 51–53. doi: 10.1145/3411501.3419431.

[5] ITe-Cheng Hsu, Shing-Tzuo Liou, Yun-Ping Wang, and C.-L. Yung-Shun Huang, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada*.

[6] R. Srujana, "Evaluating the Effectiveness of Classification Algorithms Based on CCI," vol. 3, no. 9, pp. 15909–15916, 2014, doi: 10.15680/IJIRSET.2014.0309017.

[7] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer*

Science, vol. 2, no. 3. Springer, May 01, 2021. doi: 10.1007/s42979-021-00592-x.

[8] A. Gupta, "Classification of Complex UCI Datasets Using Machine Learning Algorithms Using Hadoop," *International Journal of Scetific & Techology Research*, vol. 4, no. 05, pp. 85–94, 2015.

[9] M. Hall and E. Frank, "Combining Naive Bayes and Decision Tables," *Intelligence*, pp. 2–3, 2008.

[10] R. Naseem *et al.*, "electronics Article," 2021, doi: 10.3390/electronics.

[11] Y. Zhao and Y. Zhang, "Comparison of decision tree methods for finding active objects," *Advances in Space Research*, vol. 41, no. 12, pp. 1955–1959, 2008, doi: 10.1016/j.asr.2007.07.020.

[12] A. Baicoianu, "A comparative study of some classification algorithms using WEKA and LAD algorithm," *Annals of the Tiberiu Popoviciu Seminar*, vol. 12, pp. 73–81, 2007.

[13] A. Keramati and N. Yousefi, "A proposed classification of data mining techniques in credit scoring," *Proc. 2011 Int. Conf. on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia*, pp. 416–424, 2011, [Online]. Available: http://www.iieom.org/ieom2011/pdfs/IEOM061.pdf

[14] W. a Li and J. b Liao, "An empirical study on credit scoring model for credit card by using data mining technology," *Proceedings - 2011 7th International Conference on Computational Intelligence and Security, CIS 2011*, pp. 1279–1282, 2011, doi: 10.1109/CIS.2011.283.

[15] C. I.-F. C. I-Fei, "Evaluate the performance of cardholders' repayment behaviors using artificial neural networks and data envelopment analysis," *Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on*, pp. 478–483, 2010.

[16] S. Imtiaz and A. J. Brimicombe, "A Better Comparison Summary of Credit Scoring Classification," 2017. [Online]. Available: www.ijacsa.thesai.org

[17] M. de M. Sousa and R. S. Figueiredo, "Credit Analysis Using Data Mining: Application in the Case of a Credit Union," *Journal of Information Systems and Technology Management*, vol. 11, no. 2, pp. 379–396, 2014, doi: 10.4301/S1807-17752014000200009.

[18] O. Ghailan, H. M. O. Mokhtar, and O. Hegazy, "Improving Credit Scorecard Modeling Through Applying Text Analysis," 2016. [Online]. Available: www.ijacsa.thesai.org

[19] M. Hossin and M. N. Sulaiman, "a Review on Evaluation Metrics for Data Classification Evaluations," vol. 5, no. 2, pp. 1–11, 2015.

[20] J. Stefanowski, "Data mining for imbalanced data : Improving classifiers by selective pre-processing of examples," *Technology (Singap World Sci)*, 2008.

[21] C. H. Yeh, I. C., & Lien, "UCI Machine Learning Repository: default of credit card clients Data Set." http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[22] F. Thabtah, F. Kamalov, S. Hammoud, and S. R. Shahamiri, "Least Loss: A simplified filter method for feature selection," *Inf Sci (N Y)*, vol. 534, pp. 1–15, Sep. 2020, doi: 10.1016/j.ins.2020.05.017.

[23] A. Thakkar and R. Lohiya, "Attack classification using feature selection techniques: a comparative study," *J Ambient Intell Humaniz Comput*, vol. 12, no. 1, pp. 1249–1266, Jan. 2021, doi: 10.1007/s12652-020-02167-9.

[24] Marianne Cherrington, Fadi Thabtah, Joan Lu, and Qiang Xu, "Classification and feature selection techniques in data mining," *2019 International Conference on Computer and Information Sciences (ICCIS)*, no. May, pp. 1–4, 2019.

[25] C. Kumar and R. Sree, "Application of Ranking Based Attribute Selection Filters To Perform Automated Evaluation of Descriptive Answers Through," *ICTACT Journal on Soft Computing*, pp. 860–868, 2014.

[26] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research*, vol. 21, no. 1, pp. 2334–6043, 2011, doi: 10.2298/YJOR1101119N.