BAREKENG: Journal of Mathematics and Its Application
December 2022 Volume 16 Issue 4 Page 1337–1346
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss4pp1337-1346

CREDIT CARD FRAUD DETECTION USING LINEAR DISCRIMINANT ANALYSIS (LDA), RANDOM FOREST, AND BINARY LOGISTIC REGRESSION

Muhammad Ahsan[1*], Tabita Yuni Susanto[2], Tiza Ayu Virania[3], Andi Indra Jaya[4]

[1,2,3,4]*Departement of Statistics, Institut Teknologi Sepuluh Nopember*
Jl. Teknik Kimia, Keputih, Kec. Sukolilo, Surabaya, 60111, Indonesia

Corresponding author's e-mail: [1] muh.ahsan@its.ac.id*

Abstract. *The growth of electronic payment usage makes the monetary tension of credit-card deception is changing into major defiance for finance and technology companies. Therefore, pressuring them to continuously advance their fraud detection system is crucial. In this research, we describe fraud detection as a classification issue by comparing three methods. The method used is Linear Discriminant Analysis (LDA), Random Forest, and Binary Logistic Regression. The dataset used is a dataset containing transactions made by credit cards. The challenge in this analysis is that the dataset is highly unbalanced, so Synthetic Minority Oversampling Technique (SMOTE) must perform better on the data. The dataset contains only continuous features that are transformed into Principal Component Scores (PCs). The results show that the binary regression algorithm, the Random Forest algorithm, and the Linear Discriminant Analysis with variables that have SMOTE have Area Under Curve (AUC) values greater than using the original variables. The largest AUC value was obtained by binary logistic regression with 90:10 separation data and Random Forest Algorithm with 60:40 separation data.*

Keywords: *binary logistics regression, credit card, fraud, linear discriminant analysis, random forests.*

Article info:

Submitted: 9th July 2022 *Accepted: 18th October 2022*

How to cite this article:
M. Ahsan, T. Y. Susanto, T. A. Virania and A. I. Jaya, "CREDIT CARD FRAUD DETECTION USING LINEAR DISCRIMINANT ANALYSIS (LDA), RANDOM FOREST, AND BINARY LOGISTIC REGRESSION", *BAREKENG: J. Math. & App.*, vol. 16, iss. 4, pp. 1337-1346, Dec., 2022.

1. INTRODUCTION

The main activities of the bank include lending, credit cards, investments, mortgages, and others. Credit cards are one of the most booming payment instruments in recent years. Credit cards are payment instruments that are quite easy to use, where customers can make payments only by showing cards that have been issued by certain banks when making transactions.

The ease of using a credit card causes many people to use a credit card. This triggers the emergence of credit card fraud. Credit card fraud events often occur and cause big financial disadvantages. Villains can apply some methods (phishing or trojans) to swipe the credit card data from the customer. Hence, developing a powerful fraud detection system is very crucial because it can detect fraud when hackers use stolen cards for consumption. For these purposes, machine learning techniques such as Linear Discriminant Analysis (LDA), Random Forests (RF), and Binary Logistics Regression (LR) can be applied.

Random Forests were first developed by Leo Breiman. The RF algorithms can be employed for both categorical and numerical response variables. Likewise, the independent variables can be in both numerical and categorical forms. From a computational point of view (PoV), the algorithms of RF are fascinating because they can tackle either regression or classification (binary or multiclass). This classifier algorithm form forests with a random number of trees [1].

The LDA algorithms are introduced to convert the variable into a lower-dimensional form. This approach can maximize the ratio of the variance of between-class to the variance of within-class, thereby guaranteeing maximum class separability [2]. On the other hand, the Logistic Regression (LR) method is one of the most important data mining methods employed by researchers. This method can tackle the classification from the binary and multiclass datasets.

Carcillo, et al. proposed combining unsupervised and supervised learning in credit card fraud detection [3]. Husejinovic analyzed with the title Credit Card Fraud Detection Using Naïve Bayes and C4.5 Decision Tree Classifiers [4]. Dhankhad, et al performed credit fraud analysis with the title Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection [5]. Malini and Pushpa did a credit fraud analysis with the title Analysis on Credit Card Fraud Identification Techniques Based on KNN and Outlier Detection [6]. Patil, Nermade, and Soni did credit fraud analysis with the title Predictive Modelling for Credit Card Fraud Detection Using Data Analytics [7]. Modi and Dayma did a credit fraud analysis with the title Review on Fraud Detection Methods in Credit Card Transactions [8]. Awoyemi, Adetunmbi, and Oluwadare use several machine learning techniques for fraud detection [9] Fu, Cheng, Tu, and Zhang [10] applied a convolutional neural network in detecting fraud.

In this paper, the dataset used is a dataset containing transaction records made by credit cardholders in Europe. This dataset consists of two days of transaction records. From the dataset, it is found 492 frauds out of 248,807 records. The dataset is very imbalanced, the ratio of frauds account (positive class) is 0.173% of all records. Due to privacy concerns, the input features are transformed into PCs.

To overcome an imbalance issue in the credit card dataset, the sampling is applied using SMOTE. In this paper, a comparison is made between the accuracy of the classification results with the original variables and variables that have been in SMOTE. Also, this paper is carried out variations of the experiment on the splitting data which are divided into 60:40, 70:30, 80:20, and 90:10.

2. RESEARCH METHOD

In this section, a brief explanation of data sources, research variables, performance metrics, and analysis steps are presented.

2.1 Dataset Description

The dataset bears two days of transactions made by a cardholder in Europe. The total of records is 284,807 which there are 492 (0.173%) transactions are labeled as fraud. This dataset is very imbalanced. Because presenting transaction records of a customer can be considered as a privacy issue so that most of the variables in the data are transformed into PCs. In this dataset, we have $V_1, V_2, \cdots$, and V_{28} PCs features, and

the remaining variables are 'time', 'amount', and 'class'. A detailed explanation of this matter is tabulated in Table 1.

Table 1. Variables of Credit Card Fraud Dataset

Feature	Description
Time	Time (in seconds) to specify the duration between the first transaction and the newest transaction
Amount	Number of Transaction
Class or Label	0 = No Fraud 1 = Fraud

2.2 Performance of Classification Model

The accuracy of a classifier can be stated as the ratio of accurately predicting the class (positive and negative class) in the dataset [13]. Accuracy measures for binary problem classification can be described in terms of four terms: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [14]. These terms can be arranged in a 2×2 matrix called confusion matrix as tabulated in Table 2.

Table 2. Confusion Matrix of Binary Classification

	Prediction (+)	Prediction (-)
Actual (+)	TP (True Positives)	FN (False Negative)
Actual (-)	FP (False Positives)	TN (True Negative)

The sensitivity can be explained as the ratio of TP over the total sample of the positive class. The formula is defined as follows:

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

The specificity can be explained as the ratio of TN over the total sample of the negative class. The formula is given as follows:

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

Furthermore, the accuracy and precision of the classifier can be computed from the following expressions.

$$Accuracy = \frac{TP+TN}{TP+RN+FP+FN} \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

The holdout method is one of the approaches in evaluating the accuracy of a model. The method consists of the process of splitting the data records into two datasets (the training and testing datasets). Regularly, two-thirds of the samples can be used for the training data and the remaining can be reserved for the testing dataset. In this paper, we adopted several proportions such as 60:40. 70:30, 80:20, and 90:10.

When evaluating the model selection for a binary class, we have some difficulties when estimating and calculating the best threshold for classifier with a categorical response. One reasonable method is to select or calculate a threshold that is a compromise between the number of FP and the number of FN. To find the threshold, the receiver operating characteristic (ROC) can be applied. The ROC curve is a graphical form of the TP rate (also known as sensitivity) as a function of the FP rate [15].

2.3 Steps Analysis

The analysis steps carried out in this paper are presented as follows.

a. Pre-processing data

b. Splitting data into training data and testing data

c. Resampling training data with SMOTE

d. Classifying using the LDA method, random forests, and logistic regression.

e. Comparing the results of the classification of the three methods by looking at the results of the classification accuracy and AUC.

f. Determine the best classification method.

g. Make conclusions and suggestions.

3. RESULTS AND DISCUSSION

In the current section, exploratory data analysis, classification analysis using linear discriminant analysis, random forest methods, and binary logistic regression are discussed.

3.1 Exploratory Data Analysis

To find out the characteristics of the data, descriptive statistical analysis was performed.

Table 3. Descriptive Statistic of Attributes

Class	*Mean*		*Minimum*		*Maximum*	
	Amount	***Time***	**Amount**	**Time**	**Amount**	**Time**
0 (No Fraud)	88.3	94838	0	0	25691	172792
1 (Fraud)	122.0	80747	0	406	2126	170348

Based on Table 3, we can find out that there are differences in the average "time" and "amount" for the class "fraud" and "not fraud". The average number of transactions (“amount”) on fraudulent credit cards is 122, which is higher than the number of transactions on credit cards that have not been fraudulent. The time on a fraudulent credit card transaction is smaller than a normal transaction.

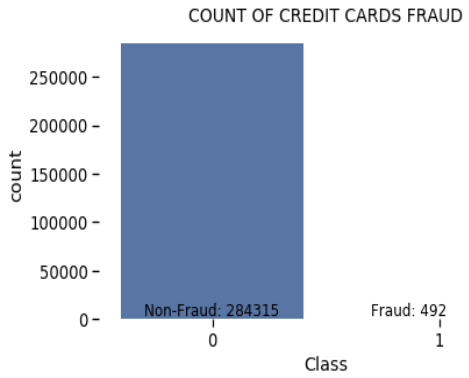

Figure 1. Bar Chart Credit Card Fraud

Figure 1 shows that 284,315 out of 284,807 (99.827%) credit card transactions did not occur fraudulently. Data has an imbalance (imbalance) in the target class category. To overcome this imbalance, resampling with SMOTE is performed so that the following class of fraud and non-fraud comparison is as follows.

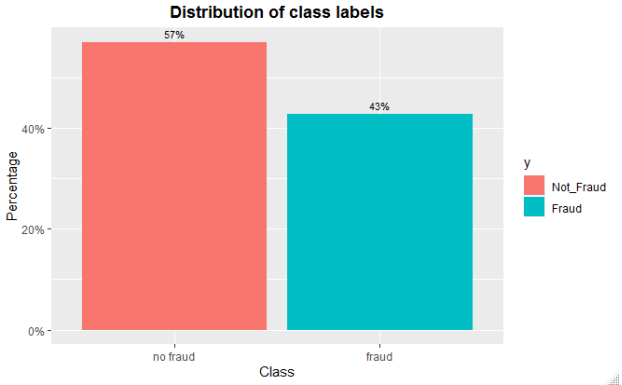

Figure 2. Bar Chart Credit Card Fraud After SMOTE

In Figure 2, it can be seen after the SMOTE is done on the data, 43% of data are fraudulent transactions while the other 57% are normal transactions. This shows that the imbalance in the data has been resolved.

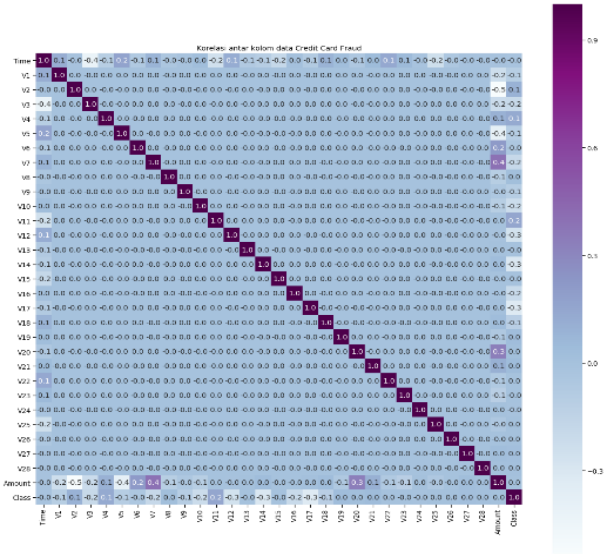

Figure 3. Correlation plot between Features

Based on Figure 3 it is known that the highest correlation is in the correlation between V_2 and Amount, which is the correlation value of -0.5, which means that V_2 and amount have an inverse relationship. A positive correlation exists in the correlation between the amount and the V_7 variable. In features V_1 to V_{28}, there is no relationship between the variables.

3.2 Data Pre-processing

Before processing, it is necessary to pre-process data. The first step in pre-processing is data cleaning, which is to find out the missing value and outlier. From 30 variables in the dataset, there are no variables that contain missing values, but there are variables that contain outliers. The number of outliers in each variable can be seen in Table 4.

Table 4. Outlier of The Attributes

Variable	Outlier	Variable	Outlier
TIME	0	V14	0.012%
AMOUNT	0.014%	V15	0.004%
V1	0.013%	V16	0.007%
V2	0.015%	V17	0.008%
V3	0.007%	V18	0.006%
V4	0.011%	V19	0.012%
V5	0.010%	V20	0.016%
V6	0.016%	V21	0.014%
V7	0.012%	V22	0.004%

Variable	Outlier	Variable	Outlier
V8	0.015%	V23	0.012%
V9	0.002%	V24	0.002%
V10	0.012%	V25	0.009%
V11	0.002%	V26	0.003%
V12	0.012%	V27	0.017%
V13	0.004%	V28	0.011%

Based on Table 4, it can be seen that the number of outliers in all variables is less than 5% so outliers do not need to be solved.

3.3 Classification Using Random Forest, Linear Discriminant Analysis, and Binary Logistic Regression

Following are the results of classification analysis on 3 methods with 2 scenarios variables without SMOTE and variables with SMOTE. All of that is compared to find which method produces good classification accuracy.

3.3.1. Binary Logistic Regression

In the Binary Logistic Regression algorithm, the analysis is performed with two scenarios, the first is the original data are classified by the random forest algorithm and the second scenario is classified on the data that has been done by SMOTE resampling. In this paper, we try out some of the training testing data sharing of 90:10, 80:20, 70:30, and 60:40. The following is the performance of each scenario.

Table 5. Performance of Logistic Regression

Sampling	Performance	90:10	80:20	70:30	60:40
Without SMOTE	Accuracy	0.999	0.999	0.999	0.999
	Specificity	0.571	0.571	0.621	0.609
	Sensitivity	1.000	0.999	0.999	0.999
	AUC	0.973	0.974	0.966	0.968
SMOTE	Accuracy	0.980	0.977	0.976	0.973
	Specificity	0.072	0.066	0.061	0.056
	Sensitivity	0.999	0.999	0. 999	0.999
	AUC	0.981	0.980	0.968	0.974

Table 5 shows that using the variables that have been resampled with SMOTE obtained accuracy and the greatest AUC is splitting data into 90% training and 10% testing. The accuracy obtained at 90:10 data splitting is 98% and the AUC value is 0.981. Whereas when using original variables or without SMOTE obtained greatest AUC is splitting data into 80% training and 20% testing. The accuracy obtained at 80:20 data splitting is 99.9% and the AUC value is 97.4%. The following is the ROC of Binary Logistic Regression with SMOTE.

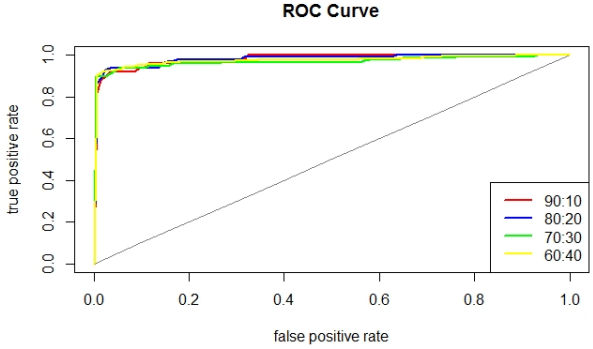

Figure 4. ROC Curve's Logistic Regression with SMOTE

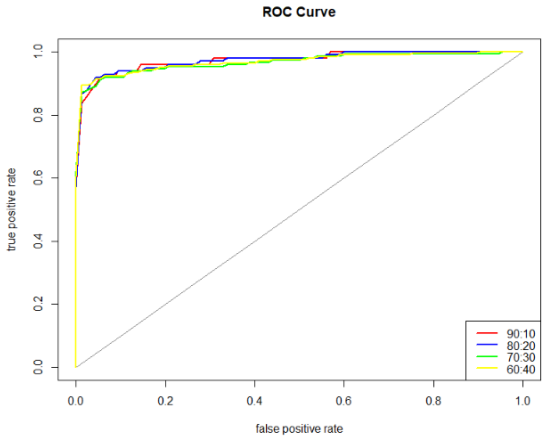

Figure 5. ROC Curve's Logistic Regression without SMOTE

3.3.2. Random Forest

In the random forest algorithm, the analysis is performed with two scenarios, the first is the original data are classified by the random forest algorithm and the second scenario is classified on the data that has been done by SMOTE resampling. In this paper, we try out some of the training testing data sharing of 90:10, 80:20, 70:30, and 60:40. The following is the performance of each scenario.

Table 6. Performance of Random Forest

Sampling	Performance	90:10	80:20	70:30	60:40
Without SMOTE	Accuracy	0.999	0.999	0.999	0.999
	Specificity	0.714	0.724	0.770	0.761
	Sensitivity	1.000	0.999	0.999	0.999
	AUC	0.975	0.976	0.980	0.976
SMOTE	Accuracy	0.992	0.992	0.992	0.993
	Specificity	0.992	0.867	0.892	0.904
	Sensitivity	0.857	0.992	0. 992	0.993
	AUC	0.978	0.980	0.980	0.981

Table 6 shows that using the variables that have been resampled with SMOTE obtained accuracy and the greatest AUC is splitting data into 60% training and 40% testing. The accuracy obtained at 60:40 data splitting is 99.3% and the AUC value is 0.981. %. Whereas when using original variables or without SMOTE obtained greatest AUC is splitting data into 90% training and 10% testing. The accuracy obtained at 90:10 data splitting is 99% and the AUC value is 97.3%. The following is the ROC of the Random Forest algorithm.

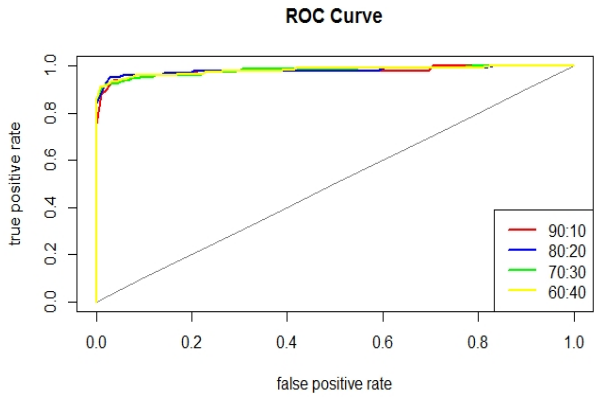

Figure 6. ROC Curve's Random Forest Algorithm with SMOTE

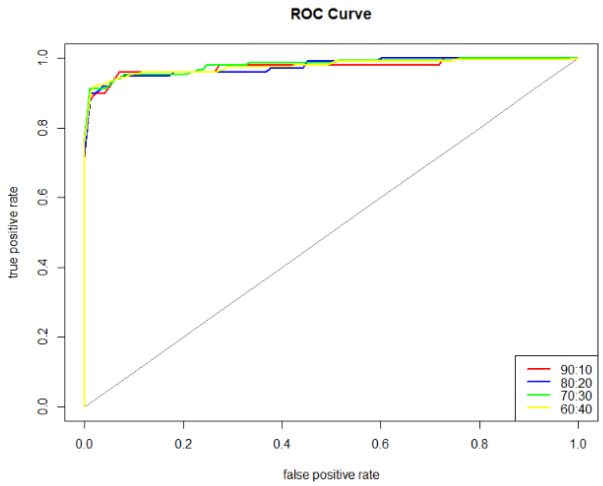

Figure 7. ROC Curve's Random Forest Algorithm without SMOTE

3.3.2.1.1. Linear Discriminant Analysis

The analysis is performed with two scenarios, the first is the original data are classified by the LDA method and the second scenario is classified on the data that has been done by SMOTE resampling using the LDA method. In this paper, we try out some of the training testing data sharing of 90:10, 80:20, 70:30, and 60:40. The following is the performance of each scenario.

Table 7. Performance of Linear Discriminant Analysis

Sampling	**Performance**	**90:10**	**80:20**	**70:30**	**60:40**
Without SMOTE	Accuracy	0.999	0.999	0.999	0.999
	Specificity	0.735	0.704	0.764	0.756
	Sensitivity	0.999	0.999	0.999	0.999
	AUC	0.867	0.852	0.882	0.878
SMOTE	Accuracy	0.916	0.989	0.990	0.990
	Specificity	0.820	0.786	0.831	0.838
	Sensitivity	0.989	0.989	0.990	0.990
	AUC	0.904	0.887	0.911	0.914

Table 7 shows that using the variables that have been resampled without SMOTE obtained the greatest AUC is splitting data into 70% training and 30% testing. The accuracy obtained at 70:30 data splitting is 99.9% and the AUC value is 88.2%. Whereas when using variables that have been resampled with SMOTE obtained greatest AUC is splitting data into 60% training and 40% testing. The accuracy obtained at 60:40 data splitting is 99% and the AUC value is 91.4%. The following is the ROC of the LDA.

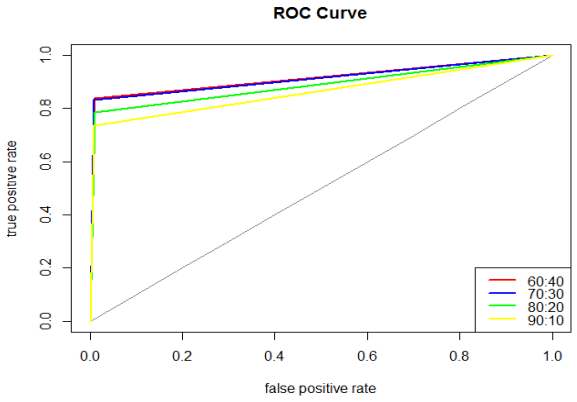

Figure 8. ROC Curve's Linear Discriminant Analysis with SMOTE

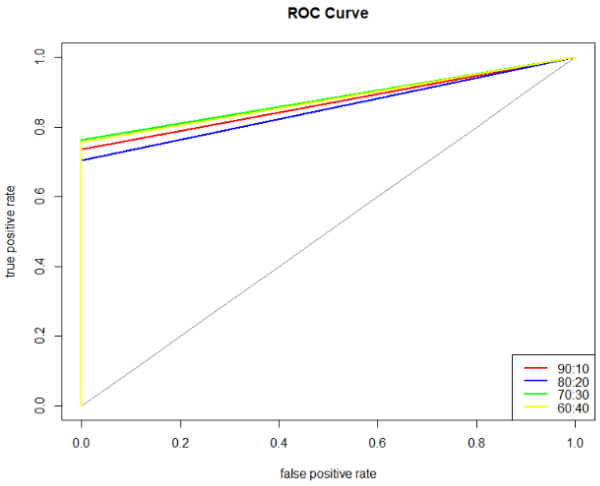

Figure 9. ROC Curve's Linear Discriminant Analysis without SMOTE

4. CONCLUSIONS

Based on the analysis that has been done on the credit card fraud dataset, the result shows that the binary logistic regression algorithm, Random Forest algorithm, and Linear Discriminant Analysis with variables that have been SMOTE have an AUC value greater than using original variables. Among the three methods, the greatest AUC value is obtained in binary logistic regression with splitting data 90:10 and in Random Forest Algorithm with splitting data 60:40. In Binary Logistic Regression Logistic with splitting data 90:10, the AUC value was obtained 98.1% and the resulting accuracy was 98.0%. While in Random Forest Algorithm with splitting data 60:40, the AUC value was obtained 98.1% and the resulting accuracy was 99.3%. The suggestion for further research is to use several other methods to detect credit card fraud and use other methods to tackle the imbalance in real-world data.

REFERENCES

[1] M. Ahsan, A. K. Anam, E. Julian, and A. I. Jaya, "Interpretable Predictive Model of Network Intrusion Using Several Machine Learning Algorithms," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 16, no. 1, pp. 57–64, 2022.

[2] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *Ai Commun.*, vol. 30, pp. 169–190, 2017, doi: 10.3233/AIC-170729.

[3] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Inf. Sci. (Ny).*, vol. 557, pp. 317–331, 2021.

[4] A. Husejinovic, "Credit card fraud detection using naive Bayesian and C4. 5 decision tree classifiers," *Husejinovic, A.(2020). Credit card Fraud Detect. using naive Bayesian C*, vol. 4, pp. 1–5, 2020.

[5] S. Dhankhad, E. Mohammed, and B. Far, "Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 122–125.

[6] N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2017, pp. 255–258.

[7] S. Patil, V. Nemade, and P. K. Soni, "Predictive modelling for credit card fraud detection using data analytics," *Procedia Comput. Sci.*, vol. 132, pp. 385–395, 2018.

[8] K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, pp. 1–5.

[9] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, 2017, pp. 1–9, doi: 10.1109/ICCNI.2017.8123782.

[10] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," in *International conference on neural information processing*, 2016, pp. 483–490.

[11] M. Komorowski, D. Marshall, J. Salciccioli, and Y. Crutain, "Exploratory Data Analysis," in *Secondary Analysis of Electronic Health Records*, 2016, pp. 185–203.

[12] R. A. Johnson and G. K. Bhattacharyya, *Statistics: principles and methods*. John Wiley & Sons, 2019.

[13] A. C. Sitepu, W. Wanayumini, and Z. Situmorang, "Determining Bullying Text Classification Using Naive Bayes Classification on Social Media," *J. Varian*, vol. 4, no. 2, pp. 133–140, 2021.
[14] I. K. P. Suniantara, G. Suwardika, and S. Soraya, "Peningkatan Akurasi Klasifikasi Ketidaktepatan Waktu Kelulusan Mahasiswa Menggunakan Metode Boosting Neural Network," *J. Varian*, vol. 3, no. 2, pp. 95–102, 2020.
[15] P. Galdi and R. Tagliaferri, "Data Mining: Accuracy and Error Measures for Classification and Prediction," 2018.