pIJmath

Pattimura International Journal of Mathematics (PIJMath)

Volume 02 issue 01, pages 07-16, May 2023
E-ISSN: 2830-6791

RESEARCH ARTICLE ◙ *OPEN ACCESS*

DOI: https://doi.org/10.30598/pijmathvol2iss1pp07-16

Comparison of Adaboost Application to C4.5 and C5.0 Algorithms in Student Graduation Classification

Yuveinsiana Crismayella[1], Neva Satyahadewi[2*], Hendra Perdana[3]

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

Corresponding author's e-mail: [2] neva.satya@math.untan.ac.id*

ABSTRACT

Article History

Received : 16th February 2023
Accepted : 5th April 2023
Published : 10th April 2023

Keywords

Study period;
C4.5 Algorithm;
C5.0 Algorithm;
Adaptive Boosting;

Students become a benchmark used to assess quality and evaluate college learning plans. Therefore, students who graduate not on time can have an effect on accreditation assessment. The characteristics of students who graduate on time or not on time in determining student graduation can be analyzed using classification techniques in data mining, namely the C4.5 and C5.0 algorithms. The purpose of this study is to compare the application of the Adaboost Algorithm to the C4.5 and C5.0 Algorithms in the classification of student graduation. The data used is the graduation data of students of the Statistics Study Program at Tanjungpura University Period I of the 2017/2018 Academic Year to Period II of the 2022/2023 Academic Year. The analysis begins by calculating the entropy, gain and gain ratio values. After that, each data was given the same initial weight and iterated 100 times. Based on the classification results using the C5.0 Algorithm, the attribute that has the highest gain ratio value is school accreditation, meaning that the school accreditation attribute has the most influence in the classification of student graduation. The application of the Adaboost Algorithm to the C5.0 Algorithm is better than the C4.5 Algorithm in classifying the graduation of students of the Untan Statistics Study Program. The Adaboost algorithm was able to increase the accuracy of the C5.0 Algorithm by 12.14%. While in the C4.5 Algorithm, the Adaboost Algorithm increases accuracy by 10.71%.

[1]***How to cite this article:***
Y. Crismayella, N. Satyahadewi, and H. Perdana, "COMPARISON OF ADABOOST APPLICATION TO C4.5 AND C5.0 ALGORITHMS IN STUDENT GRADUATION CLASSIFICATION", *Pattimura Int. J. Math. (PIJMATH).*, vol. 02, iss. 01, pp. 07-16, May 2023.

1. Introduction

Universities are providers of academic education for students [1]. The quality of the college can be determined from the percentage of the student's learning completion rate during his studies. Students become a benchmark used to assess quality and evaluate college learning plans. Regulation of the Minister of Research and Technology of Higher Education Number 44 of 2015 states that the maximum standard learning process in the study period is 7 years, the GPA is above 2.0, the minimum credit is 144 for the Bachelor program [2].

Tanjungpura University (Untan) is one of the public universities located in Pontianak City, West Kalimantan. One of the study programs at the Faculty of Mathematics and Natural Sciences (FMIPA) is the Statistics Study Program. Students of the S1 (Bachelor) program of FMIPA Untan are said to graduate on time if they can complete studies less than or equal to four years with a minimum study load of 144 credits. Students who have the potential to graduate untimely can be analyzed on the evaluation of student success at the end of the first four semesters. In the Statistics Study Program, there are still many students who do not graduate on time so that it will affect the accreditation assessment. Therefore, the characteristics of students who graduate on time or not on time in determining student graduation can be analyzed using techniques in data mining.

Data mining is the process of finding information or patterns using statistical techniques obtained by extracting and identifying useful and interesting patterns from various large databases. Classification is a method that is often used in data mining. The classification method that is widely used by researchers is the decision tree. The C4.5 and C5.0 algorithms are classification algorithms in data mining that are used to build decision trees. Decision trees can solve neural network problems of handling over-fitting, handling continuous attributes, choosing the right ones for attribute selection, handling training data with missing attribute values, and improving computational efficiency [3]. But the Decision tree method has a drawback in the high degree of class imbalance.

The distribution of the imbalance class can be marked as something that has more cases than some other classes. The balance problem is one where one class is represented by a large sample, while the other is represented by only a few samples [3]. Class imbalances affect classification performance, so a method is needed that can overcome it so that it can achieve and improve better decision tree classification performance. Adaptive Boosting (Adaboost) is one of the supervised algorithms in data mining that is used to build classification models and can handle class imbalances. The Boosting algorithm gives weight to the distribution of training data in each iteration with different values. Therefore, research was conducted on the comparison of the application of the Adaboost Algorithm to the C4.5 and C5.0 Algorithms in the graduation classification of Untan Statistics Study Program students. The purpose of this study is to compare the application of the Adaboost Algorithm to the C4.5 and C5.0 Algorithms.

2. Research Methods

2.1 Data Mining

Data mining is the process of finding patterns or information using certain techniques or methods in a data. The techniques and methods in data mining are very varied. The selection of the right method or algorithm greatly affects the overall purpose and process of Knowledge Discovery in Database (KDD) [4]. KDD is the process of determining useful information contained in the data [5]. Data mining uses several techniques with the aim of obtaining knowledge and information related to large databases. Data mining is a series of processes that are divided into seven stages [6] namely data cleaning or data cleaning, data integration, data selection, data transformation or data transformation, data mining or data mining, pattern evaluation or pattern evaluation, knowledge presentation or knowledge presentation. According to [7] there are six data mining functions, namely description, estimation, prediction, classification, grouping, and association.

2.2 Classification

The heading at the third level follows the style of the second level heading. Avoid using headings more than three levels Classification is a method in data mining used to find models that describe data. The purpose of classification models in data mining is as descriptive modeling that can explain the differences between objects with different classes and predictive modeling that can predict a class label whose record is not yet known.

2.3 Decision tree

Decision tree is one of the popular classification methods because it can be easily interpreted [8]. A decision tree is a flowchart like a tree where each node shows a test on an attribute, the results are represented by each test branch, and the classes are represented by leaf nodes. Decision trees are used to explore data that has passed the preprocessing stage and find a hidden model of data with a target variable, so that it can be used to divide large data sets into smaller record sets with regard to their destination variables [9].

2.2 C4.5 algorithm

The C4.5 algorithm is used to build a decision tree that is easy to understand, flexible, and interesting because it can be visualized in the form of images [10]. The C4.5 algorithm was introduced by Quinlan in 1996 as an improved version of the ID3 [11]. The improvements are the ability to handle features with numerical types, overcome missing values, and the ability to prunning decision trees.

The root attribute is selected based on the highest gain value of the existing attributes. Here is the formula used to calculate the gain value:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropy(S) \tag{1}$$

with $|A|$: the sum of the entire sample in the dataset, A : independent variables, $|S_i|$: number of samples for category *i-th* and : the number of categories on the independent variable A. Meanwhile, the calculation of the entropy value is presented in the following **Equation (2)**:

$$Entropy(S) = -\sum_{i=1}^{N} P_i log_2 P_i \tag{2}$$

with S: dataset, N: the number of classes on the dependent variable, and P_i: Proportion of the number of i-th class data on the dataset.

2.3 C5.0 algorithm

Decision trees are used to explore data that has passed the preprocessing stage. C5.0 is an algorithm in data mining that is used to build decision trees. C5.0 is an improvement over the ID3 and C4.5 algorithms. C5.0 provides the best accuracy rate and less execution time compared to other classification algorithms [12]. The algorithm begins with all the data being made the root of the decision tree and the selected attribute will be the divisor for that sample. To calculate the entropy can use equation (2), while to get the value of information gain can use the information gain equation in the C4.5 algorithm. The attribute with the highest gain ratio value will be the root node. Here's the formula used to calculate the gain ratio:

$$Gain\ Ratio = \frac{Gain(S,A)}{SplitInfo(S,A_i)} \tag{3}$$

with $Gain(S,A)$: gain information on independent variables A and $SplitInfo(S,A_i)$: split information on the independent variable A class i-th.

2.4 Adaptive Boosting Algorithm (Adaboost)

Adaptive Boosting (Adaboost) is a variant of the boosting algorithm [13]. The Adaboost algorithm builds a powerful classifier by combining it with a number of simple (weak) classifiers. At the beginning of the classification, each sample on the Adaboost Algorithm is assigned the same weight. After each classification, the weight of the incorrect result increases and the weight of the correct result decreases. This process is repeated until it reaches a threshold or maximum number of cycles [14]. The steps in the Adaboost Algorithm are as follows [12]:

a. The sample N dataset set has two classes labeled y∈{0,1}.
b. In the dataset, the initial weight of each sample is set equally.

$$w_i^1 = \frac{1}{N} \tag{4}$$

With w_i^1: initial weight on the first iteration for all samples and $i = 1,2,\dots,N$.

c. For each iteration t =1, 2, …, T. Where T is the maximum iteration, do the following process:
   1. Find a learner $f(x)$ (predictive model or classifier) of a resampled data set. Learner $f(x)$ applied to the dataset. If the sample is not classified correctly, then "false = 1" and if "true= 0".
   2. The number of weighted errors from all samples is calculated as follows:

$$error^t = \sum_{i=1}^{N} (w_i^t \times error_i^t) \tag{5}$$

   3. The confidence index of the learner $f(x)$ is calculated as follows:

$$\alpha^t = \frac{1}{2}\ln\left(\frac{1-error^t}{error^t}\right) \tag{6}$$

   The trust index of the *learner f(x)* depends on weighted error.

   4. Update weights for all original training samples:

$$w_i^{t+1} = w_i^t \times \begin{cases} \exp(-\alpha^t) & to\ decrease\ weight \\ \exp(\alpha^t) & to\ increase\ weight \end{cases} \quad (7)$$

If the sample is properly verified, its weight will decrease. Whereas if the sample is classified incorrectly, then its weight increases.

5. Weights are normalized as follows:

$$w_i^{t+1} = w_i^t \times \begin{cases} \exp(-\alpha^t) & to\ decrease\ weight \\ \exp(\alpha^t) & to\ increase\ weight \end{cases} \quad (8)$$

So that

$$w_i^{*\,t} = \frac{w_i^t}{\sum_i^N w_i^t} \quad (9)$$

6. If $error < 0{,}5$, or $t < T$, repeat steps 1-5; otherwise, the process stops.
7. Prediction models $f^t(x), t = 1,2, \dots, T$ obtained after T iteration, $t = 1,2, \dots T$.The final prediction for case j is derived from the model's prediction T combined using a voting approach:

$$y_j = sign \sum_{t=1}^{T} \alpha^t f^t(x) \quad (10)$$

2.5 Evaluation of the Classification Model

The performance of the classification model is evaluated using a confusion matrix. Table 1 is a confusion matrix describing true positive (TP), false negative (FN), false positive (FP), and true negative (TN). TP is a positive instance correctly classified as positive, FN is a positive instance classified as negative, FP is a negative instance classified as positive and TN is a negative instance that is correctly classified as negative **[12]**. Some evaluations defined based on the values contained in the confusion matrix are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

Accuracy is a metric that is usually used to evaluate classification results. However, when working with an unbalanced dataset, accuracy alone is not enough because the resulting value is dominated by the majority class, i.e. the negative class **[15]**.

Table 1. Confusion Matrix

		Actual	
		Positive	***Negative***
Predicted	***Positive***	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

3. Results And Discussion

The data used in the study was the graduation data of students of the Untan Statistics Study Program Period I of the 2017/2018 Academic Year to Period II of the 2022/2023 Academic Year. The number of samples used was 140 samples. The dependent attribute used is the student's graduation status (Y), while the independent attribute is presented in Table 2. It is known that students who graduated untimely were 84 people and those who were on time were 56 people. The number of men is 31 people and women are 109 people. The GPA attribute uses a GPA from semester 1 to semester 4 and is categorized into two, namely < 3 and ≥ 3. For the attributes of the Domicile Area of Origin, it is categorized into two, namely the district and city where the alumni came from when they were in college. School Accreditation attributes are categorized into two, namely A and other than A.

Table 2. Classification of dependent and independent attributes

No	Attribute	Category	Total	Percentage
1	Student's graduation status	1. Not On Time	84	60%
		2. On Time	56	40%
2	Sex	1. Male	31	22,1%
		2. Female	109	77,9%
3	GPA of the 1st semester	1. < 3	52	37,1%
		2. ≥ 3	88	62,9%

No	Attribute	Category	Total	Percentage
4	GPA of the 2nd semester	1. < 3	61	43,6%
		2. ≥ 3	79	56,4%
5	GPA of the 3th semester	1. < 3	68	48,6%
		2. ≥ 3	72	51,4%
6	GPA of the 4th semester	1. < 3	65	46,4%
		2. ≥ 3	75	53,6%
7	Domicile Region	1. Regency	100	71,4%
		2. City	40	28,6%
8	School Accreditation	1. A	31	22,1%
		2. Other Than A	109	77,9%
9	University Admission Method	1. SNMPTN	84	60,0%
		2. Other Than SNMPTN	56	40,0%
10	Scholarship	1. No	78	55,7%
		2. Yes	62	44,3%
11	First TUTEP Test Pass Status	1. Graduate	12	8,6%
		2. No	123	87,9%

3.1. Formation of C4.5 Algorithm and C5.0 Algorithm

Table 3 presents the results of calculating the entropy, gain, and gain ratio values of the ten attributes used. Based on **Table 3** of the ten independent attributes used in this study, the attribute that has the highest gain ratio value is the school accreditation attribute, which is 0.04553, which means that the school accreditation attribute has the most influence in the classification of student graduation. While the attributes GPA of the 1st semester (X_1), GPA of the 2nd semester (X_3), GPA of the 3th semester (X_4), GPA of the 4th semester (X_5), dan University Admission Method (X_8) does not greatly affect the predicted results in the classification. In forming the C4.5 and C5.0 Algorithm, it is necessary to calculate the entropy, gain, and gain ratio using **Equations (2), (1),** and **(3)**.

Table 3. Calculation of entropy, gain, and gain ratio values

Attribute	Category	*Entropy*	*Gain*	*Gain Ratio*
Student's graduation status	1. Male	0,96124	0,00042	0,00022
	2. Female	0,97602		
Sex	1. < 3	0,95593	0,00122	0,00063
	2. ≥ 3	0,98038		
GPA of the 1st semester	1. < 3	0,92312	0,01087	0,00567
	2. ≥ 3	0,99498		
GPA of the 2nd semester	1. < 3	0,90770	0,01552	0,00815
	2. ≥ 3	0,99679		
GPA of the 3rd semester	1. < 3	0,94268	0,01189	0,00612
	2. ≥ 3	1,00000		
GPA of the 4th semester	1. Regency	0,96290	0,00014	0,00007
	2. City	0,97306		
Domicile Region	1. A	0,99848	0,07072	**0,04553**
	2. Other Than A	0,55478		
School Accreditation	1. SNMPTN	0,99632	0,00825	0,00426
	2. Other Than SNMPTN	0,94029		
University Admission Method	1. No	0,99700	0,01095	0,00568
	2. Yes	0,93059		
First TUTEP Test Pass Status	1. Passed	0,81128	0,04141	0,02313
	2. No	0,97887		

Based on **Figure 1** it can be concluded that:

1. If the student is from a school with accreditation other than A then the student is classified as untimely graduation status.
2. If the student is from an A-accredited school, has never received a scholarship during college, and a GPA of the 1st semester ≥ 3, then the student is classified as untimely in graduation status.
3. If the student is from a school with A accreditation, never received a scholarship during college, GPA of the 1st semester < 3, dan GPA of the 4th semester < 3, then the student is classified as untimely in graduation status.
4. If the student is from an A-accredited school, never received a scholarship during college, GPA of the 1st semester < 3, dan GPA of the 4th semester ≥ 3, then the student is classified as graduating status on time.
5. If the student is from a school with A accreditation, has received a scholarship during college, and a GPA of the 1st semester < 3, then the student is classified as untimely in graduation status.
6. If the student is from a school with A accreditation, has received a scholarship during college, and a GPA of the 1st semester ≥ 3, then the student is classified as graduating status on time.

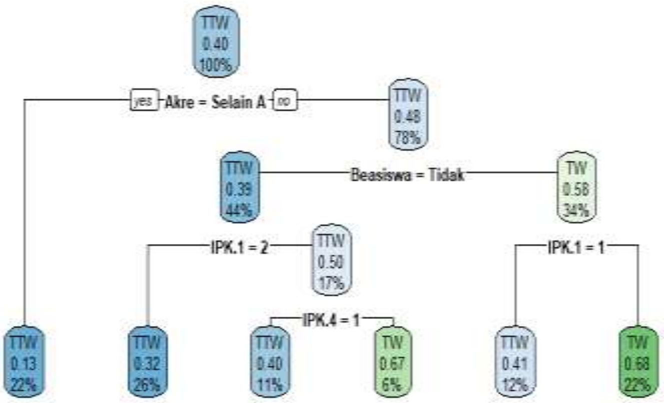

Figure 1. Decision Tree Algorithm C4.5

Based on **Figure 2** it can be concluded that:
1. If the student is from a school with accreditation other than A and the home county is a district, then the student is classified as untimely in graduation status.
2. If the student is from a school with accreditation other than A and his home county is a city, then the student is classified as graduating on time.
3. If the student is from an A-accredited school and has never received a scholarship while in college, then the student is classified as untimely in graduation status.
4. If the student is from an A-accredited school, has received a scholarship while in college and passed the first TUTEP, then the student is classified as untimely graduation status.
5. If the student is from an A-accredited school, has received a scholarship while in college and did not pass the first TUTEP, then the student is classified as graduating on time.

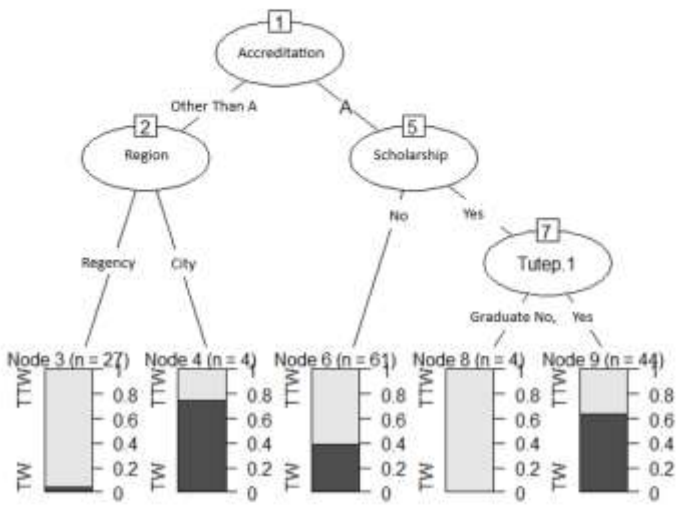

Figure 2. Decision Tree Algorithm C5.0

After the decision tree of the C4.5 and C5.0 algorithms is built, the next step is to evaluate the model using a confusion matrix. Based on the results of the evaluation of the classification model of the C4.5 Algorithm using the confusion matrix in **Table 4**, so that the accuracy value using **Equation (11)** was obtained by 70%.

Table 4. Confusion Matrix C4.5 Algorithm

		Actual	
		Not On Time	**On Time**
Predicted	**Not On Time**	71	29
	On Time	13	27

Based on the results of the evaluation of the C5.0 Algorithm classification model using the confusion matrix in **Table 5**, the accuracy value using **Equation (11)** was also obtained by 70%.

Table 5. Confusion Matrix C5.0 Algorithm

		Actual	
		Not On Time	**On Time**
Predicted	**Not On Time**	67	25
	On Time	17	31

3.2. Application of Adaptive Boosting Algorithm (Adaboost)

The initialization value of the weights from the data in the first iteration using **Equation (4)** with a maximum iteration of 100 is 0.007143. It found data that did not match the original class in the first iteration for C4.5 algorithms and C5.0 was 42 data. So, the next step is to calculate the error of the research data using **Equation (5)** obtained the data error value for the first iteration is 0.300006. After calculating the error on the research data, the next step is to determine the weight of the data using **Equation (6)**. The data weight value is 0.423635. Using **Equation (7)**, an update of the weight of the data is performed. Obtained the weight for the correctly classified data is 0.00468 and the weight for the misclassified data is 0.010911. **Table 6** is the result of the weight update on the first iteration.

Table 6. Results of Weight Update on C4.5 and C5.0 Algorithms

No	**Classification**	**Starting weight**	***Weight update***	**Classification**
1	TTW	TTW	0,0071	0,00468
2	TTW	TTW	0,0071	0,00468
3	TTW	TTW	0,0071	0,00468
4	TTW	TW	0,0071	0,01091
…	…	…	…	…
137	TW	TTW	0,0071	0,01091
138	TTW	TTW	0,0071	0,00468
139	TTW	TTW	0,0071	0,00468
140	TW	TW	0,0071	0,00468

After updating the data weights, the next step is to normalize the weights using **Equation (8)**. The weight normalization result for correctly classified data is 0.00510 and the weight for misclassified data is 0.01190. The calculation continues to be repeated until the error value ≥ 0.5 and the iteration has reached the maximum iteration. After that the process can stop. Next is to evaluate the model of the Adaboost Algorithm using a confusion matrix.

Based on the results of the evaluation of the C4.5 Algorithm classification model after being boosted using the Adaboost Algorithm using the confusion matrix in **Table 7**, the accuracy value using **Equation (11)** was obtained at 80.71%.

Table 7. Confusion Matrix Application of *Adaboost* to C4.5 Algorithm

		Actual	
		Not On Time	**On Time**
Predicted	**Not On Time**	73	16
	On Time	11	40

Based on the results of the evaluation of the C5.0 Algorithm classification model after being boosted using the Adaboost Algorithm using the confusion matrix in **Table 8**, the accuracy value using **Equation (11)** was obtained at 82.14%.

Table 8. Confusion Matrix Application of *Adaboost* to C5.0 Algorithm

		Actual	
		Not On Time	**On Time**
Predicted	**Not On Time**	74	15
	On Time	10	41

After Boosting using the Adaboost Algorithm, the accuracy value of the C4.5 Algorithm increased by 10.71%, while the C5.0 Algorithm increased by 12.14%. This accuracy value is relatively good because it can predict classification results for both classes into the categories of timely and not on time based on existing datasets and it is proven that the application of the Adaboost Algorithm to the C5.0 Algorithm is better than the C4.5 Algorithm.

4. Conclusions

Based on the results and discussion in this study, it can be concluded that the application of the Adaboost Algorithm to the C5.0 Algorithm is better than the C4.5 Algorithm in classifying the graduation of students of the Untan Statistics Study Program. The Adaboost algorithm was able to increase the accuracy of the C5.0 Algorithm by 12.14%. While in the C4.5 Algorithm, the Adaboost Algorithm increases accuracy by 10.71%.

References

[1] M. Kadafi, "Metode Naïve Bayes Classifier (NBC) Untuk Memprediksi Tingkat Kelulusan Mahasiswa Tepat Waktu," *JSI J. Sist. Inf. E-J*, vol. 12, no. 2, Oct 2020, doi: 10.36706/jsi.v12i2.12179.

[2] A. F. A. Rahman and S. Wartulas, "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus Di Universitas Peradaban)," *Indonesian Journal of Informatics and Research*, vol.1, no. 2, p. 70-77, 2020.

[3] A. Bisri and R. S. Wahono, "Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree," *J. Intell. Syst*, vol. 1, no. 1, pp. 27-32, 2015.

[4] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform*, vol. 2, no. 2, pp. 213–219, Feb. 2017, doi: 10.22202/ei.2016.v2i2.1465.

[5] W. D. Ramadana, N. Satyahadewi, and H. Perdana, "Penerapan Market Basket Analysis Pada Pola Pembelian Barang Oleh Konsumen Menggunakan Metode Algoritma Apriori," *Bimaster*, vol.11, no. 3, p. 431-438, 2020.

[6] J. Han, M. Kamber, dan J. Pei, *Data Mining Concepts And Taechniques, 4 Edition*. Walthmann: Morgan Kaufmann Publishers, 2001.

[7] M. S. Mustafa, M. R. Ramadhan, and A. P. Thenata, "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Creat. Inf. Technol. J*, vol. 4, no. 2, p. 151-162, Jan. 2018, doi: 10.24076/citec.2017v4i2.106.

[8] P. B. N. Setio, D. R. S. Saputro, and B. Winarno, "Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5," *in Prosiding Seminar Nasional Matematika*., pp. 64-71, 2020.

[9] P. A. Jusia, "Analisis Komparasi Pemodelan Algoritma Decision Tree Menggunakan Metode Particle Swarm Optimization Dan Metode Adaboost Untuk Prediksi Awal Penyakit Jantung", *in Prosiding Seminar Nasional Sistem Informasi*., pp. 64-71, 2018.

[10] A. Rohman and A. Rufiyanto, "Implementasi Data Mining Dengan Algoritma Decision Tree C4.5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandanaran", *in Proceeding SINTAK*., pp. 134-139, 2019.

[11] M. Chair, Y. N. Nasution, and N. A. Rizki, "Aplikasi Klasifikasi Algoritma C4.5 (Studi Kasus Masa Studi Mahasiswa Fakultas Matematika Dan Ilmu Pengetahuan Alam Universitas Mulawarman Angkatan 2008)," *Jurnal Informatika* Mulawarman, vol. 12, no. 1, p. 50-55.

[12] T. Tanti, P. Sirait, and A. Andri, "Optimalisasi Kinerja Klasifikasi Melalui Seleksi Fitur dan AdaBoost dalam Penanganan Ketidakseimbangan Kelas," *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 4, p. 1377-1385, Oct 2021, doi: 10.30865/mib.v5i4.3280.

[13] A. I. Prianti, R. Santoso, dan A. R. Hakim, "Perbandingan Metode K-Nearest Neighbor Dan Adaptive Boosting Pada Kasus Klasifikasi Multi Kelas," *J. Gaussian*, vol. 9, no. 3, pp. 346–354, Aug. 2020, doi: 10.14710/j.gauss.v9i3.28924.

[14] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets", *in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*., vol. 285, T. Herawan, M. M. Deris, and J. Abawajy, Eds. Singapore: Springer Singapore, 2014, pp. 13–22. doi: 10.1007/978-981-4585-18-7_2.

[15] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27-81, Dec 2019, doi: 10.1186/s40537-019-0192-5.