BAREKENG: Journal of Mathematics and Its Applications
December 2023 Volume 17 Issue 4 Page 2243–2252
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss4pp2243-2252

APPLICATION OF BAGGING CART IN THE CLASSIFICATION OF ON-TIME GRADUATION OF STUDENTS IN THE STATISTICS STUDY PROGRAM OF TANJUNGPURA UNIVERSITY

Widad Imtiyaz [1], Neva Satyahadewi [2*], Hendra Perdana [3]

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

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

Article History:

Received: 26th July 2023
Revised: 8th October 2023
Accepted: 8th November 2023

Keywords:

Bagging CART;
CART;
Ensemble;
Graduation.

ABSTRACT

The timeliness of graduation is used as the success of students in pursuing education which can be seen from the time taken and measured by the predicate of graduation obtained. The characteristics of students who tend to graduate not or on time can be analyzed using classification techniques. Classification and Regression Tree (CART) is one of the classification tree methods. There is a weakness in CART, which is less stable in predicting a single classification tree. The weaknesses in CART can be improved by using Ensemble methods, one of which is Bootstrap Aggregating (Bagging) which can reduce classification errors and increase accuracy in a single classification model. This study aims to classify and determine the accuracy of Bagging CART in the case of the accuracy of student graduation classification. The number of samples used is 140 data on the graduation status of Untan Statistics Study Program students from Period I of the 2017/2018 academic year to Period II of the 2022/2023 academic year. The variables used are the timeliness of graduation which is categorized into two namely Not and On Time, Gender, Semester 1 GPA, Semester 2 GPA, Semester 3 GPA, Semester 4 GPA, Region of Origin Domicile, High School Accreditation, Entry Path, Scholarship, and first TUTEP. A good classification can be seen from the accuracy value. The CART method obtained an accuracy value of 70%. While using the CART Bagging method obtained an accuracy value of 85.71%. Based on the accuracy value obtained, the application of the CART Bagging method can increase accuracy and correct classification errors on a single CART classification tree by 15.71% by resampling 25 times.

How to cite this article:

W. Imtiyaz, N. Satyahadewi, and H. Perdana, "APPLICATION OF BAGGING CART IN THE CLASSIFICATION OF ON-TIME GRADUATION OF STUDENTS IN THE STATISTICS STUDY PROGRAM OF TANJUNGPURA UNIVERSITY," *BAREKENG: J. Math. & App.*, vol. 17, iss. 4, pp. 2243-2252, December, 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

An assessment of the success of a university is if the student graduation rate is on time and high every year [1]. Students' success in pursuing their education can be seen from the time taken and the predicate of graduation obtained. Undergraduate students of the Faculty of Mathematics and Natural Sciences (FMIPA), Tanjungpura University (Untan) are said to be eligible to graduate on time if they can complete a study period of less than or equal to four years with a minimum learning load of 144 credits.

Data mining, or Knowledge Discovery from Data (KDD), is the process of extracting information from data sets and then transforming it into models that are easy to understand [2]. Data mining has several techniques, one of which is classification. The decision tree is one of the classification methods that involves the construction of a decision tree. Classification and Regression Tree (CART) is one of the classification tree methods that has been widely used in classification analysis because it is proven to provide a small classification error rate [3] and is easy to interpret [4].

CART belongs to the group of nonparametric statistical methods because no assumptions must be met [5]. The instability of the CART tree can be influenced by overfitting the model [6]. To improve the performance of CART classification, an ensemble method can be applied, namely bootstrap aggregating (bagging). The bagging method can reduce the error rate in the classification produced by a single CART classification model [3]. In the research of Nick Z. Zacharis (2018) examines predictive modeling in blended learning by classifying students into groups of students who successfully passed and students who failed, using the CART technique which obtained very high accuracy results of 99.1%. As in the research of Agwil et al (2020) which discusses the timeliness of student graduation in the S1 Mathematics study program using the CART Bagging method. The application of the CART Bagging method provides an accuracy value of 85.71% higher than the CART method with an accuracy value of 77.3% and resampling is done 50 times. While in the author's research this time using resampling 25 times.

Based on this description, the author analyzed by combining the Bagging method with the CART algorithm. The Bagging CART method aims to reduce the classification error produced by a single classification model (CART) in obtaining an overview of the characteristics and increasing accuracy in predicting the accuracy of student candidates who graduate not and on time graduation of the Statistics Study Program students.

2. RESEARCH METHODS

Data mining is extracting and transforming information from a data set into an understandable structure or model. Data mining has three main components: clustering or classification, association rules, and sequence analysis. Classification is used to classify each item in a data set into a predefined group [7]. One of the techniques in classification is decision trees [8]. Decision trees can be used to select the most relevant variables that can be used to form a model. The purpose of a decision tree is to obtain information that is useful in making a decision [9].

2.1 Algorithm Classification and Regression Tree (CART)

CART is one of the nonparametric machine learning methods [10]. The CART method uses response variables in the form of numeric and categorical data. Response variables in the form of categorical data are called classification trees. In contrast, response variables in the form of numeric data are called regression trees [4]. The basic idea of this method is to select predictor variables that have the greatest interaction with the response variable [3]. To see the greatest interaction on the response variable is known in the node sorting process based on goodness of split (best sorting criteria) [5]. Analysis in the CART method with the following stages:

The sorting process starts from the main node, which consists of the data to be sorted [5]. Sorting is done to sort the data into two groups of parts: the group that goes to the left node and the right node. The selection process for each parent node is based on the goodness of split (best selection criteria) [11]. The goodness of split is formed based on a heterogeneity function that aims to measure the level of heterogeneity of a class from a particular node in the classification tree using the Gini Index $i(t)$ in **Equation (1) [5]**:

$$i(t) = \sum_{\substack{j=1 \\ j \neq k}}^{J} P(j|t)\, P(k|t) \quad (1)$$

with $P(j|t)$ being the proportion of class j at node *t*, $P(k|t)$ being the proportion of class k at node *t*.

The goodness of split is an evaluation of the selection by the *s*-partitioner at node *t*. Goodness of split can also be defined as a decrease in heterogeneity. The value of $\phi(s,t)$ is used as a test of the goodness of split criterion (best splitting test criteria) [5]. Development on the tree by looking for all possible partitions at node t_1 so that a partition s^* is found that provides the highest value of heterogeneity reduction which can be seen in Equation (2) [12]:

$$\phi(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (2)$$

with P_L is the proportion of the number of objects that belong to t_L and at t_L is the left node. The parser that produces higher $\phi(s,t)$ is the best parser because it is able to reduce heterogeneity higher.

Class labeling is done from the beginning of node selection until the final node is formed because each node formed can become the final node. The labeling of each end node is based on the rule of the largest number of class members calculated using give an explanation for each symbol:

$$P(j_0|t) = max_j P(j|t) = max_j \frac{N_j(t)}{N(t)} \quad (3)$$

A node t will be the final node or will not be re-sorted if there is only one observation in each child node, all observations in each child node have identical response variable distributions, and there is a limit to the maximum number of tree depths determined [11].

The maximum classification tree that is formed is likely to be very large. The more sorting that is done, the higher the accuracy rate. However, a huge size will make it easier to understand, causing overfitting (complex value matching). This problem can be overcome by pruning the maximum classification tree to obtain an optimal classification tree. The pruning size used to obtain the optimal tree size uses Equation (5) [5]:

$$R(t) \quad = R(t_R) + R(t_L) \quad (4)$$

2.2 Bootstrap Aggregating (Bagging)

Bootstrap aggregating (Bagging) is an ensemble method that stabilizes and improves classification performance [13]. Bagging can also be used in some classification and regression methods to reduce the variance of a predictor and thus predict the estimation process [5]. There are two stages in the bagging process, namely bootstrapping, which is a sampling of the data owned (resampling), and aggregating is merging many conjecture values into one conjecture [14]. The process of making conjectures by bagging using trees is as follows [5]:

1. Bootstrapping steps

   a. Draw a random sample with size recovery from the data cluster.

   b. Construct the best tree based on the data.

   c. Repeat steps a-b for B times to obtain B classification trees.

2. Aggregating stage

   Using the majority vote rule, perform a combined estimation based on the B classification trees.

The use of bagging is especially helpful in overcoming the instability of classification trees and regression trees. In many data clusters, the bagging method can reduce the misclassification rate in classification cases [5].

2.3 Classification Accuracy

Apparent Error Rate (APER) is used to express the value of the proportion of samples that are misclassified in the classification process [15]. Calculation of classification accuracy to calculate the value of 1-APER will result in a small chance of misclassification. Calculation of classification results with a

small chance is said to be the best classification method. APER calculation can be calculated using the confusion matrix [16]. The confusion matrix is a method used to measure the performance of a classification method [17]. Then the form of the confusion matrix is shown in Table 1.

Table 1. Confusion Matrix

Prediction	Actual		
	Class 1	Class 2	
Class 1	n_{11}	n_{12}	N_1
Class 2	n_{21}	n_{22}	N_2
	N_1	N_2	N

$$APER = \left(\frac{n_{11} + n_{22}}{N}\right) \quad (5)$$

$$Sensitivity = \frac{n_{11}}{N_1} \times 100\% \quad (6)$$

$$Specificity = \frac{n_{22}}{N_2} \times 100\% \quad (7)$$

$$Accuracy = 1 - (APER) = 1 - \left(\frac{n_{11} + n_{22}}{N}\right) \times 100\% \quad (8)$$

The APER, sensitivity, specificity and accuracy values are obtained from the confusion matrix calculation using Equation (5), Equation (6), Equation (7), and Equation (8) [14]. Accuracy can express the level of accuracy in classifying [3].

3. RESULTS AND DISCUSSION

The data used in this study are primary data in the form of questionnaires and secondary data obtained from the academic FMIPA Untan and PDDikti Untan websites. The number of samples used was 140 data. The response attribute used is the graduation status of Untan Statistics Study Program students from Period I of the 2017/2018 academic year to Period II of the 2022/2023 academic year.

3.1 Classification Tree Formation CART

3.1.1 Splitting Node

First, compile all the variables that make up the candidate split. Then calculate the candidate branch value for the probability of each branch P_L (proportion of the number of objects that belong to t_L as the left node) and P_R (proportion of the number of objects that belong to t_R as the right node). After getting the value P_L and P_R, calculate $P(j|t)$ for each left and right branch. Then calculate the goodness of split using Equation (3) as follows:

$$\begin{aligned} \phi(s,t) &= i(t) - P_L i(t_L) - P_R \times i(t_R) \\ &= 0.48 - (0.779 \times 0.499) - (0.221 \times 0.225) \\ &= 0.042 \end{aligned}$$

Next, calculate P_L, P_R, and $P(j|t)$ for the left and right branches and the goodness of split similarly for all candidate branches. The highest Goodness of Split is chosen as the branch. The calculation results can be seen in Table 2.

Table 2. The Goodness of Split Value

No	Variable	Sorter		*Goodness of Split*
		Left Node	Right Node	
1.	High School Accreditation	A	Other Than A	0.042
2.	4th Semester GPA	<3	≥3	0.010
3.	Domicile Area of Origin	District	City	0.008
4.	3rd Semester GPA	<3	≥3	0.007
5.	Scholarship	No	Yes	0.007
6.	Entry Point	SNMPTN	Othet Than SNMPTN	0.006
7.	First TUTEP	Pass	Did not Pass	0.003
8.	2nd Semester GPA	<3	≥3	0.001
9.	Gender	Male	Female	0.000
10.	1st Semester GPA	<3	≥3	0.000

Based on **Table 2**, the results of the calculation of goodness of split suitability show that the highest branch candidate value is 0.042, namely the left branch of High School Accreditation A and the right branch of High School Accreditation other than A, so this branch candidate is chosen to be the root node. Then the other branches will be calculated similarly using the next iteration up to a maximum of four depths.

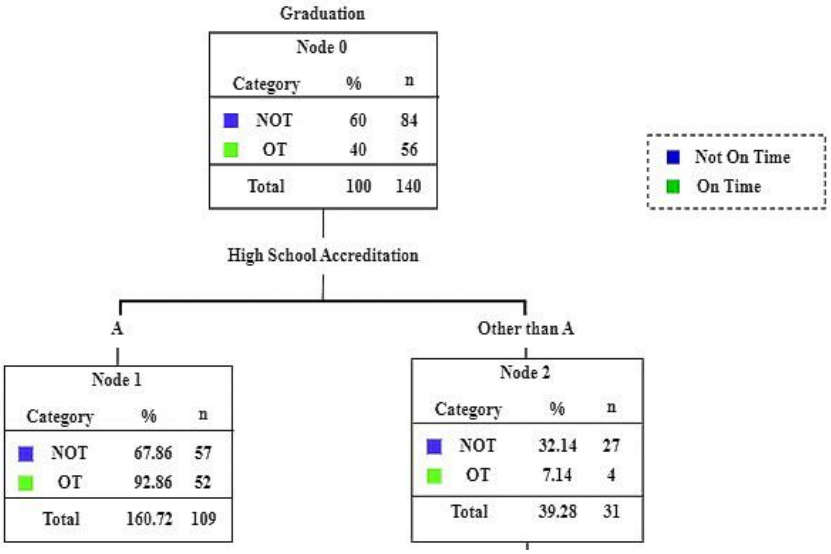

Figure 1. Root Node Sorting in The First Classification Tree

3.1.2 Class Assignment

The class labeling process on the vertices formed based on the rule of the highest number of class members. If $P(j_0|t) = max_j P(j|t)$, then $j_0 = j$ with j is graduated not on time and graduated on time. For example, on node 1, using **Equation (4)** as follows:

$$P(\text{Graduating Not On Time}|\text{Node 1}) = \frac{57}{109} = 0.523$$

$$P(\text{Graduate On Time}|\text{Node 1}) = \frac{52}{109} = 0.477$$

3.1.3 Stop the Splitting

The first maximal classification tree obtained has 4 inner nodes and 6 end nodes. The termination process can be seen in **Figure 2**. The termination process is at node 9 and node 10. At node 9 there are 65 observations in the same class (homogeneous), and at node 10 there are 75 observations at the maximum depth of 4 depths, so the node sorting process is stopped.

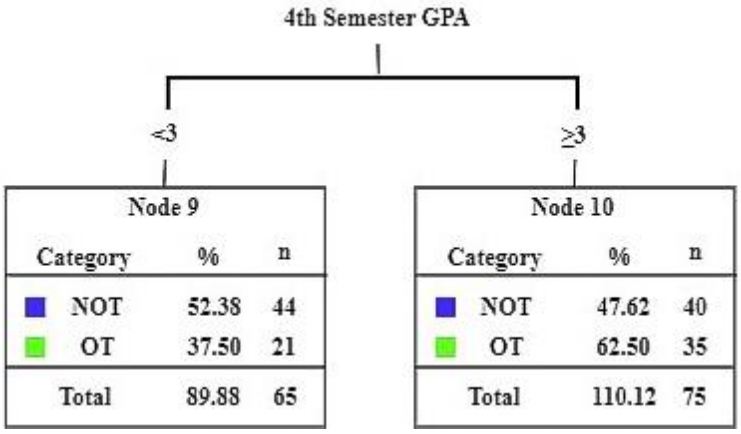

Figure 2. Node 9 and 10 in The First Maximal Classification Tree

3.1.4 Pruning the Classification Tree

The pruning process begins by taking t_L (left node) and t_R (right node) from the maximal tree generated from the parent node t. If nodes and the parent node satisfy the Equation $R(t) = R(t_L) + R(t_R)$, then node t_L and t_R are pruned. The process is repeated until no more pruning is possible. At this stage, node 0, node 1 and node 2 are used to be pruned, and the results are obtained using **Equation (5)** as follows:

$$\begin{aligned} R(t) &= R(t_L) + R(t_R), \text{by} \\ R(node_1) + R(node_2) &= 0.37 + 0.03 \\ R(t) &= 0.4 \quad R(node_0) \end{aligned}$$

This means that it is fulfilled. Therefore, it is done at node 1 and node 2. At node 0, the first classification tree is pruned. The pruning process is carried out until there is no more possibility of pruning. After the pruning process stops, the optimal classification tree can be seen in **Figure 3**.

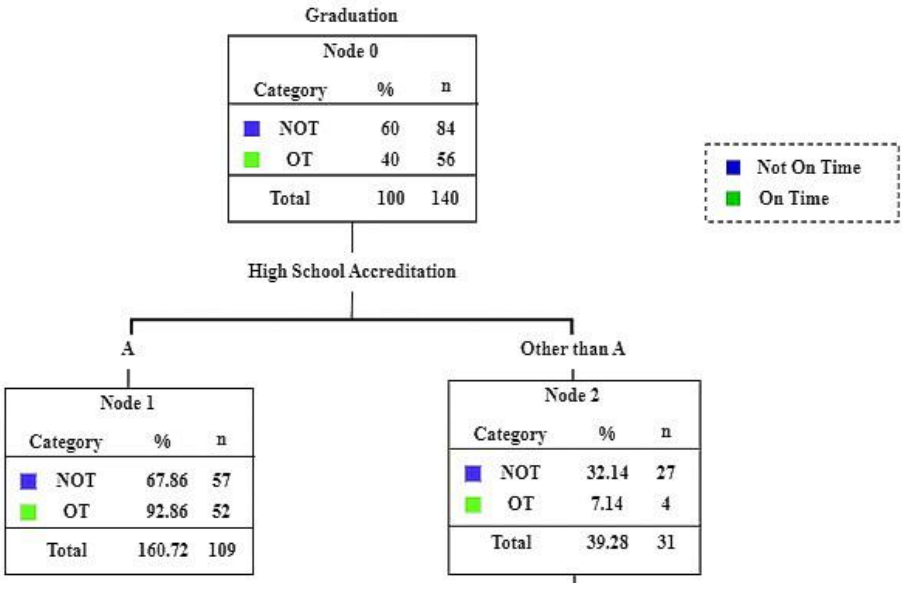

Figure 3. Node 0 in The First Pruned Maximal Classification Tree

Figure 3 is the result of the classification tree pruning process, which is the first optimal classification tree. Graduating on time is influenced by High School Accreditation.

3.2 Classification Interpretation on CART

The single classification tree prediction results obtained will be used to test the accuracy of classification using **Equation (8)** in the CART method first. Based on the prediction results of the CART classification tree, 71 data are predicted to fall into the class not on time, while 27 are predicted to fall into the class on time. The final result of the estimation is based on the most votes, so the final prediction falls into the class not on time. The results of the classification accuracy test on the CART classification tree can be seen in **Table 3**.

Table 3. CART Classification Result

Prediction	Actual		
	Not On Time	On Time	
Not on Time	71	29	100
On Time	13	27	40
	84	56	140

$$
\begin{aligned}
APER &= \left(1 - \frac{n_{11}+n_{22}}{N}\right) \times 100\% \\
&= \left(1 - \frac{71+27}{140}\right) \times 100\% \\
&= (1 - 0.7) \times 100\% \\
&= 0.3 \times 100\% = 30\% \\
Accuracy &= 1 - (APER) = 70\%\%
\end{aligned}
$$

Based on the calculation from **Table 3**, the APER value (misclassified value) is obtained at 30% with an accuracy value of 70%. The accuracy value illustrates that the overall accuracy of the classification produced by the CART model is 70%.

3.3 Application of Bagging CART

The bootstrap aggregating (bagging) method is performed to improve the accuracy of the previously known CART. After performing the bootstrap process 25 times to form a classification tree and predict data on each classification tree. Repetition is carried out 25 times, because several times doing repeated sampling results in the same accuracy so that the resampling is used 25 times. The next step is to aggregate predictions based on 25 guesses on the data with majority vote rules. Based on the prediction results from 25 classification trees, 71 data are predicted to fall into the not on time class, while 49 data are predicted to fall into the on time class. The final result of the prediction is based on the majority vote, so the final prediction falls into the class not on time.

3.4 Classification Interpretation on Bagging CART

The combined prediction results on CART and bagging will be used to test the accuracy of classification using **Equation (8)** on the combination of the CART bagging method. The results of the classification accuracy test on the observation classification tree can be seen in **Table 4**.

Table 4. Bagging CART Classification Results

Observation	Prediction		
	Not On Time	On Time	
Not On Time	71	7	78
On Time	13	49	62
	84	56	140

$$APER = \left(1 - \frac{n_{11}+n_{22}}{N}\right) \times 100\%$$
$$= \left(1 - \frac{71+49}{140}\right) \times 100\%$$
$$= (1 - 0.8571) \times 100\%$$
$$= 0.1429 \times 100\% = 14.29\%$$
$$Accuracy = 1 - (APER) = 85.71\%$$

Based on the calculations from **Table 4**, the APER value is 14.29% with an accuracy value of 85.71%, and this accuracy value illustrates that the overall classification accuracy generated by the CART Bagging model is 85.71%.

3.5 Classification Accuracy on CART and Bagging CART

The results of comparing the classification accuracy value on the initial CART tree with the classification accuracy value after using bagging CART can be seen in **Table 5**.

Table 5. Classification Accuracy Result

	Classification Accuracy Value (%)
First Classification Tree (CART)	70
Classification Tree with Bagging (25 times)	85.71
Improvement	15.71

Based on **Table 5**, it can be seen that the classification accuracy value by applying the bagging technique is 85.71%. Using bagging techniques can increase classification accuracy from 70% in CART to 85.71% in bagging CART.

The Bagging CART method's accuracy value is better than the CART single classification method because the bagging CART method is resampled 25 times, and modeling is carried out. Predictions are made from each sample formed. Representative predictions are selected using the majority vote or choosing the most votes. With the repetition of sampling 25 times, this results in the convergence of the prediction results in **Table 5**.

4. CONCLUSIONS

Based on the discussion results described, determining the accuracy value of the graduation classification results of Tanjungpura University Statistics Study Program students obtained a value of 85.71% using the bagging CART method. The bagging CART method can increase classification accuracy from 70% in the initial classification tree to 85.71% in bagging CART. It is better than the classification tree without bagging because it can increase accuracy by 15.71%.

REFERENCES

[1] A. Wibowo, D. Manongga, and H. D. Purnomo, "The Utilization of Naive Bayes and C.45 in Predicting The Timeliness of Students' Graduation," *Scientific Journal of Informatics*, vol. 7, no. 1, pp. 99–112, 2020, doi: 10.15294/sji.v7i1.24241.

[2] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[3] W. Agwil, H. Fransiska, and N. Hidayati, "Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging Cart," *FIBONACCI: Jurnal Pendidikan Matematika dan Matematika*, vol. 6, no. 2, p. 155, 2020, doi: 10.24853/fbc.6.2.155-166.

[4] N. Z. Zacharis, "Classification and regression trees (CART) for predictive modeling in blended learning," *International Journal of Intelligent Systems and Applications*, vol. 10, no. 3, pp. 1–9, 2018, doi: 10.5815/ijisa.2018.03.01.

[5] D. Ratnaningrum, M. A. Mukid, and T. Wuryandari, "Analisis Klasifikasi Nasabah Kredit Menggunakan Bootstrap Aggregating Classification And Regression Trees (Bagging CART)," *Jurnal Gaussian*, vol. 5, no. 1, pp. 81–90, 2016, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[6] S. Gocheva-Ilieva, H. Kulina, and A. Ivanov, "Assessment of students' achievements and competencies in mathematics using cart and cart ensembles and bagging with combined model improvement by mars," *Mathematics*, vol. 9, no. 1, pp. 1–17, 2021, doi: 10.3390/math9010062.

[7] G. Kesavaraj and S. Sukumaran, "06726842," 2013.

[8] A. Maesya and T. Hendiyanti, "Forecasting Student Graduation with Classification and Regression Tree (CART) Algorithm," *IOP Conference Series: Materials Science and Engineering*, vol. 621, no. 1, 2019, doi: 10.1088/1757-899X/621/1/012005.

[9] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130–135, 2015, doi: 10.11919/j.issn.1002-0829.215044.

[10] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105400, 2020, doi: 10.1016/j.cmpb.2020.105400.

[11] S. Innassuraiya, T. Widiharih, and I. T. Utami, "ANALISIS KLASIFIKASI MENGGUNAKAN METODE REGRESI LOGISTIK BINER DAN BOOTSTRAP AGGREGATING CLASSIFICATION AND REGRESSION TREES (BAGGING CART) (Studi Kasus: Nasabah Koperasi Simpan Pinjam Dan Pembiayaan Syariah (KSPPS))," vol. 11, no. 2, pp. 183–194, 2022, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/

[12] S. H. Sumartini, "Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya," *Jurnal Sains dan Seni ITS*, vol. 4, no. 2, pp. 211–216, 2015, [Online]. Available: https://www.neliti.com/publications/15687/penggunaan-metode-classification-and-regression-trees-cart-untuk-klasifikasi-re

[13] X. Li, X. Liu, and P. Gong, "Integrating ensemble-urban cellular automata model with an uncertainty map to improve the performance of a single model," *International Journal of Geographical Information Science*, vol. 29, no. 5, pp. 762–785, 2015, doi: 10.1080/13658816.2014.997237.

[14] P. Radha and B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques," vol. 1, no. 6, pp. 334–339, 2014.

[15] D. Kusnandar, N. N. Debataraja, S. W. Rizki, and E. Saputri, "Water quality mapping in pontianak city using multiple discriminant analysis," *AIP Conference Proceedings*, vol. 2268, no. September, 2020, doi: 10.1063/5.0016809.

[16] Y. T. Samuel, J. J. Hutapea, and B. Jonathan, "Predicting the timeliness of student graduation using decision tree c4.5 algorithm in universitas advent Indonesia," *Proceedings of 2019 International Conference on Information and Communication Technology and Systems, ICTS 2019*, pp. 276–280, 2019, doi: 10.1109/ICTS.2019.8850948.

[17] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification," *IOP Conference Series: Materials Science and Engineering*, vol. 879, no. 1, 2020, doi: 10.1088/1757-899X/879/1/012076.