INTENSIF, Vol.6 No.1 February 2022
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v6i1.16156

Analysis of CART and Random Forest on Statistics Student Status at Universitas Terbuka

Received:
7 July 2021
Accepted:
26 January 2022
Published:
11 February 2022

[1*]Siti Hadijah Hasanah, [2]Eka Julianti
[1]Statistika, Universitas Terbuka
[2]Sistem Informasi, Universitas Terbuka
E-mail: [1]sitihadijah@campus.ut.ac.id, ekajulianti@ecampus.ut.ac.id

*Corresponding Author

Abstract— CART and Random Forest are part of machine learning which is an essential part of the purpose of this research. CART is used to determine student status indicators, and Random Forest improves classification accuracy results. Based on the results of CART, three parameters can affect student status, namely the year of initial registration, number of rolls, and credits. Meanwhile, based on the classification accuracy results, RF can improve the accuracy performance on student status data with a difference in the percentage of CART by 1.44% in training data and testing data by 2.24%.

Keywords— CART; Distance Learning; Ensemble; Machine Learning; Random Forest

Corresponding Author:

Siti Hadijah Hasanah,
Statistika,
Universitas Terbuka,
Email: sitihadijah@campus.ut.ac.id

I. INTRODUCTION

The Open University (UT) has several faculties, one of which is the Faculty of Science and Technology (FST). This faculty was previously known as the Faculty of Mathematics and Natural Sciences (FMIPA). FST offers a Bachelor of Statistics study program. This study program was established in 1994 and has succeeded in producing Statistics graduates from various characteristics students possess. UT has an open and distance learning system, so it is one of UT's advantages compared to other universities in Indonesia by opening 39 service offices spread throughout Indonesia. UT does not limit the period of study completion, and there is no application of a dropout system. There is no limitation on the year of diploma graduation or age. The time for registration or registration is free throughout the year [1] [2]. Based on the advantages and convenience of studying at UT, there is one problem that must be faced by us, namely the active status of students. It happens because students are given the convenience of registering at any time, so many UT students are inactive in a particular semester. They do not know when the student will become active again [1].

This study aims to determine what indicators can affect student status to classify the status of active and inactive students. From the results of this study, a policy will also be made as a solution for students who can become inactive students in the next semester to reduce the number of inactive students in the UT Statistics study program. Several statistical methods used in the classification technique are CART and Random Forest (RF). Classification and Regression Trees (CART) is a data exploration method based on decision tree techniques. A regression tree is generated when the response variable is numeric, while a classification tree is generated when the response variable is categorical. The tree formed from the binary recursive sorting process in a data cluster makes the response variable values in each data cluster more homogeneous. CART is sensitive to new data so that if there is a slight change in a data set, it can result in a significant change in decision trees [3]. The way to solve this problem is to use an Ensemble method. This method is a classifier set that is trained individually. The prediction results obtained are combined when classifying new data [4] [5]. Some of the Ensemble techniques include Bagging [6][7], Boosting [7][8], and Random forest [9][10][11]. Random Forest (RF) is a development of the CART method by applying bootstrap aggregating (bagging) and random feature selection methods [9]. In RF, many trees are grown to form like a forest. The analysis is carried out by collecting these trees in a data set consisting of n observations and p explanatory variables.

Several studies have shown that RF is better than some machine learning methods, namely empirical studies using residential apartment data. These studies indicate that RF works better than the CHAID, CART, KNN, Multiple Regression, ANN (MLP) methods, and RBF and

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v6i1.16156

Boosted Trees [12]. In the study of solar radiation, they are forecasting one day to 6 days ahead using MARS, CART, M5, and RF models. RF gives the best accuracy results while CART produces the lowest accuracy results [11]. Research in predicting the delay or progress of ship arrivals using three approaches, namely ANN Backpropagation, CART, and RF, from the results of the three methods obtained, RF is better than BP and CART [13]. Research on many attributes for early cancer detection so that dimension reduction is needed, then the CART and RF methods are applied. After analysis, it is found that RF is the method that produces the best performance compared to CART [14].

II. RESEARCH METHOD

A. Data and Variables

This research data comes from UT Statistics Study Program students in 2019, totaling 1493, divided into 2, namely 70% training data and 30% testing data with nine independent variables and one response variable. These nine independent and one response were taken based on the data collection results, which we generally use to determine student characteristics. Characteristics of UT Statistics Study Program students consist of:

Table 1. CHARACTERISTICS STUDENTS AT DEPARTMENT OF STATISTICS, UNIVERSITAS TERBUKA

Variable	Information	Scale	Category
X1	Gender	Nominal	1 = Man 2 = Woman
X2	Age	Interval	
X3	Education	Ordinal	1 = Senior High School 2 = Associate Degree 3 = bachelor’s degree 4 = master’s degree 5 = Doctoral Degree
X4	Marital Status	Nominal	1 = Single 2 = Married
X5	Job	Nominal	1 = Unemployment 2 = Private Employees 3 = Entrepreneur 4 = Civil Servants 5 = Army/Police
X6	Initial Registration Year	Interval	
X7	Number of Registrations	Interval	
X8	Credits	Interval	
X9	GPA	Interval	
Y	Student Status	Nominal	0 = Not Active 1 = Active

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v6i1.16156

B. Data analysis

The construction of the CART classification tree includes three things [15], [16] :

a. Selection of sorters (split)

   Each split only depends on the value originating from one independent variable.

b. Terminal node determination

   node t can be used as a terminal node if there is no significant decrease in heterogeneity in splitting, there is only one observation on each child node, or there is a minimum limit of n. There is a limit on the number of maximum tree levels or depth.

c. Class label marking

   Based on the rules of the most significant number of class members.

   The formation of the classification tree stops if there is only one observation in each child node or there is a minimum limit of n. All observations in each child node are identical, and there is a limit to the maximum number of tree levels. After the maximum tree is formed, the tree pruning stage is carried out, which aims to prevent the formation of vast and complex classification trees so that a decent tree size is obtained based on cost complexity pruning.

RF is done in the following way [17] :

a. Perform a random sampling of size n

b. with recovery on cluster data. This stage is the bootstrap stage.

   In this stage, the tree is constructed until it reaches its maximum size (without pruning). At each node, the disaggregation is done by selecting m explanatory variables at random, where $m << p$. The best disaggregation is chosen from them explanatory variables. This stage is the stage of random feature selection.

c. Repeat steps a and b k times so that a forest of k trees is formed. The observation response is predicted by aggregating the expected results of k trees, and the classification is based on the majority vote.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v6i1.16156

III. RESULT AND DISCUSSION

The following is a descriptive analysis that describes student status based on the following characteristics of the data:

Table 2. DESCRIPTIVE STATISTICS

Characteristics	Information	Not Active (%)	Active (%)
Gender	Man	35,9	64,1
	Woman	29,1	70,9
Education	Senior High School	31,9	68,1
	Associate Degree	35,9	64,1
	Bachelor's degree	29,0	71,0
	Master's degree	00,0	100,0
	Doctoral Degree	100,0	00,0
Marital Status	Single	26,1	73,9
	Married	38,0	62,0
Job	Unemployment	27,0	73,0
	Private Employees	30,5	69,5
	Entrepreneur	26,4	73,6
	Civil Servants	41,2	58,8
	Army/Police	100,0	00,0
Age	≤ 25	8,50	91,5
	26 - 35	39,1	60,9
	36 - 45	48,8	51,2
	46 - 55	44,1	55,9
	≥ 56	40,0	60,0

The table above it shows that the percentage of women who are active as students are 70.9%, the percentage of women is more significant than 64.1% of men. So it can be concluded that women are more dominant in taking the Statistics study program than men.Percentages based on education that S1 and S2 are active as students are 71% and 100%, respectively. It shows that the status of active students who have S1 and S2 education is easier to attend lectures because they have studied Statistics during previous lectures and took Statistics as a support to improve competence in terms of data analysis and processing. While those with S3 education were not active 100% after the analysis, it was found that there was only one doctoral student who was not active as a student.

The marital status of unmarried students of 73.9% shows that unmarried students are easier to follow the lecture process in Statistics Study Program. At the same time, the employment status of students who work as entrepreneurs is 73.6% and not working 73.0%. It shows that students who work as entrepreneurs and those who do not work will find it easier to follow the lecture process at the Statistics Study Program. The highest percentage of age as active students was in the age group 25 years at 91.5%, followed by the age group 26-35 years, 56 years, 46-55 years, and 36-45 years. So it can be concluded that the age group ≤ 25 years is the student who mostly takes Statistics Study Program because they are fresh graduates from high school and above.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v6i1.16156

CART can produce a model that is simple and easy to interpret. The resulting CART model is based on the variables. The variables affect the response and work as markers for forming a node.

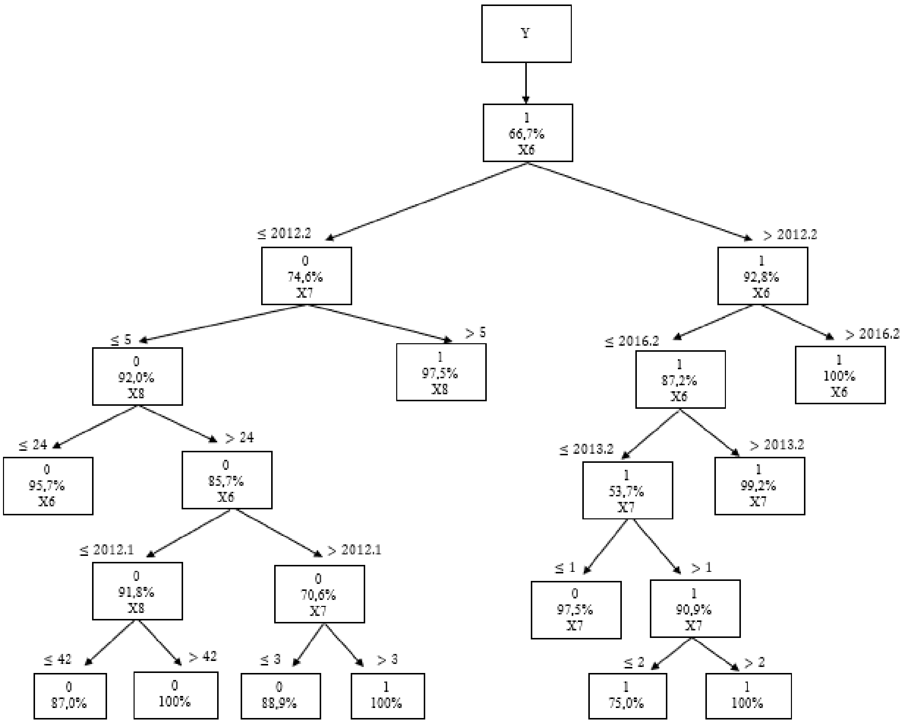

Figure 1. CART RESULT

In Figure 1, it can be explained that, Previously, nine independent variables affected the response of the student status of the Statistics Study Program UT. Still, after an analysis using CART, it was found that three independent variables that could affect student status (Y) were the year of initial registration (X6), the number of registrations (X7), and the number of registrations (X7). Credits (X8). The initial registration year for Statistics students is divided into two groups, namely the batch below 2012.2 and those above 2012.2. The class of students above 2012.2 has a relatively large active percentage of 92.8% compared to the collection below 2012.2 of 25.4%; it can be concluded that the batch of students above 2012.2 has the ability and great opportunity to complete their Bachelor of Statistics education at UT. Some of the criteria for student status are said to be active according to Figure 1 as follows:

1. Student status with the criteria for the initial registration year 2012.1-2012.2 taking more than 24 credits, and the number of registrations is 3-5 times, the student status is 100% active.
2. Student status with the initial registration year 2012.2 and the number of registrations is five times. The student status is 97.5% active.
3. Student status with the initial registration year above 2016.2 then the student status is 100% active.

Based on this research data, it can also be concluded that the saturation point of students in studying at the Statistics Study Program is students who have taken lectures for more than seven years. After analyzing using the CART method, the authors build the Ensemble method, which is used to improve the accuracy of the classification of student status by using the development of the CART method, namely RF.

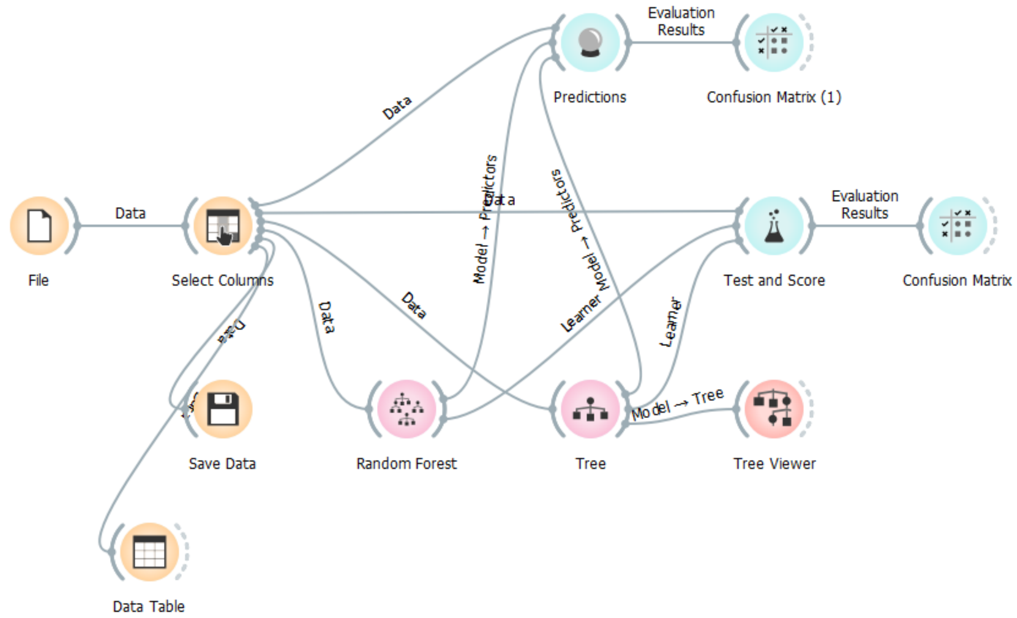

Figure 2. ORANGE APP ON CART AND RF

The picture above explains that the application used in this study is the Orange application. Orange is an open-source data mining software. Figure 2 illustrates that this study aims to apply the CART and RF methods and compare the two methods. The two methods have good accuracy in classifying the active status of Statistics students based on cross-validation results on training data and testing data.

Table 3. RESULTS OF RF AND CART DATA TRAINING

Student Status	**RF (%)**		**CART (%)**	
	Not Active	***Active***	***Not Active***	***Active***
Not Active	97,99	2,01	93,97	6,03
Active	3,72	96,28	3,87	96,13

Based on the results above, it can be explained that the RF training data is better in determining the accuracy level of classification of inactive students by 97.99% than CART by 93.97%, and it is also better in determining the accuracy level of classification of active students by 96.28%. of CART by 96.13%.

Table 4. PERFORMANCE MEASURES CLASSIFICATION METHOD OF TRAINING DATA

Performance Measures Classification Method of Training Data	**Method**	
	RF	***CART***
Sensitivity	0,9292	0,9237
Specificity	0,9897	0,9697
1-APER	0,9685	0,9541
Press's Q	918,1644	862,8107
AUC	0,9850	0,9630

Based on the evaluation model results in table 5, it can be explained that in the training data, the RF sensitivity value is more significant than CART at 0.9292. It means that the student status data is classified as inactive at 92.92%. The RF specificity value is more significant than CART at 0.9897, which means that the student status is classified as active at 98.97%. The value of 1-APER RF can increase the accuracy of CART classification in general. The value of Press's Q on RF and CART are 918.1644 and 862.8107. It is known that the critical value at a significance level of 0.01 is 6.63, so that Press's Q is greater than the critical value, so it can be concluded that the prediction results of RF and CART are significant at the 0.01 level. The AUC values on RF and CART are 0.9850 and 0.9630. Respectively, these values are in the range of 0.90 -1.00, which are included in the excellent classification criteria.

Table 5. RESULTS OF RF AND CART DATA TESTING

Student Status	**RF (%)**		**CART (%)**	
	Not Active	***Active***	***Not Active***	***Active***
Not Active	94,67	5,33	90,67	9,33
Active	2,69	97,31	4,04	95,96

Based on the above results, it can be explained that the RF testing data is better in determining the accuracy level of classification of inactive students by 94.67% than CART by 90.67%, and it is also better in determining the accuracy level of classification of active students by 97.31%. of CART by 95.96%.

Table 6. PERFORMANCE MEASURES CLASSIFICATION METHOD OF TESTING DATA

Performance Measures Classification Method of Testing Data	Metode	
	RF	CART
Sensitivity	0,9467	0,9189
Specificity	0,9731	0,9532
1-APER	0,9642	0,9418
Press's Q	385,2908	349,0492
AUC	0,9830	0,9500

Based on the evaluation model results in table 6, it can be explained that in the testing data, the RF sensitivity value is more significant than CART at 0.9467. It means that the student status data is classified as inactive at 94.67%. The RF specificity value is more significant than CART at 0.9731, which means that the data student status is rightly classified as active by 97.31%. The value of 1-APER RF can increase the accuracy of CART classification in general. The value of Press's Q on RF and CART are 385.2908 and 349.0492, respectively. It is known that the critical value at a significance level of 0.01 is 6.63, so Press's Q is greater than the critical value, so it can be concluded that the prediction results of RF and CART are significant at the 0.01 level. The AUC values on RF and CART are 0.9830 and 0.9500, respectively. These values are in the range of 0.90 -1.00, included in the excellent classification criteria.

IV. CONCLUSION

The parameters in classifying the status of Statistics Study Program students using the CART method are nine parameters consisting of gender, age, education, marital status, job, initial registration year, number of registrations, credits, and GPA. Based on the results of the CART analysis, parameters that affect student status are obtained, namely the year of initial registration, number of registrations, and credits. The results of the analysis of RF training data are better in determining the accuracy level of classification of inactive students by 97.99% than CART, which is 93.97%, and RF has a classification accuracy of active students of 96.28% and better than CART of 96, 13%. Likewise, the RF testing data is better in determining the accuracy level of classification of inactive students by 94.67% than CART 90.67%. It is also better in determining the classification accuracy of active students 97.31% from CART of 95.96%. So based on these results, RF can improve the accuracy performance on student status data with a difference in the percentage of CART by 1.44% in training data and testing data by 2.24%.

REFERENCES

[1] S. Hasanah and S. Permatasari, “Metode Klasifikasi Jaringan Syaraf Tiruan Backpropagation Pada Mahasiswa Statistika Universitas Terbuka,” vol. 14, no. 2, pp. 243–252, 2020, doi: 10.30598/barekengvol14iss2pp249-258.

[2] S. H. Hasanah, "Multivariate Adaptive Regression Splines (MARS) for Modeling The Student Status at Universitas Terbuka," vol. 7, no. 1, pp. 51–58, 2021, doi: https://doi.org/10.15642/mantik.2021.7.1.51-58.

[3] C. D. Sutton, "Classification and Regression Trees, Bagging, and Boosting," Handb. Stat., vol. 24, no. 04, pp. 303–329, 2005, doi: 10.1016/S0169-7161(04)24011-1.

[4] R. Maclin, "Popular Ensemble Methods : An Empirical Study Popular Ensemble Methods : An Empirical Study," J. Artif. Intell. Res., vol. 11, no. July, pp. 169–198, 2016.

[5] M. van Wezel and R. Potharst, "Improved customer choice predictions using ensemble methods," Eur. J. Oper. Res., vol. 181, no. 1, pp. 436–452, 2007, doi: 10.1016/j.ejor.2006.05.029.

[6] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996, doi:

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v6i1.16156

10.1007/bf00058655.
[7] H. Inoue and R. Inoue, "A very large platform for floating offshore facilities," Coast. Ocean Sp. Util. III. Proc. Symp. Genoa, 1993, pp. 533–551, 1995.
[8] R. E. Schapire, "Explaining adaboost," Empir. Inference Festschrift Honor Vladimir N. Vapnik, pp. 37–52, 2013, doi: 10.1007/978-3-642-41136-6_5.
[9] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[10] A. Davies and Z. Ghahramani, "The Random Forest Kernel and other kernels for big data from random partitions," 2014, [Online]. Available: http://arxiv.org/abs/1402.4293.
[11] R. Srivastava, A. N. Tiwari, and V. K. Giri, "Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India," Heliyon, vol. 5, no. 10, p. e02692, 2019, doi: 10.1016/j.heliyon.2019.e02692.
[12] E. A. Antipov and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics," Expert Syst. Appl., vol. 39, no. 2, pp. 1772–1778, 2012, doi: 10.1016/j.eswa.2011.08.077.
[13] J. Yu et al., "Ship arrival prediction and its value on daily container terminal operation," Ocean Eng., vol. 157, no. January, pp. 73–86, 2018, doi: 10.1016/j.oceaneng.2018.03.038.
[14] R. Chairunisa, “Perbandingan CART dan Random Forest untuk Deteksi Kanker berbasis Klasifikasi Data Microarray,” vol. 1, no. 1, pp. 19–25, 2017.
[15] N. Z. Zacharis, "Classification and regression trees (CART) for predictive modeling in blended learning," Int. J. Intell. Syst. Appl., vol. 10, no. 3, pp. 1–9, 2018, doi: 10.5815/ijisa.2018.03.01.
[16] A. Hartati, I. Zain, and S. Suprih, “Kepala Rumah Tangga di Jawa Timur,” J. Sains Dan Seni Its, vol. 1, no. 1, pp. 101–105, 2012.
[17] V. Y. Kullarni and P. K. Sinha, "Random Forest Classifier: A Survey and Future Research Directions," Int. J. Adv. Comput., vol. 36, no. 1, pp. 1144–1156, 2013.