BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1785–1794
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1785-1794

CLASSIFICATION OF STUDENT GRADUATION STATUS USING *XGBOOST* ALGORITHM

Maria Welita Dwinanda [1], **Neva Satyahadewi** [2*], **Wirda Andani** [3]

[1,2,3] *Statistics Study Program, Faculty of Mathematics and Natural Sciences, Tanjungpura University*
Prof. Dr. H. Hadari Nawawi St, Pontianak, 78124, Indonesia

*Corresponding author's e-mail: ** ***neva.satya@math.untan.ac.id***

ABSTRACT

Article History:

Received: 01st June 2023
Revised: 19th August 2023
Accepted: 29th August 2023

Keywords:

Classification;
Student graduation status;
XGBoost

College is an optional final stage in formal education. At this time, universities are required to have good quality by utilizing all the resources they have. Therefore, efforts are needed to help the faculty and study program make policies and decisions. One of the efforts that can be made is to classify student graduation status as early as possible to increase the number of students graduating on time. Thus, a classification algorithm is needed to avoid overfitting and produce good accuracy. The purpose of this study was to classified the student graduation status of the Statistics Untan Study Program using the XGBoost algorithm. XGBoost is an ensemble algorithm obtained through the development of gradient boosting. XGBoost has several features that can be used to prevent overfitting, but it can only process numerical data. Therefore, 140 numerical data were adjusted using the dummy technique in this study. The resulting XGBoost classification model is optimal at the number of rounds is 3 and the number of folds is 5. Based on the performance evaluation of the XGBoost algorithm, an accuracy of 75,00%, precision of 88,89%, sensitivity of 76,19% and specificity of 71,43% were obtained. Thus, the performance of the XGBoost algorithm is classified as good.

How to cite this article:

Dwinanda. M. W, Satyahadewi. N, and Andani. W., "Classification of Student Graduation Status Using XGBoost Algorithm (Case Study: Statistics Study Program of Tanjungpura University)," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1785-1794, September , 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

College is an optional final stage in formal education [1]. At this time, universities are required to have good quality by utilizing all the resources they have. The quality of a university in each study program in Indonesia is measured based on accreditation organized by the National Accreditation Board of Higher Education (BAN-PT). The government formed BAN-PT to conduct and develop college accreditation independently. Accreditation is an external quality assurance system as part of the Higher Education Quality Assurance System, which is carried out based on interactions between standards or criteria in Higher Education Standards [2].

There are nine criteria for assessing study program accreditation, two of which are students and tri-dharma outcomes and achievements. The tri-dharma output criteria include graduates' Grade Point Average (GPA), study period, study success and timely graduation. Thus, a university must maintain the quality of its students by paying attention to the quality of graduates, which can be reviewed through the length of the study period and timely graduation.

Tanjungpura University (Untan), located in Pontianak City, West Kalimantan, is one of the state universities in Indonesia. In Untan, there are nine faculties consisting of 96 study programs. One of the faculties at Untan is the Faculty of Mathematics and Natural Sciences (FMIPA) which houses the Statistics Study Program. In paying attention to the quality of its students, the academic department of FMIPA Untan evaluates the success of student studies at the end of each semester, the end of the first four semesters, the end of eight semesters and the end of the study program. Student graduation is said to be on time if the length of study is less than or equal to four years and a minimum learning load of 144 credits [3].

In reality, students of the Statistics Study Program of FMIPA Untan with on-time graduation status are fewer than students with untimely graduation status. The number of students with the status of graduating on time is 56, while the number of students with the status of not graduating on time is 84. This is a problem for the accreditation assessment of the Statistics Study Program. Therefore, efforts are needed to help the faculty and study program make policies and decisions. Efforts can be made to classify student graduation status as early as possible to increase the number of students graduating on time. Data mining is one method that can be utilized for classification.

Data mining is a technique that combines several sciences, such as machine learning, pattern recognition, statistics and visualization, to obtain information from large databases [4]. The technique in data mining that is often used is classification. Classification is analyzing data to produce models to classify data into specific classes. An algorithm that can be used is eXtreme Gradient Boosting (XGBoost).

XGBoost is an ensemble algorithm obtained through the development of gradient boosting. XGBoost has several features that can be used to prevent overfitting. XGBoost consists of multiple trees that produce one final tree with the best results. The main reason behind the success of the XGBoost algorithm is its ability to adapt to various situations (flexible) due to improvements from previous calculations [5]. Several studies using the XGBoost algorithm have been conducted, as in references [5] [6] [7] [8]. Meanwhile, reference [9] research predicting student graduation status. This research differs from other research that has been done, namely the development of dummy techniques in adjusting the form of numerical data, and there is no need to select features in it.

In this study, the classification of the graduation status of students of the Untan Statistics Study Program was carried out and evaluated the performance of the XGBoost algorithm was using the confusion matrix. The performance evaluation of the XGBoost algorithm includes accuracy, precision, sensitivity and specificity.

2. RESEARCH METHODS

2.1 Classification

Classification is a data mining method that machines use to sort or categorize objects based on specific characteristics, such as humans trying to distinguish one object from another [10]. According to [11], classification is finding a model that can explain and distinguish data classes based on their classification structure used in categorical labels. There are four components in classification as follows:

1. Class

Class is categorical data representing the label contained in the object after classification.

2. Attributes

   Attributes are independent variables of the model to be classified.

3. Training Data

   Training data is a collection of data used to train the model in finding the appropriate class based on the appropriate classification.

4. Test Data

   Test data is a set of new data that will be classified by the previously formed model so that the classification algorithm can be evaluated.

2.2 Binary Sigmoid Function

In classification trees with binary classes, the probability value on the output leaf can be updated using a sigmoid function. The sigmoid function is a mathematical function with an interval between 0 and 1. If the value is more than 0,50, the instance is classified as class 1. If the sigmoid value is less than 0,50, the instance is classified as class 0 [12].

2.3 eXtreme Gradient Boosting

Gradient Boosting, also known as eXtreme Gradient Boosting, was first proposed by Dr. Tianqi Chen of the University of Washington in 2014 [13]. XGBoost is an improvement of the Gradient boosting algorithm that can build classification trees efficiently and operate in parallel [14]. In XGBoost, the new model predicts the residuals of the previous model and then sums them up to obtain the final prediction. The XGBoost model starts with a leaf, with the initial leaf filled with probability values of the predicted attributes [15]. The steps in building an XGBoost tree are as follows:

a. Initialize the initial probability of prediction (Pr_i^1), with $i = 1,2, \dots, n$.

b. Calculate the residuals with the following formula:

$$Residual_i^t = Y_i - Pr_i^t \tag{1}$$

c. Calculate the cover value of the attribute with the following formula:

$$Cover(A) = \sum_{i=1}^{n} (Pr_i^t(1 - Pr_i^t)) \tag{2}$$

d. Calculate the similarity score (SS) with the following formula:

$$SS_{node} = \frac{(\sum_{i=1}^{n} Residual_i)^2}{\sum_{i=1}^{n}(Pr_i^t(1 - Pr_i^t)) + \lambda} \tag{3}$$

e. Calculate the attribute gain value with the following formula:

$$Gain(A) = SS_{left} + SS_{right} - SS_{root} \tag{4}$$

f. Calculate the leaf output value with the following formula:

$$Output(A)_i = \frac{\sum_{i=1}^{n} Residual_i}{\sum_{i=1}^{n}(Pr_i(1 - Pr_i)) + \lambda} \tag{5}$$

g. Calculate the log odds value as follows:

$$Log\ odds_i^t = log\left(\frac{Pr_i^t}{1 - Pr_i^t}\right) \tag{6}$$

h. Update the probability value to be normalized with the following formula:

$$Pr_i^{t+1} = \log\ odds_i^t + (\eta \times output(A)_i) \tag{7}$$

i. Normalize the probability value using the binary sigmoid function as follows:

$$Sigmoid(Pr_i^{t+1}) = \frac{exp^{Pr_i^{t+1}}}{1 + exp^{Pr_i^{t+1}}} \tag{8}$$

j. Repeating step b up to i.

k. Evaluate the performance of the classification algorithm.

2.4 Classification Algorithm Performance Evaluation

Confusion Matrix (CM) is one of the methods used to evaluate the accuracy of the model that has been formed. CM contains comparison information between the model output and actual classification results. The four components that represent the classification results in CM are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), which can be seen in **Table 1 [16]**.

Table 1. Confusion Matrix

Confusion Matrix		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positive (TP) Correct result	False Positive (FP) Unexpected result
	Negative	False Negative (FN) Missing result	True Negative (TN) Correct absence of result

According to **Table 1**, TP is a positive instance that is true, TN is a negative instance that is true, FP is a negative instance that is classified as positive, and FN is a positive instance that is classified as a negative instance. The formulation for calculating the accuracy, precision, sensitivity, and specificity values in the performance evaluation of classification algorithms is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{12}$$

2.5 Research Data

The data used in this study are primary and secondary. Primary data is obtained through questionnaires distributed to alums of Untan Statistics Study Program students. In contrast, secondary data is data on student graduation from Untan Statistics Study Program Period I Academic Year 2017/2018 to Period II Academic Year 2022/2023 obtained through academic FMIPA Untan and PDDIKTI Untan sites. The amount of data used amounted to 140 instances.

2.6 Research Attributes

This study used 11 attributes consisting of ten independent attributes and one dependent attribute. An explanation of the research attributes can be seen in **Table 2**.

Table 2. Research Attributes

Attribute	Name	Data Type	Category
X_1	Grade Point Average 1st Semester	Numeric	-
X_2	Grade Point Average 2nd Semester	Numeric	-
X_3	Grade Point Average 3rd Semester	Numeric	-
X_4	Grade Point Average 4th Semester	Numeric	-
X_5	Region of Origin Domicile	Categorical	0. District 1. City
X_6	Gender	Categorical	0. Male 1. Female
X_7	High School Accreditation	Categorical	0. < A 1. A

X_8	Entry Path	Categorical	0. Other than SNMPTN 1. SNMPTN
X_9	Scholarship Ownership	Categorical	0. No 1. Yes
X_{10}	Passing Status of First TUTEP Test	Categorical	0. Not Passed 1. Passed
Y	Student Graduation Status	Categorical	0. Not on Time 1. On Time

Based on the independent attributes in **Table 2**, it can be seen that there are six attributes with categorical data types. In this research with the XGBoost algorithm, it is necessary to transform categorical data into numerical data because the XGBoost algorithm can process numeric data only.

3. RESULTS AND DISCUSSION

3.1 Division of Training Data and Test Data

In this study, the dataset was divided into two parts, namely training data and test data. Data division is used to avoid overfitting. Overfitting is a condition when all data that has gone through the training process achieves excellent accuracy, but there is a mismatch in the test data prediction process. The algorithm is taught using training data when developing models, and its performance is evaluated using test data. The results of the dataset division are presented in **Table 3**.

Table 3. Dataset Distribution

Number of Data	Training Data	Test Data
140 (100%)	112 (80%)	28 (20%)

The division of the dataset is done by randomizing the entire data and then splitting the data with the proportion of training data:test data of 80:20. After obtaining the results of dividing the dataset into training data and test data, based on **Table 3**, the XGBoost model was formed using 112 training data.

3.2 Implementation of XGBoost Algorithm

3.2.1 XGBoost Tree Formation

After obtaining the training data, an XGBoost tree can be formed using the training data, with the manual calculation described as follows:

a. The initial prediction probability $\left(Pr_i^1\right)$ in this study is initialized at 0,50 because the classified attributes have two classes.

$$Pr_i^1 = 0{,}50$$

b. Using **Equation (1)**, the residuals, for instance one, are calculated:

$$Residual_1^1 = 0{,}00 - 0{,}50 = -0{,}50$$

c. The cover value of Grade Point Average 4th Semester (X_4) attribute is calculated using **Equation (2)**:

1. The cover value of IPS 4 < 3,18

$$Cover(X_4) = \big(0{,}50(1{,}00 - 0{,}50)\big) + \cdots + \big(0{,}50(1{,}00 - 0{,}50)\big) = 14{,}00$$

2. The cover value of IPS 4 ≥ 3,18

$$Cover(X_4) = \big(0{,}50(1{,}00 - 0{,}50)\big) + \cdots + \big(0{,}50(1{,}00 - 0{,}50)\big) = 14{,}00$$

d. After calculating the cover value, the similarity score (SS) is calculated for the Grade Point Average 4th Semester (X_4) attribute node using **Equation (3)**:

1. SS value for IPS 4 < 3,18

$$SS_{left} = \frac{(-0{,}50 - 0{,}50 - \cdots - 0{,}50)^2}{\big(0{,}50(1{,}00 - 0{,}50)\big) + \cdots + \big(0{,}50(1{,}00 - 0{,}50)\big)} = 34{,}57$$

2. SS value for IPS 4 ≥ 3,18

$$SS_{right} = \frac{(0{,}50 + 0{,}50 - \cdots - 0{,}50)^2}{\big(0{,}50(1{,}00 - 0{,}50)\big) + \cdots + \big(0{,}50(1{,}00 - 0{,}50)\big)} = 16{,}07$$

e. The gain value of the Grade Point Average 4th Semester (X_4) attribute is calculated using **Equation (4)**:

$$Gain(X_4) = 34,57 + 16,07 - 1,75 = 48,89$$

f. After obtaining the gain value, the calculation of the Grade Point Average 4th Semester (X_4) attribute output of each instance is carried out using the formula in **Equation (5)**:

$$Output(X_4)_i = \frac{(-0,50 - 0,50 - \cdots - 0,50)}{\big(0,50(1,00 - 0,50)\big) + \cdots + \big(0,50(1,00 - 0,50)\big)} = -1,74$$

g. Using **Equation (6)**, the log odds value is calculated with instance one:

$$Log\ odds_1^2 = log\left(\frac{Pr_1^1}{1 - Pr_1^1}\right) = log\left(\frac{0,50}{1,00 - 0,50}\right)$$

$$Log\ odds_1 = log1 = 0,00$$

h. The updated probability value will be normalized **Equation (7),** for instance one as follows:

$$Pr_1^2 = 0,00 + (0,30 \times (-1,74)) = -0,52$$

i. After obtaining the updated probability value, the probability normalization is performed using the sigmoid function in **Equation (8)**:

$$Sigmoid(Pr_1^2) = \frac{exp^{Pr_1^2}}{1 + exp^{Pr_1^2}} = \frac{exp^{-0,52}}{1 + exp^{-0,52}}$$

$$Sigmoid(Pr_1^2) = 0,37$$

After obtaining the gain value, leaf output and classification results using training data, the XGBoost model can be formed, as shown in **Figure 1**.

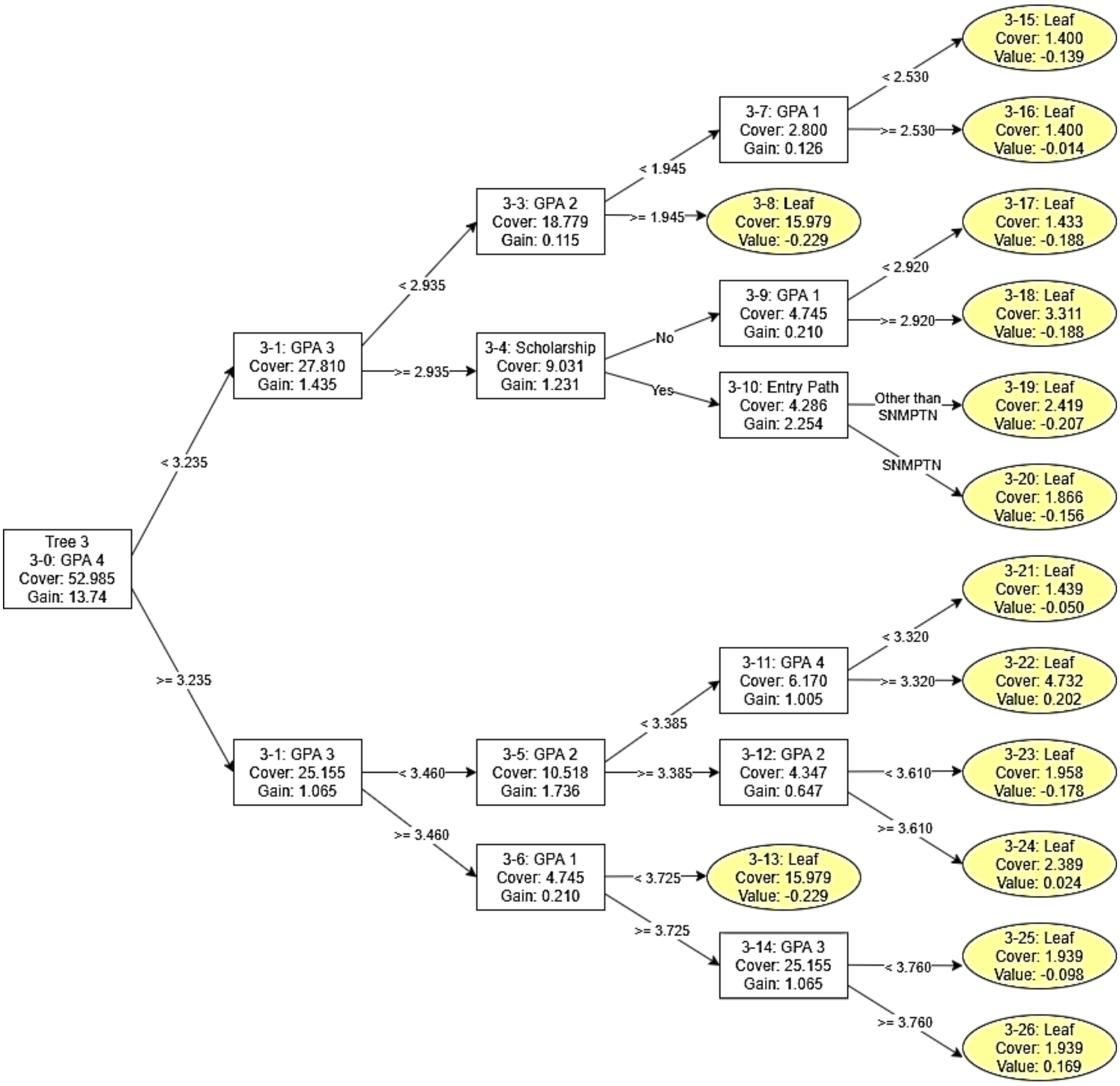

Figure 1. XGBoost Tree Model

Based on the XGBoost model in **Figure 1**, it can be seen that the initial leaf (root node) in the classification results of student graduation status is Grade Point Average 4th Semester and produces 14 leaf outputs. The leaf output has a value representing the leaf's probability value in categorizing positive classes and negative classes. Of the 14 output leaves produced, nine contribute to categorizing negative classes, and five others contribute to categorizing positive classes.

3.2.2 XGBoost Model Validation

The next step in applying the XGBoost algorithm is to validate the XGBoost model that has been formed. Several experiments were carried out on the combination of the number of rounds and folds so that accuracy, precision, sensitivity and specificity were obtained in **Table 4**.

Table 4. XGBoost Model Validation

Dataset Distibution	Round	Fold	Accuracy	Precision	Sensitivity	Specificity
Training	1	5	94.64%	91.30%	100.00%	87.76%
Test	1	5	85.71%	94.74%	85.71%	85.71%
Training	2	5	97.32%	95.45%	100.00%	93.88%
Test	2	5	78.57%	89.47%	80.95%	71.43%
Training	3	5	99.11%	98.44%	100.00%	97.96%
Test	3	5	75.00%	88.89%	76.19%	71.43%
Training	4	5	99.11%	98.44%	100.00%	97.96%
Test	4	5	75.00%	88.89%	76.19%	71.43%
Training	5	5	99.11%	98.44%	100.00%	97.96%
Test	5	5	75.00%	88.89%	76.19%	71.43%

Table 4 shows that the resulting accuracy is stable when the number of rounds is three times. Hence, the XGBoost model in **Figure 1** is optimal and can be used to classify the graduation status of Untan Statistics Study Program students.

3.3 XGBoost Classification

After obtaining the XGBoost model with the optimum number of rounds and folds, classification can be carried out using test data with the classification results in **Table 5**.

Table 5. Classification Results of XGBoost Algorithm

Instance	Status	Classification
1	Not on Time	Not on Time
2	Not on Time	On Time
3	Not on Time	Not on Time
4	Not on Time	Not on Time
⋮	⋮	⋮
25	Not on Time	Not on Time
26	Not on Time	On Time
27	Not on Time	Not on Time
28	On Time	On Time

Based on **Table 5**, the classification results were obtained with the status of student graduation, not on time in as many as 21 instances and the status of student graduation on time in as many as 7 instances. Through these classification results, the performance of the XGBoost algorithm can be evaluated.

3.4 Confusion Matrix

From the classification results obtained, a confusion matrix can be formed to evaluate the performance of the XGBoost algorithm, which can be seen in **Table 6**.

Table 6. Confusion Matrix of XGBoost Algorithm

Confusion Matrix		Actual	
		Not on Time	On Time
Predicted	Not on Time	16	2
	On Time	5	5

Based on the confusion matrix in **Table 6**, the accuracy, precision, sensitivity and specificity values in this study are calculated based on **Equation (9)** to **Equation (12)** as follows:

$$Accuracy = \frac{16+5}{(16+5+2+5)} \times 100\% = 75{,}00\%$$

$$Precision = \frac{16}{(16+2)} \times 100\% = 88{,}89\%$$

$$Sensitivity = \frac{16}{(16+5)} \times 100\% = 76{,}19\%$$

$$Specificity = \frac{5}{(5+2)} \times 100\% = 71{,}43\%$$

The accuracy, precision, sensitivity, and specificity values in this study represent the performance used to assess the benchmark for the success of the XGBoost algorithm in classifying the graduation status of Untan Statistics Study Program students. The accuracy obtained in this study was 75.00%. Based on the results of this study, the goal of the XGBoost algorithm has been achieved, namely, preventing overfitting. This finding can be supported through the validation results of the XGBoost model for training data and test data in **Table 4**, which shows that the XGBoost model is optimal and there are no discrepancies in the accuracy values of training data and test data. As in other studies, the accuracy value of the XGBoost model of 75.00% is relatively good.

4. CONCLUSIONS

In this study, the XGBoost algorithm is used, which is able to classify the graduation status of students of the Statistics Study Program FMIPA Untan in the Not on Time and On Time classes. When the model-building process is carried out, the third tree XGBoost model is formed with the optimum number of rounds and folds of 3 and 5, respectively. By evaluating the performance of the XGBoost algorithm, an accuracy value of 75,00%, precision of 88,89%, sensitivity of 76,19% and specificity of 71,43% is obtained. Therefore, the performance of the XGBoost algorithm is relatively good.

REFERENCES

[1] N. Hasanah, F. Syahfitri and T. Pujahadi, “Sosialisasi Tentang Pentingnya Pendidikan Tingkat Perguruan Tinggi Kepada Masyarakat Desa Jaring Halus,” *Jurnal Pengabdian Kepada Masyarakat,* pp. 23-29, 2021.

[2] MA BAN-PT, “Kebijakan Instumen Akreditasi BAN-PT dan LAM Berbasis SN Dikti,” BAN-PT, 2019.

[3] E. Haryatmi and S. P. Hervianti, “Penerapan Algoritma Support Vector Machine untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu,” *Jurnal RESTI (Rekayasa SIstem dan Teknologi Informasi,* vol. 5, no. 2, pp. 386-392, 2021.

[4] Y. Mardi, “Klasifikasi Menggunakan Algoritma C4.5,” *Jurnal Edik Informatika Penelitian Bidang Komputer Sains dan Pendidikan Informatika,* vol. 2, no. 2, pp. 213-219, 2017.

[5] S. E. H. Yulianti, O. Soesanto and Y. Sukmawaty, “Penerapan Metode Extreme Gradient Boosting (XGBoost) pada Klasifikasi Nasabah Kartu Kredit,” *Journal of Mathematics: Theory and Applications,* pp. 21-26, 2022.

[6] A. A. Firdaus and A. K. Mutaqin, “Klasifikasi Pemegang Polis Menggunakan Metode XGBoost,” *Prosiding Statistika,* pp. 704-710, 2021.

[7] M. R. Givari, M. R. Sulaeman and U. Y, “Perbandingan Algoritma SVM, Random Forest dan XGBoost untuk Penentuan Persetujuan Pengajuan Kredit,” *Nuansa Informatika,* vol. 16, no. 1, pp. 141-149, 2022.

[8] M. K. Nasution, R. R. Saedudin and V. P. Widartha, “Perbandingan Akurasi Algoritma Naive Bayes dan Algoritma XGBoost pada Klasifikasi Penyakit Diabetes,” *eProceedings of Engineering,* vol. 8, no. 5, 2021.

[9] T. H. Hasibuan and D. Mahdiana, “Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma C4.5 pada UIN Syarif Hidayattullah Jakarta,” *SKANIKA: Sistem Komputer dan Teknik Informatika,* vol. 6, no. 1, pp. 61-74, 2023.

[10] A. A. Hania, “Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning,” *Jurnal Teknologi Indonesia,* vol. 1, pp. 1-6, 2017.

[11] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques Third Edition, University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University, 2012.

[12] X. Li and Z. Li, “A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm,” *Ingenierie des Systemes d'Information,* vol. 24, no. 5, pp. 525-530, 2019.

[13] D. Zhang and Y. Gong, “The Comparison of LightGBM and XGBoost Coupling Factor Analysis and Prediagnosis of Acute Liver Failure,” *IEEE Access*, vol. 8, pp. 220990-221003, 2020.

[14] T. Chen and C. Guestrin, “Xgboost: A Scalable Tree Boosting System,” *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794, 2016.

[15] S. F. N. Islam, A. Sholahuddin and A. S. Abdullah, “Extreme Gradient Boosting (XGBoost) Method in Making Forecasting Application and Analysis of USD Exchange Rates Against Rupiah,” *Journal of Physics: Conference Series,* vol. 1722, no. 1, p. 012016, 2021.

[16] F. Y. Manik and K. S. Saragih, “Klasifikasi Belimbing Menggunakan Naive Bayes Berdasarkan Fitur Warna RGB,” *IJCCS: Indonesian Journal of Computing and Cybernetics Systems*, vol. 11, no. 1, pp. 99-108, 2017.