Tugas Akhir

by Muhamad Sopiyan

Submission date: 22-Dec-2021 11:27AM (UTC+0700)
Submission ID: 1734884427
File name: uhamad_Sopiyan_183112706450094_-Universitas_Nasional_revisi.docx (297.99K)
Word count: 5274
Character count: 28782

Fraud Detection Using Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) Algorithms on Credit Cards

Muhamad Sopiyan[1],*Fauziah[2], Yunan Fauzi Wijaya[3]

[1,2,3]*Informatics, Faculty of Communication and Information Technology, Universitas Nasional, Indonesia*

[1,2,3]sofyanm625@gmail.com, fauziah@civitas.unas.ac.id, yunan.fw@civitas.unas.ac.id

Abstract - The following credit card records were used in this study of 284.807 transactions made by credit card holders in Europe for two days from the Kaggle dataset. This is a very poor data set, having 492 transactions, an imbalance of only 0.172% of the 284.807 transactions. The purpose of this study is to obtain the best model and then simulate it by electronically detecting unauthorized financial transactions in bank payment systems. The dataset for this study is unbalanced class data with 99.80% for the major class and 0.2% for the minor class. This type class-imbalanced data problem is solved by applying method a combination of minority oversampling techniques using Synthetic Minority Oversampling Technique (SMOTE). To determine the most appropriate and accurate classification in solving class balance problems, comparisons were made with the Random Forest Classifier (RFC), Logistic Regression (LGR), and Gradient Boosting Classifier (GBC) algorithms. The test results in this study are the Random Forest Classifier (RFC) algorithm is better than other algorithms because it has the highest accuracy the percentage of data-train is 100% and data-test is 99.99% and the evaluation of the AUC score as a result of algorithm testing is 0.9999.

Keywords: Data Meaning, Fraud Detection, Gradient Boosting Classifier (GBC), Logistic Regression (LGR), Random Forest Classifier (RFC).

I. INTRODUCTION

Currently, the use of computer technology as a means of supporting transaction activities is very popular according to the number of credit card users and even as a means of daily payment. Utilization of computer technology is needed for various kinds of electronic transactions.[1] In the world of technology, the term machine learning is not new. However, in the era of increasingly rapid technological developments in recent years, the term machine learning has become increasingly popular and has begun to be studied a lot. One of the functions of machine learning is to get the "value" of a data set. This has caused many companies from various industries, especially in the banking sector, to be interested in applying machine learning technology.[2] In terms of the use of machine learning technology, in this research process using a credit card fraud registry consisting of 284,807 transactions made by credit card holders in Europe for two days in progress obtained from the kaggle dataset. Information The dataset contains a highly unbalanced data set, containing 492 fraudulent transactions, which represents only (0.172% of the 284,807 transactions. For some information about the characteristics of these datasets such as V1, V2,... V28 are the main components that are obtained with PCA. The "Time" attribute contains the seconds that elapsed between each transaction in the data log. Attribute "Amount" is the number of transactions, This attribute can be used as paid learning. The attribute `Class` is a response variable and takes the value 1 if fraud occurs and 0 otherwise.[3] With the presentation of the problems faced in this study is a collection of data with unbalanced categories, which compares 99.80% of the major categories and 0.2% of the minor categories of the overall transactions that take place. This kind of unbalanced data problem will be solved by applying a combination of oversampling methods, namely the minority resampling technique using the synthetic minority oversampling technique (SMOTE).[4] In an approach to solving this type of highly unbalanced binary classification problem, the first step to take is to remove some records from the majority class, while the second adds more random copies to the minority class. Both techniques are carried out until the majority and minority classes are balanced and a balanced class distribution visualization and bar chart are produced with the same data sample in

the minority class.[5][6] The next technique is to get the best algorithm accuracy and then simulate it into the bank payment system. The fraud detection generated in electronic financial transactions in this study uses the accuracy of the machine learning algorithm performance comparison, namely Random Forest Classifier (RFC), Logistic Regression (LGR), Gradient Boosting Classifier (GBC)[7] these three types of machine learning algorithms will use a combination of SMOTE parameters for configuration and optimization to get the model with the best accuracy score and precision score. Furthermore, testing the point confused matrix data for validation (primary data) of each tested data to test the accuracy of each tested algorithm and the evaluation of the AUC score as a result of algorithm testing in ensuring the accuracy of the performance of the algorithm being tested.

II. METHOD

The method used in this research is the sample data method. The reason for using this method is because it is the best way to collect datasets from search results and learn datasets from Kaggle datasets. This research is systematically divided into several stages of research consisting of data collection (dataset), data processing and reading, modelling, experimentation and model testing as well as evaluation and validation:

A. Data collection (Dataset)

The dataset contains transactions made with credit cards by cardholders in Europe from the Kaggle dataset. This register represents transactions that occurred in the last two days, from the information the dataset has 492 fraudulent transactions, which represented only 0.172% of the 284,807 transactions. For some information about the characteristics of datasets like V1, V2,..V28 is the main component obtained by the PCA process. The "Time" attribute contains the seconds that elapsed between each transaction in the log data. Attribute "Amount" is the number of transactions, this attribute can be used as paid learning. The 'Class' feature is a response variable and takes a value of 1 if there is fraud and 0 if there is no fraud.

B. Data processing and reading

The data is processed based on the results of data collection and data cleaning processes to overcome data problems such as data anomalies, missing data values, data redundancy, and inappropriate data. The data is then selected and grouped by type and function to divide it into training and testing data so that it can be applied to the classification algorithm that will be tested. Following the approach proposed in previous studies, the first step to be taken is to apply a resampling method such as SMOTE. The next step is to model the training data. To measure the performance of this classification algorithm, it is done by using the confusion matrix obtained from the validation process. The validation results are used to measure the performance of each model. The results are obtained from the measurement of the performance of the model used.[8] The development carried out in this research is the addition of Machine Learning (Supervised Learning) algorithms, namely the Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) algorithms by applying a combination of resampling methods such as SMOTE and the comparison of the confusion matrix values used. Obtained from the validation and evaluation of the model with the Accuracy AUC value for the training and testing subset of the dataset. So that the results can be illustrated by visualizing the Area Under ROC (AUROC) as an indicator to measure the performance of the binary number classifier, which can be explained in Fig. 1.

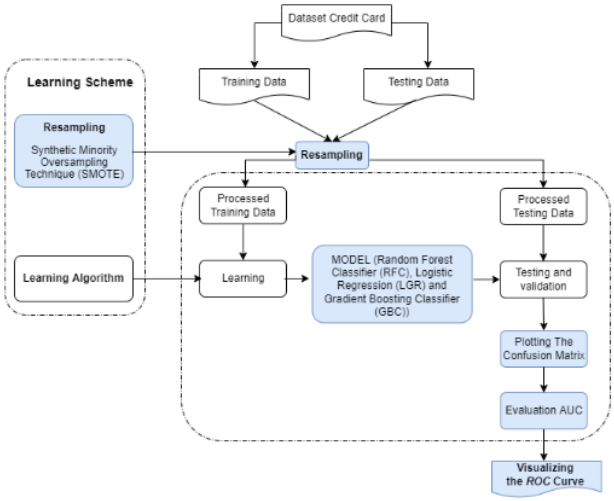

Fig. 1 Proposed Research Method

Fig. 1 explain the flow diagram of the model to be proposed. The initial step that will be carried out is adding data balancing parameters by resampling the sampling process with the oversampling method, namely taking the minority class in such a way that the proportion in the sample is greater than the original proportion, such as the use of the Synthetic Minority Oversampling Technique (SMOTE) method on credit card datasets.[9] In order for the dataset to be more balanced, the next step is to model the training data using the Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) algorithms as comparison and tested with data through a validation process. The validation results are used to measure the performance of each algorithm with the accuracy of calculating the AUC value. Comparisons are made by comparing the performance of each Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) algorithms to measure the accuracy of the performance of each resulting algorithm. Classification is also carried out by comparing the confusion matrix values obtained from the validation process based on the values of Accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate), Precision (Positive Predictive Value) and also the visualization value of the Area Under ROC (AUROC).

C. Modelling

Before the training and testing process is carried out, the data is sampled using a combination of the Synthetic Minority Oversampling Technique (SMOTE) method in dealing with class-imbalanced (class-imbalanced) fields, then training and data testing will be carried out using the Random Forest Classifier (RFC) classification method. Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) as well as the accuracy of the AUC evaluation value and the comparison of the confusion matrix values obtained from the validation process.[10]

I). Synthetic Minority Oversampling Technique (SMOTE). Here are the steps in the SMOTE technique.

1. Calculate the difference between the vectors of the first instance with the k-nearest neighbours.
2. From the difference multiplied by a random number between 0 to 1.
3. The result of the difference is added to the main vector, so it will create a new instance.

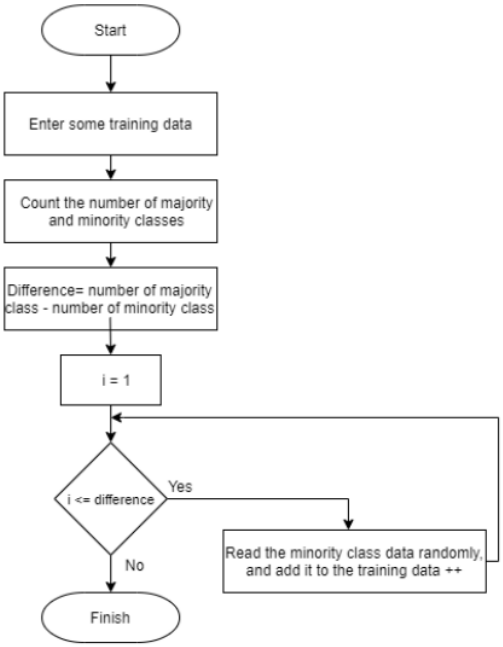

Fig. 2 Flowchart Synthetic Minority Oversampling Technique (SMOTE)

Fig. 2 describes the flow process of the dataset which is transformed using a combination of the Synthetic Minority Oversampling Technique (SMOTE) method so that it has the same sample in the minority class. This oversampling technique aims to multiply the minority class sample so that it is the same as the other majority class by duplicating the minority class sample randomly. In this method, the sample from the minority class is randomly selected and duplicated. From the resulting process only increases the size of the minority class by replicating the same information:

II). Random Forest Classifier (RFC)

The Random Forest Classifier (RFC) algorithm has been widely used in data mining research for both classification and regression because of its superior performance and simple structure. In general, the development of the Random Forest Classifier carried out from the bagging process lies in the sorter selection process.[11] In the Random Forest Classifier, the disaggregation selection involves only a few predictor variables which are taken at random. The implementation steps of the Random Forest Classifier (RFC) algorithm in this study are described as follows:

1. Use bootstrap resampling with returns with return values to extract n data samples from the initial data set.
2. Compile a classification tree from each resampling bootstrap data set, and determine the best rank based on randomly selected predictor variables. The number of randomly selected variables can be determined by calculating log2 (M + 1), where M is the number of predictors or usage, where p is the number of predictors.
3. Prediction of the classification of the sample data according to the classification tree formed.
4. Repeat steps 1 to 3 until you get the required number of classification trees. Iteration is done K times.
5. Combining the prediction results of the classification tree based on the majority Voting rule to predict the ranking of the final sample data:

III). Logistic Regression (LGR)

Logistic Regression Algorithm (LGR) is used in this study as a data analysis and statistical technique that aims to determine the relationship between several classification variables where the response variables are categorical, both nominal and ordinal in the research dataset. The mathematical modelling used in this study aims to analyse the

relationship between many variables and one binary variable in the logistic regression stages to find the logistic equation in this research which is described in the form of the equation below:[12]

$$\pi(x) = \frac{e^{\beta+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}} \qquad (1)$$

Equation "(1)" get 1 – π(x) as follows:

$$1 = \pi\,(x) = 1 - \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j}}$$

$$1 = \pi\,(x)\ \frac{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j x_j} - e^{\beta_0\sum_{j=1}^{p}\beta_j x_j}}{1+e^{\beta_0\sum_{j=1}^{p}\beta_j x_j}}$$

$$= \frac{1}{1+e^{\beta_0\sum_{j=1}^{p}\beta_j x_j}}$$

$= \frac{n\,(x)}{1=\pi\,(x)}$ as follows:

$$\frac{\pi\,(x)}{1-\pi\,(x)} = e^{\beta_0\sum_{j=1}^{p}\beta_j x_j}$$

The logistic equation:

$$g\,(x) = ln\,(\frac{\pi\,(x)}{1-\pi\,(x)}) \qquad (2)$$

$$= ln\,(e^{\beta_0\sum_{j=1}^{p}\beta_j x_j})$$

$$= \beta_0 + \sum_{j=1}^{p}\beta_j x_j$$

IV). Gradient Boosting Classifier (GBC)

The Gradient Boosting Classifier (GBC) algorithm in the process of this research method is divided into two categories, namely refractive error and variance error. Because gradient enhancement is one of the algorithms used to minimize case classification modelling errors. The Gradient Boosting Classifier (GBC) algorithm can be used to predict not only continuous target variables (as regression) but also categorical target variables (as classifiers). When used as a regression the cost function is the mean squared error (MSE). And when used as a classification, the cost function is a logarithmic loss. The following are the steps to perform the classification using the Gradient Boosting Classifier (GBC) algorithm in this study: The model is based on a subset of data.

1. Use the model to make predictions on the entire data set.
2. Calculate the error by comparing the predicted value with the actual value.
3. Create a new model using the calculated error as the target variable. The goal is to find the best clearance to minimize error.
4. The predictions from the new model are combined with the previous predictions.
5. Calculate the new error using the predicted value and the true value.
6. Repeat this process until the error function does not change or reaches the maximum estimator:

D. Experiment and Model Testing

The test model carried out in this research is using a computer specification with an Intel Core i5-6200U processor, CPU @ 2.30 GHz, 8 GB RAM, with Windows 10 64 bit operating system, and Anaconda Navigator analysis tools, distribution of Python Packages from Continuum Analytics. Experiments and algorithm testing are carried out to obtain the accuracy of the performance of each tested algorithm by testing the credit card dataset prediction model.

E. Evaluation and Validation

This study will use data-validation measurements (primary data) to test the accuracy of each model tested using the confusion matrix value obtained from the validation process based on the accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate) and Precision (Positive Predictive Value) as a measuring point for testing N parameters in the combination of Synthetic Minority Oversampling Technique (SMOTE) and depth in testing the Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) algorithms and evaluating the model with The best Accuracy AUC value so that it can guarantee the accuracy performance of each algorithm that will be tested.

III. RESULTS AND DISCUSSION

In this chapter, we will discuss the results of research conducted using the Anaconda Navigator Python distribution package application analysis tool from Continuum Analytics. The data will be processed to predict the achievement of using the accuracy of each proposed algorithm performance by testing the credit card dataset.[13] The dataset used is data from Kaggle "Credit Card" which has 429 frauds from 284,807 transactions. This study was conducted to produce the highest accuracy value for each performance of the proposed algorithm, namely Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) by comparing the performance of these algorithms and applying a combination of minority class oversampling methods using Synthetic Minority Oversampling Technique (SMOTE) to solve class-imbalanced data problems (class-imbalanced), and also data-validation measurement on (primary data) to test the accuracy of each model being tested using the confusion matrix value obtained from the validation process based on the value of Accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate), Precision (Positive Predictive Value) on the effectiveness of each performance algorithm tested. As well as evaluating the model with the Accuracy AUC value for the training and testing subset of the dataset, so that it can determine the performance of each algorithm that is most appropriate and ensures great accuracy can be used.

A. Feature Engineering and Data Modelling

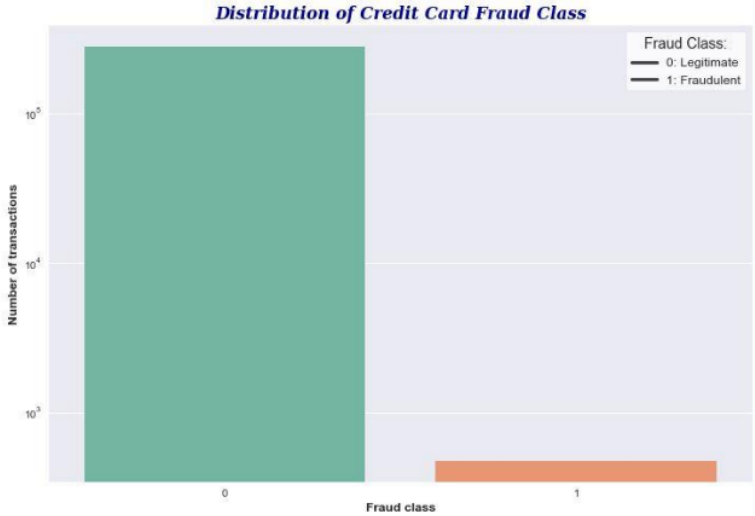

Fig. 3 Unbalanced distribution of fraud classes

Fig. 3 presents the results of the bar chart visualization, it can be seen that the number of valid credit card transactions is much higher than the number of fraudulent transactions. This is clearly to be expected because fraud detection is one of the problem domains where the class distribution is inherently unequal. If fraudulent transactions occur higher than the legal one, this indicates that the banking institution is facing a very serious security breach that can cause loss of revenue, disruption in operations and loss of reputation or customer trust in the buying, leasing and banking services. However, in dealing with this very large data imbalance, it can be solved by applying a combination of minority class oversampling methods using the Synthetic Minority Oversampling Technique (SMOTE) because otherwise it can hamper the accuracy of the classification model to be tested.

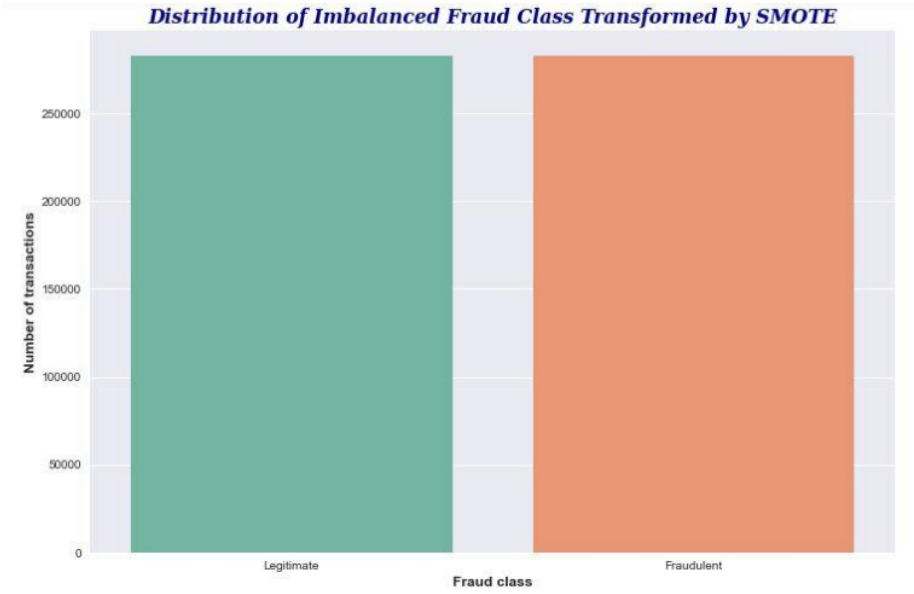

Fig. 4 Dataset Transformation Using *Synthetic Minority Oversampling Technique* (SMOTE)

Fig. 4 presents the results of the bar chart that has passed the transformation stage using the combination method of Synthetic Minority Oversampling Technique (SMOTE) as an approach to solving the class-imbalanced type of binary classification problem. One of the simplest and most widely adopted resampling techniques is oversampling because by duplicating data from the minority class, so that the number of minority classes approaches the majority class. Adding more random copies to the minority class. Both techniques 10 carried out until the majority and minority classes are balanced. The minority class oversampling method using the Synthetic Minority Oversampling Technique (SMOTE) in this study was carried out because it is a complex quality resampling technique and introduces small variations into the minority class observation copy instead of the exact copy, resulting in a more diverse synthetic sample.[14]

B. Evaluation and Validation

At this stage, evaluation and validation are carried out. The process results from a model that is measured based on the accuracy of the classification performance which is evaluated using a confusion matrix, as shown in Table I, by measuring the accuracy and AUC (Area Under Curve) values of a built model.

TABLE I
CONFUSION MATRIX

Prediction Class	Actual Class	
	No	Appropriate
No	TP	FP
Appropriate	FN	TN

Table I the predictive confusion matrix can be explained for the actual class, which consists of a True Positive (TP) component as a correctly identified positive class, then a False Positive (FP) component is a negative class that is incorrectly identified, then a False Negative (FN) component is a positive class that has been incorrectly identified and a True Negative component (TN) is a negative correctly identified class. The evaluation is calculated from the results of the confusion matrix with the formula as in "(3)" to "(9)".

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$\text{Sensitivity (SN)} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{Specificity (SP)} = \frac{TN}{TN+FP} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{Positive predictive values (PPV)} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Negative Predictive value (NPV)} = \frac{TN}{TN+FN} \quad (8)$$

$$\text{F – Measure (F)} = \frac{2TP}{2TP+FP+FN} \quad (9)$$

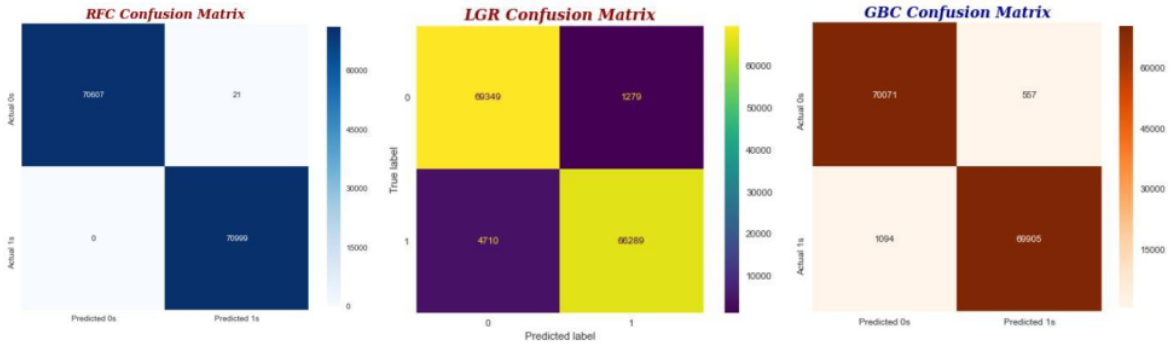

Fig. 5 Confusion Matrix Algorithm Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC)

Based on the results of Fig. 5 this study uses a plotting confusion matrix for validation of model performance in a highly unbalanced binary class data set. Validation is done to determine the performance of the algorithm used and its effectiveness by using a confusion matrix. by testing the results of the validation model based on primary data, the data visualized from the results of the confusion matrix is presented in color with darker color descriptions the smaller the correlation value and the lighter the color, the greater the correlation value. The `Class` attribute in this confusion matrix can be explained by the response variable by taking the predicted 1s if fraud occurs and the predicted 0s if not. The Confusion Matrix Algorithm Random Forest Classifier (RFC) is explained with the displayed values for True Negatives (TN) of 70607, False Positives (FP) of 21, False Negatives (FN) of 0, True Positives (TP) of 70999. Meanwhile Confusion Matrix Algorithm Logistic Regression (LGR) is explained with the displayed value of True Negatives (TN) of 69349, False Positives (FP) of 1279, False Negatives (FN) of 4710, True Positives (TP) of 66289. And Confusion Matrix the Gradient Boosting Classifier (GBC) algorithm is explained with the displayed values for True Negatives (TN) of 70071, False Positives (FP) of 557, False Negatives (FN) of 1094, True Positives (TP) of

69905. From the results of the validation test on the model based on the results of the Confusion Matrix and the statistics above, it can be seen that the model used is very sensitive, which is actually a concern for banking institutions because false negatives are more dangerous than false positives.

Of course, in a credit card fraud detection system, an effective performance algorithm with a good degree of classification accuracy should also have far fewer false positives (FP) because the error can cost the bank billions of dollars and the customer will likely not use the credit card again.

C. Results

Table II is the result of data-evaluation and data-validation of primary data with a comparison of the 3 algorithms used in this study which contains the validation value of the confusion matrix based on the value of Accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate), Precision (Positive Predictive Value) and evaluation of the AUC score as a result of algorithm testing in ensuring the accuracy of performance on the algorithm process being tested. AUC obtained from the ROC (receiver operating character) curve. The AUC value is used for classification analysis to determine the best algorithm for predicting the data. The results of this study using three algorithms for comparison can be seen in the following table:

TABLE II
ALGORITHM COMPARISON RESULTS

No	Algorithm	Accuracy		Evaluation AUC	Validation		
		Data-train	Data-test		*Sensitivity (True Positive Rate)*	*Specificity (True Negative Rate)*	*Precision (Positive Predictive Value)*
1	**Random Forest Classifier (RFC)**	100%	99.99%	0.9999	100%	99.97%	99.97%
2	**Logistic Regression (LGR)**	95.86%	95.77%	0.9916	93.36%	98.18%	98.10%
3	**Gradient Boosting Classifier (GBC)**	98.86%	98.83%	0.9994	98.45%	99.21%	99.20%

Table II it can be explained that the best algorithm performance based on the accuracy and AUC values is generated by the Random Forest Classifier (RFC) algorithm with the highest accuracy value at 100% data-train percentage and 99.99% data-test and point confused matrix data testing. for validation (primary data) Sensitivity (True Positive Rate) is 100%, Specificity (True Negative Rate) is 99.97%, Precision (Positive Predictive Value) is 99.97% and the evaluation of the AUC score as a result of algorithm testing is 0.9999, Logistic Regression (LGR) with a data-train percentage of 95.86% and a data-test of 95.77% and the evaluation of the AUC score as a result of algorithm testing of 0.9916 while the Gradient Boosting Classifier (GBC) produces an accuracy value with a data-train percentage of 98.86% and data-test of 98.83% and the evaluation of the AUC score as the result of algorithm testing is 0.9994. From the performance of the data-validation confused matrix generated such as Sensitivity, Specificity, and Precision, the best algorithm performance accuracy is owned by the Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC) algorithms. Meanwhile, logistic regression (LGR) has the worst performance from all evaluations (accuracy, sensitivity, specificity, precision, and AUC). Based on the results of the evaluation and test, Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) have significant differences, in contrast to Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC) which do not have significant differences. So the results of testing on the credit card dataset, the algorithm that has the best performance is the Random Forest Classifier (RFC) algorithm, while the Logistic Regression (LGR) algorithm have poor performance. This study confirms from previous research which explains that in solving credit card fraud classification problems the algorithm that has the best performance is the Neural Network algorithm with an accuracy value of 93.59% and an AUC score of 0.977.[15] Thus, in this study, it is explained that the results of research using the Random Forest Classifier (RFC) algorithm provide higher accuracy results than previous studies using the Neural Network algorithm with an accuracy value on the data-train percentage of 100% and data-test of 99.99% and the evaluation of the AUC score as the result of testing the algorithm is 0.999 so that it has a better performance effectiveness on the credit card fraud dataset.

D. Visualizing the ROC Curve

Receiver operating characteristics (ROC) curves are often used as indicators to measure the performance of binary number classifiers. This is not a matrix model, but rather graphical representation of true positive (TPR) and false positive (FPR) values with different rating thresholds from 0 to 1.

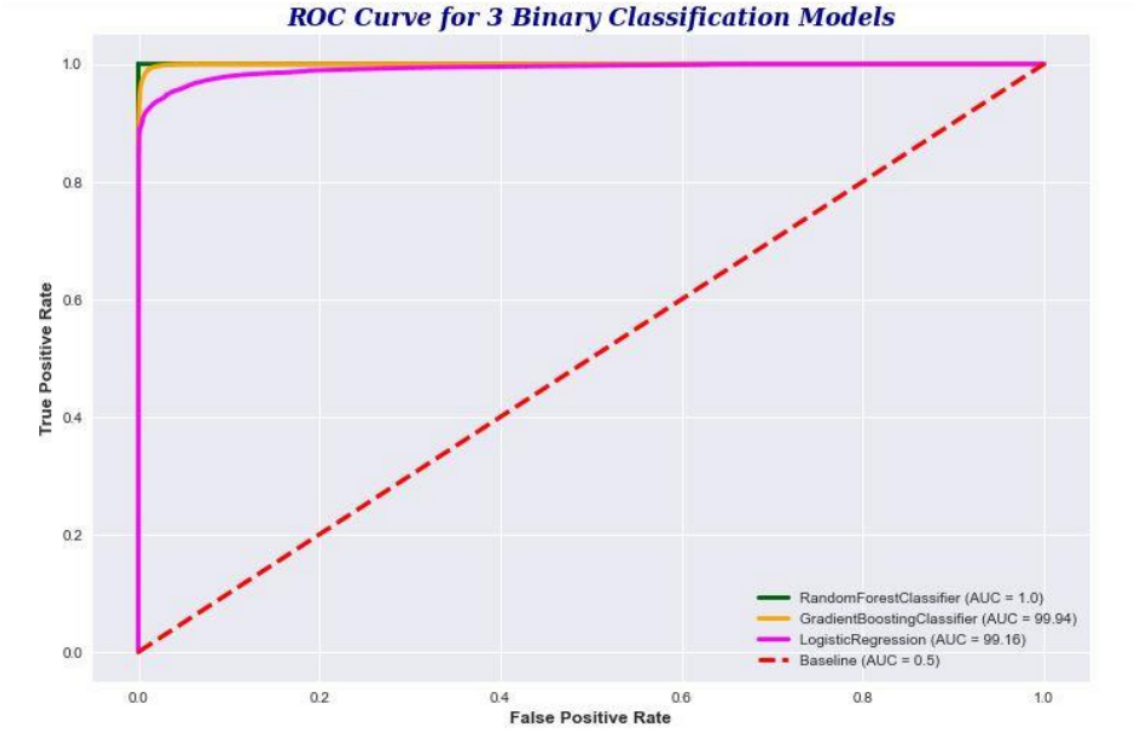

Fig. 6 Visualization with ROC Curve

Fig. 6 explain that the Area Under the Curve (AUC) value has one of the most common matrices for evaluating models, where a value close to the baseline 0.5 is equivalent to randomly guessing whether the transaction is fraudulent or not, and a value close to a predicted 1 is suggestive of a high performance model. In other words, as a general rule a good binary classification model will be as far as possible from the baseline model towards the upper left corner with the vertical line angle length and the horizontal line true positive rate reaching 100 or 1.0. The baseline value (AUC) which is in Figure 0.5 if the accuracy of the test produced by the model is less than 0.5, it illustrates that the effectiveness of the algorithm being tested is not good. From the ROC curve above, it can be seen that the Random Forest Classifier (RFC) and Gradient Boosting Classifier (GBC) algorithms have the highest AUC values while Logistic Regression (LGR) has poor performance, even though it is already above the baseline (AUC) value of all data-evaluation and data-validation of the primary data tested on the credit card dataset. With the results of the percentage evaluation of the AUC score as the result of model testing, the Random Forest Classifier (RFC) algorithm produces an AUC score of 0.9999, the Gradient Boosting Classifier (GBC) produces an AUC score of 0.9994 and Logistic Regression (LGR) 0.9916. Thus making the Random Forest Classifier (RFC) algorithm the best model to accurately predict differences between classes.

IV. CONCLUSION

The conclusions obtained in this study are the results of a comparison of the classification of three algorithms (Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC)) to classify credit card fraud datasets. The data used are 284,807, the dataset is unbalanced. So data balancing is carried out by applying a sample to the classification model created. The test results show that the Random Forest Classifier (RFC) algorithm produces the highest accuracy value, with the data-train percentage of 100% and data-test of 99.99% and the evaluation of the AUC score as the result of algorithm testing is 0.9999, Logistic Regression (LGR) with a data-train percentage of 95.86% and a data-test of 95.77% and the evaluation of the AUC score as a result of algorithm testing is 0.9916 while the Gradient Boosting Classifier (GBC) produces an accuracy value with a data-train percentage of 98.86% and data-test of 98.83% and the evaluation of the AUC score as a result of testing the algorithm is 0.9994. From the accuracy values mentioned, it can be seen that the three algorithms are not over fit, which is an indication that the model is performing well. Thus, in this case the researcher's goal is to have a model defined to beat the baseline model accuracy of 99.80% on previously unseen data. So, by comparing the test accuracy values of each model with the accuracy of the baseline model, it can be observed that the Random Forest Classifier (RFC) algorithm has a higher accuracy value than the baseline model which is assumed to predict every transaction to be non-fraudulent. And the Random Forest Classifier (RFC) algorithm seems to perform better on invisible data because it has a higher test accuracy value. Suggestions that can be given from this research for the development of further research by considering this tradeoff, it is expected to develop a model that is effective in filtering electronic transactions by using other algorithms to overcome class-imbalanced data before being applied to the selected classification algorithm. In addition, comparing the Random Forest Classifier (RFC), Logistic Regression (LGR) and Gradient Boosting Classifier (GBC) algorithms can be performed on other types of datasets to strengthen and prove the findings in this study.

REFERENCES

[1] Y. P. Anggodo, W. Cahyaningrum, A. N. Fauziyah, I. L. Khoiriyah, O. Kartikasari, and I. Cholissodin, "Hybrid K-Means Dan Particle Swarm Optimization Untuk," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 2, pp. 104–110, 2017.

[2] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[3] H. Abijono, P. Santoso, and N. L. Anggreini, "Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data," *J. Teknol. Terap. G-Tech*, vol. 4, no. 2, pp. 315–318, 2021, doi: 10.33379/gtech.v4i2.635.

[4] A. Bisri and R. Rachmatika, "Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 309, 2019, doi: 10.22146/jnteti.v8i4.529.

[5] S. Apriliana and L. Agustina, "The Analysis of Fraudulent Financial Reporting Determinant through Fraud Pentagon Approach," *J. Din. Akunt.*, vol. 9, no. 2, pp. 154–165, 2017, doi: 10.15294/jda.v7i1.4036.

[6] S. Sugiyatno and D. Lelono, "Outlier Detection Credit Card Transactions Using Local Outlier Factor Algorithm (LOF)," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 4, p. 409, 2019, doi: 10.22146/ijccs.46561.

[7] M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit Card Fraud Detection Using Random Forest Algorithm," *2019 Proc. 3rd Int. Conf. Comput. Commun. Technol. ICCCT 2019*, vol. 5, no. 2, pp. 149–153, 2019, doi: 10.1109/ICCCT2.2019.8824930.

[8] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *J. Inform.*, vol. 5, no. 2, pp. 175–185, 2018, doi: 10.31311/ji.v5i2.4158.

[9] Y. Yazid and A. Fiananta, "Mendeteksi Kecurangan Pada Transaksi Kartu Kredit Untuk Verifikasi Transaksi Menggunakan Metode Svm," *Indones. J. Appl. Informatics*, vol. 1, no. 2, pp. 61–66, 2017.

[10] L. D. Perwara, F. A. Bachtiar, and Indriati, "Penerapan Algoritma Decision Tree C4 . 5 Untuk Deteksi Fraud Pada Kartu Kredit dengan Oversampling Synthetic Minority Technique (SMOTE)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 8, pp. 2664–2669, 2020.

[11] G. Niveditha, K. Abarna, and G. V. Akshaya, "Credit Card Fraud Detection Using Random Forest Algorithm," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 301–306, 2019, doi: 10.32628/cseit195261.

[12] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *IlmuKomputer.com J. Softw. Eng.*, vol. 1, no. 1, pp. 46–53, 2015.

[13] F. Zamachsari and N. Puspitasari, "Penerapan Deep Learning dalam Deteksi Penipuan Transaksi Keuangan Secara

Elektronik," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 203–212, 2021, doi: 10.29207/resti.v5i2.2952.

[14] F. Mar'i and A. A. Supianto, "Clustering Credit Card Holder Berdasarkan Pembayaran Tagihan Menggunakan Improved K-Means dengan Particle Swarm Optimization," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 6, p. 737, 2018, doi: 10.25126/jtiik.201856858.

[15] M. Y. Sahroni, N. A. Setifani, and D. N. Fitriana, "Analisis perbandingan algoritma Naïve Bayes, k-Nearest Neighbor dan Neural Network untuk permasalahan class-imbalanced data pada kasus credit card fraud dataset," *Teknologi*, vol. 11, no. 2, pp. 69–73, 2021, doi: 10.26594/teknologi.v11i2.2393.

Tugas Akhir

ORIGINALITY REPORT

14%	10%	10%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Sandra L. Ramírez-Mora, Hanna Oktaba, Helena Gómez-Adorno, Gerardo Sierra. "Exploring the communication functions of comments during bug fixing in Open Source Software projects", Information and Software Technology, 2021 Publication	2%
2	journals.sagepub.com Internet Source	1%
3	garuda.kemdikbud.go.id Internet Source	1%
4	publikasi.dinus.ac.id Internet Source	1%
5	export.arxiv.org Internet Source	1%
6	www.ijeat.org Internet Source	1%
7	Submitted to Universitas Pamulang Student Paper	1%

8	repository.ubn.ru.nl Internet Source	1 %
9	Cheng Zhang, William Cantara, Youngmin Jeon, Karin Musier-Forsyth, Nikolaus Grigorieff, Dmitry Lyumkis. "Analysis of Local Variability and Allostery in Macromolecular Assemblies using Cryo-EM and Focused Classification", Cold Spring Harbor Laboratory, 2018 Publication	<1 %
10	Mohammad Farid Naufal, Siti Rochimah. "Software complexity metric-based defect classification using FARM with preprocessing step CFS and SMOTE a preliminary study", 2015 International Conference on Information Technology Systems and Innovation (ICITSI), 2015 Publication	<1 %
11	baixardoc.com Internet Source	<1 %
12	arxiv.org Internet Source	<1 %
13	Utomo Pujianto, Agusta Rakhmat Taufani, Luis Devvi Ratna Kus Anggraini, Deni Sutaji. "Comparative Analysis of Bagging and Boosting Algorithms on the Classification of the Popularity of Educational-themed Youtube	<1 %

	Videos", 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2021 Publication	
14	Submitted to Beirut Arab University Student Paper	<1 %
15	tunasbangsa.ac.id Internet Source	<1 %
16	www.jcreview.com Internet Source	<1 %
17	www.mdpi.com Internet Source	<1 %
18	Submitted to Aston University Student Paper	<1 %
19	Travis R Goodwin, Dina Demner-Fushman. "A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision", Journal of the American Medical Informatics Association, 2020 Publication	<1 %
20	Pijush Dutta, Shobhandeb Paul, Madhurima Majumder. "An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS", Research Square Platform LLC, 2021 Publication	<1 %

21	norma.ncirl.ie Internet Source	<1%
22	Submitted to University of Derby Student Paper	<1%
23	doaj.org Internet Source	<1%
24	jurnal.ugm.ac.id Internet Source	<1%
25	www.coursehero.com Internet Source	<1%
26	Ananto Setyo Wicaksono, Ahmad Afif. "Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction", International Journal of Advanced Computer Science and Applications, 2018 Publication	<1%
27	Saad Hikmat Haji, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Falah Y. H. Ahmed, Dilovan Asaad Zebari. "The Impact of Different Data Mining Classification Techniques in Different Datasets", 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA), 2021 Publication	<1%

28	Shengjie Min, Guangchun Luo, Zhan Gao, Jing Peng, Ke Qin. "Resonance - An Intelligence Analysis Framework for Social Connection Inference via Mining Co-Occurrence Patterns Over Multiplex Trajectories", IEEE Access, 2020 Publication	<1 %
29	Sulthan Rafif, Pramana Yoga Saputra, Moch Zawaruddin Abdullah. "Classification of Trends in Lecturer Research Fields Using Naive Bayes Method", 2021 International Conference on Electrical and Information Technology (IEIT), 2021 Publication	<1 %
30	epdf.pub Internet Source	<1 %
31	injoit.org Internet Source	<1 %
32	repositorio.unifei.edu.br Internet Source	<1 %
33	Desy Ika Puspitasari, Al Fath Riza Kholdani, Adani Dharmawati, Muhammad Edya Rosadi, Windha Mega Pradnya Dhuhita. "Stroke Disease Analysis and Classification Using Decision Tree and Random Forest Methods", 2021 Sixth International Conference on Informatics and Computing (ICIC), 2021	<1 %

Publication

34	Frits Gerit John Rupilele, Irwan Soulis, Aram Palilu, Abdurrozzaq Hasibuan et al. "Management Information System for Monitoring and Inspection of the Implementation of Universities", International Journal of Engineering & Technology, 2018 Publication	<1 %
35	Warren A Cheung, BF Francis Ouellette, Wyeth W Wasserman. "Compensating for literature annotation bias when predicting novel drug-disease relationships through Medical Subject Heading Over-representation Profile (MeSHOP) similarity", BMC Medical Genomics, 2013 Publication	<1 %
36	www.testmagzine.biz Internet Source	<1 %
37	Fendy Yulianto, Rio Arifando, Ahmad Afif Supianto. "Improved Eliminate Particle Swarm Optimization on Support Vector Machine for Freshwater Fish Classification", 2019 International Conference on Sustainable Information Engineering and Technology (SIET), 2019 Publication	<1 %
38	Submitted to Universitas Nasional Student Paper	<1 %

39 mafiadoc.com
Internet Source

<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		