CART and Random Forest Analysis on Graduation Status of Halu Oleo University Students

Gusti Arviana Rahman[1,2*], Khairil Anwar Notodiputro[3], Bagus Sartono[4], La Surimi[5]

[1]Statistic Study Program, Department of Mathematics FMIPA, Halu Oleo University, Kendari, Indonesia
[2,3,4]Department of Statistics, IPB University, Bogor, Indonesia
[5]Computer Science Study Program, Department of Mathematics FMIPA, Halu Oleo University, Kendari, Indonesia
*Corresponding author: arviana.rahman@uho.ac.id

Received: 10 October 2024
Revised: 16 September2025
Accepted: 30 October 2025

ABSTRACT – **Classification and Regression Tree (CART) is a popular classification method and it is used in various fields. The method is capable to be applied on various data conditions. An alternative method of CART is random forest. These two methods of classification were studied in this paper using graduation data of Halu Oleo University. This data was interesting due to the imbalance problem existed in the data. We compared several scenarios, namely the CART and Random Forest methods, Random Forest with oversampling, and Random Forest with undersampling. There were three explanatory variables considered in the model including Study Program, GPA, and TOEFL score. The results showed that the best method to classify the student's graduation status at Halu Oleo University is Random Forest without handling imbalanced data, as it provided the highest sensitivity. This suggests that Random Forest, even without specific adjustments for data imbalance, can effectively capture the patterns in the data and provide accurate classifications, making it a robust choice for this dataset.**

Keywords– **classification tree, imbalance data, oversampling, undersampling, statistical learning.**

I. INTRODUCTION

Classification and Regression Trees (CART) is a nonparametric classification method that is popularly used. This method has been used in solving problems in various fields such as health, marketing, social, financial, and so on. This method is able to deal with various data conditions. CART has many advantages, including being able to explore high-dimensional data with efficient computing, can be used in a combination of continuous and categorical data, and is easy to interpret. Among the many advantages of CART, the weakness of this classification tree method is that it is less stable when learning data changes which will cause major changes in the prediction results of the classification tree [1]. To overcome the weaknesses of the CART method, a method is needed that can be used to increase the prediction accuracy of an unstable classifier. According to [2] a method to increase the prediction accuracy of an unstable classifier, namely the Ensemble method.

The Ensemble method involves combining several individual classifiers, where the predictions from each classifier are aggregated into a final prediction using majority voting for classification tasks or averaging for regression tasks [3]. Previous research shows that the Ensemble method often produces more accurate predictions compared to a single classifier [4]. One of the newest Ensemble Methods is Random Forest which was developed from the Bagging process [5][6]. Random Forest was first introduced by Breiman in 2003. Random Forest has the advantage of a faster computational iteration process [7] [8].

Tertiary education represents the concluding phase of formal schooling, serving as an optional pathway [9]. In this era, universities are required to maintain high standards by utilizing their available resources. The quality of a university's various study programs in Indonesia is measured through accreditation conducted by the National Accreditation Board of Higher Education (BAN-PT). BAN-PT, established by the government, operates autonomously to oversee and enhance college accreditation processes. Accreditation functions as an external mechanism for ensuring quality within the framework of the Higher Education Quality Assurance System. It operates through the evaluation of adherence to standards outlined in Higher Education Standards, thereby fostering accountability and improvement within the academic realm [10].

There are nine criteria used to evaluate university accreditation, among which are students' outcomes and achievements in fulfilling the Tridharma responsibilities. The criteria for Tridharma outcomes entail factors such as the Grade Point Average (GPA) of graduates, the duration of their studies, their academic success, and timely completion, also the on-time graduation student percentage for each program [11]. Therefore, universities are tasked with upholding student quality by ensuring the excellence of graduates, which can be assessed through factors like the duration of their studies and their ability to graduate on time. This means that the university must pay more attention to its students so that they can complete their studies on time, thereby increasing the percentage of students who graduate on schedule.

Halu Oleo University, situated in Kendari City, Southeast Sulawesi, stands as one of Indonesia's state universities. UHO has four educational levels which encompass 17 departments which are divided into 121 study programs. It's important to be aware that the duration of study varies depending on the level of education, so we can categorize as on-time graduation for each program. The D3 program spans over 6 semesters (equivalent to 36 months), while the

undergraduate program lasts for 8 semesters (48 months). Master's programs typically run for 4 semesters (24 months), and doctoral (S3) programs extend over 8 semesters (48 months). The duration of the study will determine the student's graduation status, whether on-time or not. As an example, to ensure the excellence of its students, UHO's academic departments regularly assess the student's progress after each semester, after the first four semesters, at the end of eight semesters, and upon completion of the study program [12].

Predicting whether a student will graduate on time can be accomplished using a classification model, such as a classification tree. A key advantage of this approach is that it does not rely on specific assumptions, like the normality of data distribution. To mitigate the instability and high variance associated with individual trees, ensemble techniques can be employed to combine multiple classification trees [13]. Several prior educational studies have employed data mining algorithms, specifically utilizing a classification tree as the fundamental learning model. [1] introduced a prediction system using a classification tree algorithm. However, their study was restricted by a small dataset and the use of only one type of classifier. Subsequently, [2] used a larger dataset to predict timely student graduations within engineering faculties at various private universities in Indonesia, also employing a classification tree. Similar to the previous study, they used only one type of classifier but compared validation results across different amounts of testing data.

Furthermore, previous research on using CART for predicting student graduation has been applied to classify on-time graduations of students in the Statistics Study Program at Tanjungpura University [14]. Another study utilizing CART focused on forecasting student graduations at Pakuan University [15]. [16] achieved CART classification accuracy of 94.2% and binary logistic regression classification accuracy of 86.7% using student profile data from the Faculty of Mathematics and Natural Sciences at Brawijaya University.

In contrast, [17] compared four classification methods (J48, PART, Random Forest, and Bayes Network) to predict student performance across three colleges in India. Similarly, [18] used seven classifiers (J48, Random Forest, Rap Tree, Logistic Model Tree (LMT), Naïve Bayes, BayesNet, and PART) to predict academic performance at an engineering college in India, finding Random Forest to be the most effective. [19] analyzed CART and Random Forest to predict the status of Statistics students at Universitas Terbuka. [20] studied the prediction of undergraduate students' study completion status using MissForest imputation in Random Forest and XGBoost models. [21] applied various machine learning models, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, XGBoost, and CatBoost, to predict on-time graduation, with Random Forest identified as the best model. Additionally, [22] explored the use of Random Oversampling Techniques to improve graduation time accuracy prediction using the Random Forest algorithm.

Similarly, [23] predicted students' academic performance at an engineering college in India using seven different classifiers: J48, Random Forest, Rap Tree, Logistic Model Tree (LMT), Naïve Bayes, BayesNet, and PART. They found Random Forest to be the most efficient algorithm. [24] applied several machine learning algorithms to predict student performance in China, including Extreme Gradient Boosting (XGBoost), Random Forest, Lasso, Elastic Net, Support Vector Machine, and Classification Tree. XGBoost outperformed the other models. [25] researched the prediction of undergraduate students' study completion status using MissForest imputation in Random Forest and XGBoost models. [26] analyzed CART and Random Forest for predicting the status of Statistics students at Universitas Terbuka. Additionally, [27] explored the application of Random Oversampling Techniques to predict graduation time accuracy using the Random Forest algorithm.

The main objective of this study is comparing several scenarios, namely the CART which is capable to be applied on various data conditions and Random Forest methods which is consist of Random Forest without handling, Random Forest with oversampling, and Random Forest with undersampling. These methods of classification were applied in this paper using graduation data from Halu Oleo University in October 2022 until October 2023.

II. LITERATURE REVIEW

A. *Classification and Regression Trees (CART)*

Classification and Regression Trees are classification methods using historical data to build a decision tree. The CART methodology began to be developed in the 80s by Breiman, Friedman, Olshen, and Stone in their paper entitled "Classification and Regression Trees" (1984). CART is a non-parametric discriminant analysis designed to present decision rules in the form of a binary tree that divides data into learning samples within univariate linear limits. This analysis produces hierarchical data groups starting from the root node for the entire learning sample and ending in small groups of homogeneous observations. Each terminal node is given a class label or predicted value, resulting in a tree structure that can be interpreted as a decision tree [11]. Development The CART classification tree includes three things, namely:

a. Selection of sorters (split)
b. Determination of terminal nodes
c. Class label marking

B. *Random Forest*

The Random Forest (RF) method is a machine learning method that uses ensemble learning techniques, where several decision trees are combined to make more accurate predictions. RF applies bootstrap aggregating (bagging) and random feature selection methods in forming its decision tree. In the case of classification, the final decision is used

as a prediction from the RF model using a majority vote system. Random Forest (RF) is a powerful machine learning method for making predictions and might face challenges with imbalanced datasets.

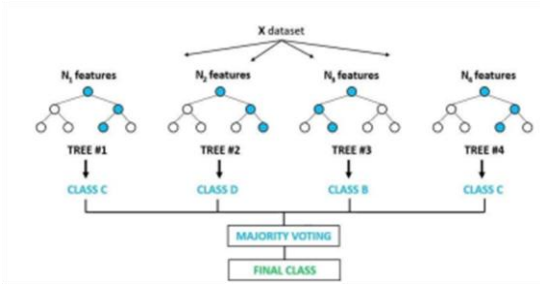

Figure 1 Random Forest

The Random Forest method is a method based on decision trees. During Random Forest training, many decision trees will be created so that from the samples in the training set several trees will be produced. Random Forest requires a combination of multiple decision trees to accurately predict outcomes. When using a random forest as a classifier, each decision tree can produce the same or different answers. For example, decision trees A, B, E, and F predict the result 1. Meanwhile, decision trees C and D predict the result 0. Because there are many alternative answers in the decision tree and the probability is high, the random forest takes the predicted results. The results of multiple decision trees are based on majority voting and more accurate outcome predictions.

C. *Confusion Matrix*

When calculating the accuracy of the classification algorithm, the Confusion Matrix method is used. This method produces accuracy, precision, and recall values. Accuracy is the percentage of accuracy in classifying data that is classified correctly after testing [2]. This research measures accuracy using the confusion matrix method as follows.

Table 1. Confusion Matrix

Categorical		**Reference**	
		True	False
Prediction	True	TP	FP
	False	FN	TN

where

TP : The classification is correct on the prediction and correct on the actual value

FP : The classification is correct on the prediction and incorrect on the actual value

FN : The classification is incorrect on the prediction and correct on the actual value

TN : The classification is wrong on the prediction and wrong on the actual value

Based on the values in the Confusion Matrix, the Accuracy, Sensitivity and Specificity values can be calculated. Accuracy states the level of accuracy of the classifier in classifying observations. The following is the formula used to see classification performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

III. METHODOLOGY

A. *Data*

The data used in this research is data originating from the UPT Center for Technology, Information and Communication, and Academic Information Systems of Halu Oleo University (https://siakadbeta.uho.ac.id/). The data consists of 850 graduate students who completed their studies from October 2022 to October 2023, especially in the Faculty of Mathematics and Natural Science. There are thirteen variables used in this research, namely one predictor variable and twelve response variables

Table 2. Variables

Variable	**Information**	**Scale**

Student graduation status (Y)	0 : not on time 1 : on time	Nominal
Gender (X1)	1 : Female 2 : Male	Nominal
Study Program (X2)	Bachelor of Mathematics Bachelor of Physics Bachelor of Chemistry Bachelor of Biology Bachelor of Biotechnology Bachelor of Statistics Bachelor of Computer Science etc	Ordinal
GPA (X3)		Interval
Father's Education (X4)	1: No school 2: elementary school, middle school, high school 3 : D1, D2, D3 4 : S1, S2, S3	Ordinal
Mother's Education(X5)	1 : No school 2 : elementary school, middle school, high school 3 : D1, D2, D3 4 : S1, S2, S3	Ordinal
Father's Occupation (X6)	1 : Not Working 2 : Working	Nominal
Mother's Occupation (X7)	1 : Not Working 2 : Working	Nominal
Father's Income (X8)	1 : Rp.0 / No income 2 : Less than Rp. 500,000-Rp. 3,000,000 3 : Rp. 3,000,001-etc	Nominal
Mother's Income (X9)	1 : Rp.0 / No income 2 : Less than Rp. 500,000-Rp. 3,000,000 3 : Rp. 3,000,001-etc	Nominal
Organizational Activity Status (X10)	1 : Not active 2 : Active	Nominal
Residence status (X11)	1 : Separated 2: With parents/family	Nominal
TOEFL score (X12)		Interval

B. Method

The data analysis procedures carried out in this research are:

1) Pre-process data
   At this stage, an examination of the data is carried out. If there is incomplete data, the observation is removed. The data used in this research were 850 observations, because there was incomplete data, only 827 observations were taken.
2) Processing and analysis of data
   - The data is divided into two types of data, namely training data and testing data. In this research, the library (createDataPartition) in the R program is used. In this research, there are three scenarios used, see Table 2.
   - The training data is then used to create a classification algorithm, namely CART and Random Forest.
   - In analysis with Random Forest, the analysis is carried out in three parts, namely without handling, with handling using oversampling, and with handling using undersampling.
   - After obtaining a model, either from CART or Random Forest, the classification model obtained is tested using testing data, or this stage is referred to as model validity.
   - Obtain test results on the model, either using CART or Random Forest and compare the results obtained.
   - Determine the variable importance.
3) Interpretation is carried out.

IV. RESULTS AND DISCUSSIONS

A. Data Exploration

The data used in this research is student graduation status at Halu Oleo University with a sample size of 850 observations. After preprocessing the data, there were found 23 observations contained missing values, so the dataset used in this research is 827 observations. The data proportion of not on-time student graduation status is 84.52% (699

observations) and the proportion of on-time student graduation status is 15.48% (128 observations) as can be seen in Figure 2.

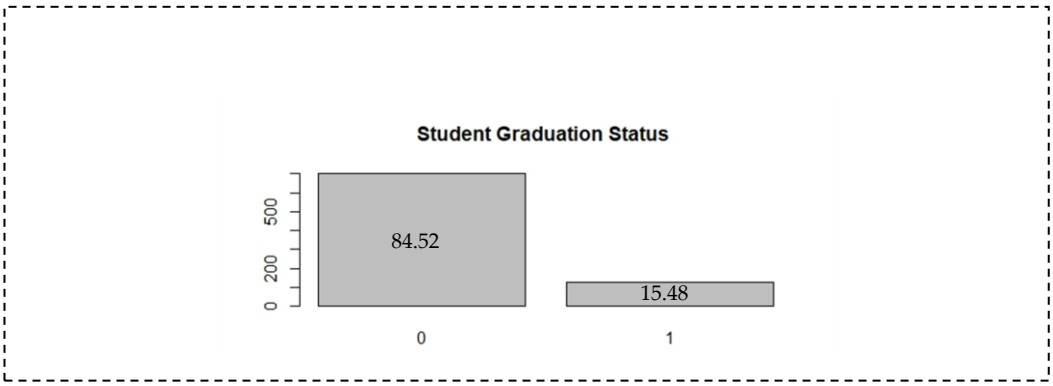

Figure 2 Barplot for student graduation status at Halu Oleo University

The data used for the predictor variable is shown in Figure 2. It shows that the data is imbalanced, where the minority class receives the family program, and the majority class does not graduate on-time. this is called imbalanced data. Apart from that, data exploration was also carried out to see the distribution of explanatory variables in each class of response variables, as in Figure 3.

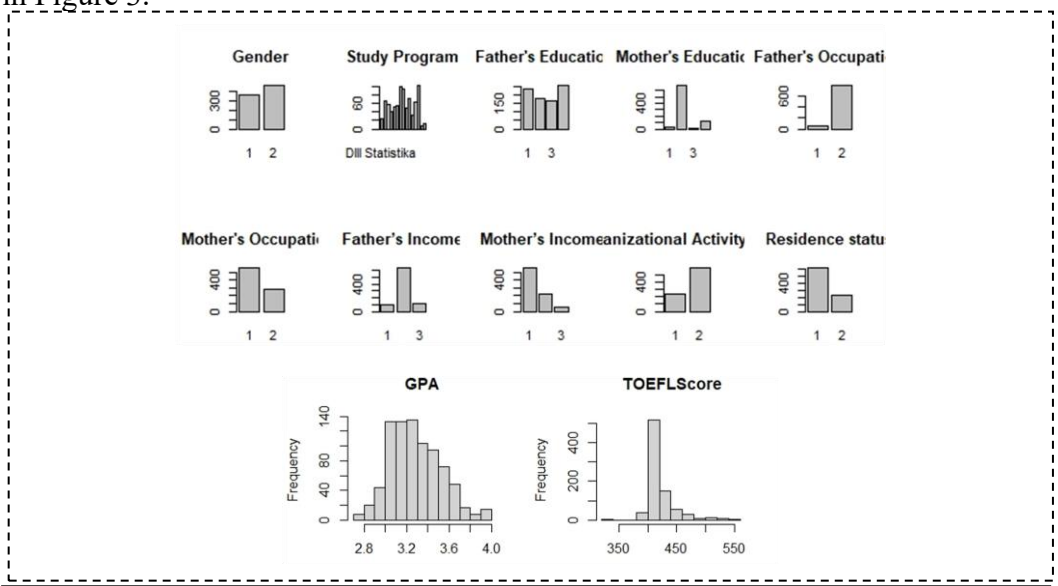

Figure 3 Explanatory variables for student graduation status at Halu Oleo University

B. Data Partition

In this paper, the data will be divided into training and testing data to show which scenario will give the best result for this case. It is shown in Table 2.

Table 2 Scenarios for Spillting Data

Scenarios	Percentage (%)		Data Amount	
	Training	Testing	Training	Testing
I	70	30	578	249
II	80	20	661	166
III	90	10	744	83

The scenario used in this research consists of three scenarios, which are (i) 70% training data and 30% testing data, (ii) 80% training data and 20% testing data, also (iii) 90% training data and 10% testing data. All the scenarios will be applied in both, the CART method and random forest, including handling the imbalanced data.

C. Resampling Data

The problem of observational data shows that there is class imbalance data. This problem should be handled by the resampling method to obtain a training dataset with a more balanced class ratio. Further, oversampling and undersampling handling are only used in the Random Forest method, not for the CART method. Data comparisons between majority and minority classes in the basic training data and the resampling are shown in Table 3.

Table 3 Scenarios for Splitting Data

Sampling Methods	Scenarios	Data Amount	
		0	1
Original	I	489	83
	II	540	96
	III	593	107
Oversampling	I	489	489
	II	540	540
	III	593	593
Undersampling	I	83	83
	II	96	96
	III	107	107

Table 3 presents the comparison of training data that are used to random forest analysis not only without handling but also with handling using oversampling and undersampling for each scenario. Oversampling and undersampling processes using the same amount for both on-time graduation status and non-on-time graduation status.

D. CART Analysis

The CART classification method uses a decision tree algorithm. The results of the CART analysis in the form of a classification tree in the case of not on timely graduation of Halu Oleo University students with a graduation time of October 2022 to October 2023 are shown in Figure 4. The structure of the confusion matrix can be seen in Table 1.

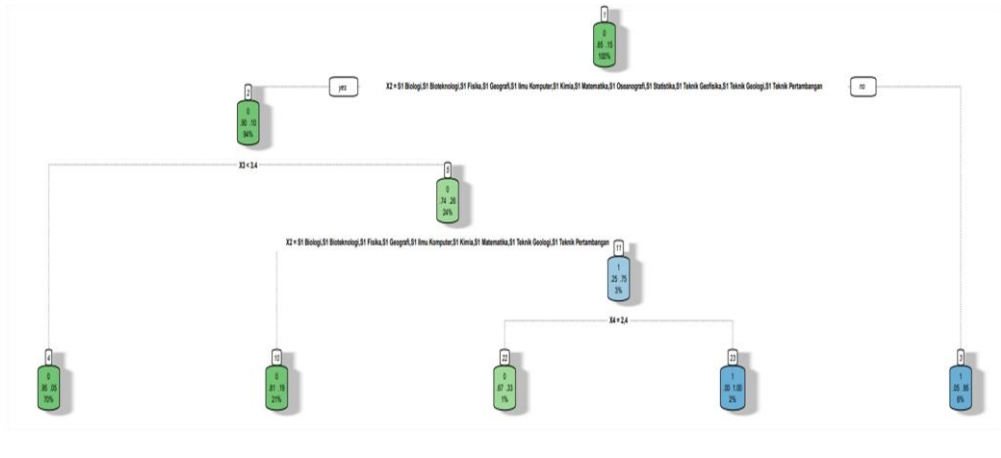

Figure 4 CART of Graduation Status at Halu Oleo University

Based on the classification tree in Figure 4, it is known that the variables used to sort the CART classification tree and which most determine graduation time are Study Program (X2), GPA (X3), and Father's Education (X4). The confusion matrix of CART is presented in Table 4 which consist of three scenarios. The accuracy and sensitivity values of the confusion matrix can be obtained by using equations (1) and (2) as we noticed before in Chapter 2.

Table 4 Confusion matrix of CART

Scenario	Prediction	Actual		Value	
		1	0	Accuracy	Sensitivity
I	1	18	3	91.16	98.58
	0	19	209		
II	1	11	0	89.76	96.73
	0	17	138		
III	1	5	0	89.16	95.31
	0	9	69		
			Average	90.03	96.97

Table 4 shows that the 1st scenario (70% training data and 30% testing data) is given the highest accuracy value, which

is 91.16%. the lowest accuracy value is given by 3rd scenario, which is 98.58%. This can be seen from the TN value is the highest value from the confusion matrix. Table 4 also shows that the TN value for each scenario has the highest value than others. This means that the CART model used in all of the scenarios only has high accuracy in predicting the majority class but cannot predict the minority class well.

E. Random Forest Classification Analysis

1. Random Forest without handling

Classification with Random Forest is a modification of Bagging CART. The selection of selectors in a random forest does not involve all variables but only some of them are taken randomly in each selection. For each sorting, 3 variables from 11 variables are randomly selected and the best sorter is then searched for from these 3 variables. The results of the random forest classification tree classification carried out on data on inaccurate graduation times for Halu Oleo University students are almost the same as using CART where 3 variables are most often used as tree sorters and most determine the accuracy of graduation sequentially, namely Study Program (X2), GPA (X3), and TOEFL Score (X12).

In addition to Study Program, GPA, and TOEFL Score, which are dominant factors in determining students' timely graduation, other variables such as Organizational Activity Status and Parents' Occupation are also relevant in this context. Organizational Activity Status refers to students' participation in organizational activities during their studies. Engagement in such activities can contribute to the development of soft skills, such as time management, communication, and teamwork, which are essential for both academic and non-academic success. However, excessive involvement without proper time management may disrupt study priorities and extend the duration of study. Therefore, universities can facilitate time management training for students actively involved in organizations to ensure they can still complete their studies on time.

Meanwhile, Parents' Occupation can influence the support provided to students during their studies. Parents with more stable jobs or higher incomes are often better able to provide educational facilities, such as access to books, technology, and additional courses. On the other hand, if parents' jobs demand significant amounts of their time, their emotional support or involvement in their children's academic progress might be limited. In this regard, universities could initiate mentorship programs for students from families with limited support to help bridge this gap.

Three scenarios were used in this part, combined with two types of mtry, and three types of ntree. We use mtry=4 and mtry=10 which are obtained from tuning parameters and grid search. The ntrees which are used in these random forest analysis consist of 100, 500, and 1000. The accuracy and sensitivity values of the confusion matrix can be obtained by using equations (1) and (2) as we noticed before in Chapter 2.

Table 5 Confusion matrix of Random Forest without handling

Scenario	mtry	ntree	Prediction	Actual		Value (%)	
				1	0	Accuracy	Sensitivity
I	4	100	1	19	3	88.63	98.57
			0	26	207		
		500	1	19	3	88.63	98.57
			0	26	207		
		1000	1	19	2	89.02	99.05
			0	26	208		
	10	100	1	21	6	88.24	97.14
			0	24	204		
		500	1	19	3	88.63	96.67
			0	26	207		
		1000	1	23	7	88.63	96.67
			0	22	203		
II	4	100	1	12	3	87.96	98.11
			0	20	156		
		500	1	12	3	87.96	98.11
			0	20	156		
		1000	1	12	3	87.96	98.11
			0	20	156		
	10	100	1	15	5	88.48	96.86
			0	17	154		
		500	1	17	5	89.53	96.86
			0	15	154		
		1000	1	16	5	89.01	96.86
			0	16	154		
III	4	100	1	10	5	87.40	95.28
			0	11	101		

		500	1	7	3	86.61	97.17
			0	14	103		
		1000	1	10	5	87.40	95.28
			0	11	101		
	10	100	1	10	5	87.40	95.28
			0	11	101		
		500	1	10	5	87.40	95.28
			0	11	101		
		1000	1	10	6	86.61	94.34
			0	11	100		
Average						**88.08**	**96.90**

Table 5 shows that the highest accuracy is obtained in 2nd scenario with mtry=10 and ntree=500, with an accuracy value is 89,53%, besides the lowest accuracy is obtained in the 3rd scenario, which is 86.61% (mtry=4, ntree=500). According to the sensitivity value, the highest sensitivity is obtained from the 1st scenario with mtry=4 and ntree=1000, which is 99.05%. the lowest sensitivity value is found in the 3rd scenario with mtry=10 and ntree=1000, which is 94.34%. The average value of accuracy in this part is 88.08%, besides the average value of sensitivity is 96.90%.

2. Random Forest with handling imbalanced data with oversampling

According to Figure 3, we know that there is imbalanced data in this case, so it must be handled. One method for handling imbalanced data is oversampling. Three scenarios were used in this part, combined with two types of mtry, and three types of ntree. We use mtry=4 and mtry=10 which are obtained from tuning parameters and grid search. The ntrees which are used in these random forest analysis consist of 100, 500, and 1000. The accuracy and sensitivity values of the confusion matrix can be obtained by using equations (1) and (2) as we noticed before in Chapter 2. The following is a confusion matrix which is the result of random forest classification by oversampling.

Table 6. Random Forest Classification Result with Oversampling

Scenario	mtry	ntree	Prediction	Actual		Value (%)	
				1	0	Accuracy	Sensitivity
I	4	100	1	22	10	87.06	95.24
			0	23	200		
		500	1	23	8	88.24	96.19
			0	22	202		
		1000	1	22	10	87.45	95.24
			0	23	200		
	10	100	1	24	9	88.24	95.71
			0	21	201		
		500	1	27	10	89.02	95.24
			0	18	200		
		1000	1	24	8	88.63	96.19
			0	21	202		
II	4	100	1	15	7	87.43	95.60
			0	17	152		
		500	1	15	6	87.96	96.23
			0	17	153		
		1000	1	14	4	87.43	96.23
			0	18	153		
	10	100	1	18	8	88.48	94.97
			0	14	151		
		500	1	15	6	87.96	96.23
			0	17	153		
		1000	1	15	8	86.91	94.97
			0	17	151		
III	4	100	1	10	6	86.61	94.34
			0	11	100		
		500	1	9	7	85.04	93.40
			0	12	99		
		1000	1	11	5	88.19	95.28
			0	10	101		
	10	100	1	10	6	86.61	94.34

			0	11	100		
		500	1	9	5	86.61	95.28
			0	12	101		
		1000	1	10	6	86.61	94.34
			0	11	100		
Average						**87.47**	**95.28**

Table 6 shows that the highest accuracy is obtained in 1st scenario with mtry=10 and ntree=500, with an accuracy value is 89,02%, besides the lowest accuracy is obtained in the 3rd scenario, which is 85.04% (mtry=4, ntree=500). According to the sensitivity value, the highest sensitivity is obtained from the 2nd scenario. The value is 96.23% which is obtained from (i) mtry=4 and ntree=500, (2) mtry=4 and ntree=1000, and (iii) mtry=10 and ntree=500. The lowest sensitivity value is found in the 3rd scenario with mtry=4 and ntree=500, which is 93.40%. The average value of accuracy in this part which is used oversampling to handle the imbalanced data is 87.47%, besides the average value of sensitivity is 95.28%.

3. Random Forest with handling imbalanced data with undersampling

Another method for handling the imbalanced data is undersampling. Three scenarios were used in this part, combined with two types of mtry, and three types of ntree. We use mtry=4 and mtry=10 which are obtained from tuning parameters and grid search. The ntrees which are used in these random forest analysis consist of 100, 500, and 1000. The accuracy and sensitivity values of the confusion matrix can be obtained by using equations (1) and (2) as we noticed before in Chapter 2. The following is a confusion matrix which is the result of random forest classification by oversampling.

Table 7 Random Forest Classification Result with Undersampling

Scenario	mtry	ntree	Prediction	Actual		Value (%)	
				1	0	Accuracy	Sensitivity
I	4	100	1	37	51	76.68	75.71
			0	8	159		
		500	1	36	68	69.80	67.62
			0	9	142		
		1000	1	38	51	76.68	75.24
			0	7	158		
	10	100	1	39	6	78.43	76.67
			0	6	161		
		500	1	35	59	72.94	71.90
			0	10	151		
		1000	1	37	50	77.25	76.19
			0	8	160		
II	4	100	1	26	40	75.92	84.84
			0	6	119		
		500	1	25	45	72.77	71.70
			0	7	114		
		1000	1	25	38	76.44	76.10
			0	7	121		
	10	100	1	26	35	78.53	77.99
			0	6	124		
		500	1	26	41	75.39	74.21
			0	6	118		
		1000	1	26	33	79.58	79.25
			0	6	126		
III	4	100	1	17	29	74.02	72.64
			0	4	77		
		500	1	20	30	75.59	71.70
			0	1	76		
		1000	1	19	29	75.59	72.64
			0	2	77		
	10	100	1	26	35	78.53	77.99
			0	6	124		
		500	1	19	23	80.31	78.30
			0	2	83		
		1000	1	26	33	79.58	79.25
			0	6	126		
Average						**76.34**	**75.55**

Table 7 shows that the highest accuracy is obtained in 3rd scenario with mtry=10 and ntree=500, with an accuracy value is 80.31%, besides the lowest accuracy is obtained in the 1st scenario, which is 69.80% (mtry=4, ntree=500). According to the sensitivity value, the highest sensitivity is obtained from the 3rd scenario. The value is 84.84% which is obtained from mtry=4 and mtree=100. The lowest sensitivity value is found in the 1st scenario with mtry=4 and ntree=500, which is 67.62%. The average value of accuracy in this part which is used oversampling to handle the imbalanced data is 76.34%, besides the average value of sensitivity is 75.55%.

F. *Classification Comparison*

The performance of the Random Forest method is measured by prediction accuracy and sensitivity. The comparison of four method will be presented in two tables, which are based on accuracy and sensitivity separately.

Table 8 Performance of Classification Methods Based on Accuracy

Classification Method	Scenario	mtry	ntree	Accuracy (%)
CART	I	-	-	91.16
Random Forest without handling	II	10	500	89.53
Random Forest with oversampling	I	10	500	89.02
Random Forest with undersampling	III	10	500	80.31

Table 8 shows the performance of each method used in this case according to accuracy value. The best method based on accuracy is the CART method, because it has the highest accuracy value, which is 91.16%. The lowest accuracy value is held by random forest with undersampling. From the results of these accuracy calculations it can be seen that although usually random forests are more robust and less susceptible to overfitting than single trees, in the case of simple data, the use of many trees can introduce unnecessary complexity, which sometimes actually reduces accuracy. whereas as a single decision tree method, CART tends to be simpler and easier to interpret. In cases where the data is not very complex or the amount of data is relatively small, CART can be better at capturing patterns without overfitting.

Table 9 Performance of Classification Methods Based on Sensitivity

Classification Method	Scenario	mtry	ntree	Sensitivity (%)
CART	I	-	-	98.58
Random Forest without handling	I	4	1000	99.05
Random Forest with oversampling	II, III	4,10	500,1000	96.23
Random Forest with undersampling	II	4	100	84.84

Based on Table 9, the highest sensitivity value was obtained from the Random Forest method without treatment, where the value was equal to 99.05%. It is followed respectively by CART, oversampling, and undersampling. Even without special treatment, the random forest without handling imbalanced data can be more robust to class imbalance than CART. This is because combining predictions from multiple trees can help ensure that minority classes get adequate representation. This can be seen in the results of sensitivity calculations, where the value for random forest without handling is higher when compared with CART or with oversampling and undersampling. Random Forest with oversampling/undersampling is used to overcome imbalances but can introduce new problems. Oversampling can lead to overfitting of the minority class while undersampling can remove important information from the majority class.

G. *Variable Importance*

Based on Table 9, the variable importance in this research is taken from the random forest without handling, because it can help the university to increase the number of on-time graduation student status. It is shown in Figure 5.

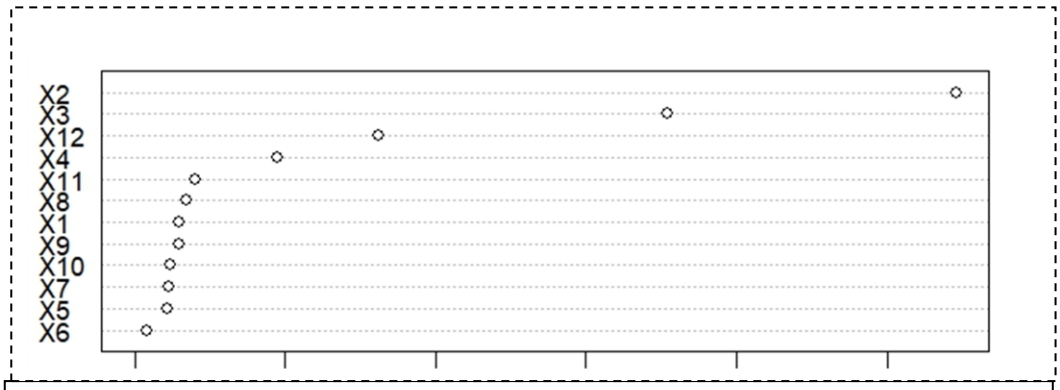

Figure 5 Variable Importance

Figure 5 shows that the three variables that have the most influence in determining the graduation status of students at Halu Oleo University are the Study Program (X2), GPA (X3) and TOEFL Score (X12). The lowest variable that influences a student's graduation status is the Father's Occupation (X6). If the university wants to further increase the number of students graduating on time, it needs to carry out further evaluations regarding other variables.

V. CONCLUSIONS AND SUGGESTIONS

There are three variables that determine the student's graduation status: Study Program (X2), GPA (X3), and TOEFL Score (X12). Random Forest tends to provide better sensitivity to the majority class in imbalanced data because dominant patterns in the majority class are easier for the algorithm to learn and detect. However, to obtain balanced performance between the majority and minority classes, it is important to apply techniques for handling imbalanced data or adjust the model parameters according to the needs of the problem being studied. The results showed that the best method to classify the student's graduation status at Halu Oleo University is Random Forest without handling imbalanced data, as it provided the highest sensitivity. This suggests that Random Forest, even without specific adjustments for data imbalance, can effectively capture the patterns in the data and provide accurate classifications, making it a robust choice for this dataset.

Furthermore, the results of this study provide valuable insights for universities to enhance their academic strategies. The identification of key variables such as GPA, TOEFL score, and Study Program can guide the development of targeted interventions to improve students' timely graduation rates. Academic Coaching Programs: Based on the importance of GPA and TOEFL scores, universities can design coaching programs to support students academically. For instance, early intervention programs can be established for students with low GPAs, offering academic advising, tutoring, or mentorship to help improve their performance. Curriculum Development for Timely Graduation: The influence of Study Program on graduation outcomes suggests a need to review and optimize curricula. By identifying bottlenecks in course availability or prerequisites, universities can streamline program structures to support students in completing their studies within the expected timeframe. English Proficiency Enhancement: Given the significance of TOEFL scores, universities should consider expanding English language support services, such as intensive language workshops, online modules, or peer-assisted learning programs. Enhancing students' English proficiency not only improves their academic performance but also prepares them for global career opportunities. These practical applications align with the university's goal of increasing the percentage of on-time graduates and fostering academic excellence. By leveraging these findings, Halu Oleo University can implement evidence-based strategies to improve both academic and institutional outcomes.

While this study provides valuable insights into the prediction of timely graduation rates among Halu Oleo University students, there are several limitations that should be acknowledged:

1. The analysis relies heavily on the variables available within the university's academic information system. This dependence may omit external factors, such as socio-economic conditions or external learning resources, which could also significantly impact student graduation outcomes.

2. Although oversampling was used to address the class imbalance, it introduces a risk of overfitting the minority class, as synthetic data points may overly represent certain patterns in the training data. This could affect the model's performance on unseen data.

3. This study primarily focused on CART and Random Forest methods. While Random Forest showed robust results, other advanced machine learning models, such as Gradient Boosting Machines (e.g., XGBoost or LightGBM), may offer alternative approaches with potentially better performance. Future research could explore these models.

By acknowledging these limitations, future studies can expand on this work by incorporating more diverse datasets, exploring additional predictive variables, and testing advanced modeling techniques to further enhance the generalizability and robustness of the findings.

ACKNOWLEDGEMET

This work was supported by LPDP Scholarship under Kementerian Keuangan Republik Indonesia. Thank you for Halu Oleo University to give permission to use the data in this research.

REFERENCES

[1] L. Breiman, Classification and Regression Trees. Monterey, CA: Wadsworth International Group, 1984.
[2] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

[4] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann, 2011.
[5] Han, W. Wang, and B. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Advances in Intelligent Computing. Springer, 2005, pp. 878–887.
[6] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[7] Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.
[8] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.
[9] Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011.
[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.
[11] X. Wang, S. Smith, and T. Hyndman, "Improving CART with ensemble methods for imbalanced datasets," Journal of Statistical Computation and Simulation, vol. 89, no. 3, pp. 423–439, 2019.
[12] Indonesian Ministry of Education and Culture, "Standards for Higher Education Accreditation," BAN-PT, Jakarta, 2020.
[13] S. Safitri, B. Setiawan, and M. Kurniawan, "Applying CART for timely graduation prediction at Tanjungpura University," Indonesian Journal of Educational Research, vol. 8, no. 2, pp. 122–130, 2020.
[14] P. Dewi and E. R. Putri, "A study on student graduation prediction using CART at Pakuan University," Education and Data Mining Journal, vol. 5, no. 1, pp. 77–89, 2021.
[15] T. Abdullah, "Predicting student success using CART and Random Forest at Universitas Terbuka," Open Education Review, vol. 6, no. 4, pp. 215–230, 2021.
[16] Huang and Y. Li, "Using MissForest and XGBoost for timely graduation prediction," Educational Data Science Journal, vol. 4, no. 1, pp. 54–65, 2020.
[17] S. S. Rao and N. Gupta, "A comparison of classification methods for academic performance prediction," International Journal of Data Science and Analytics, vol. 7, no. 3, pp. 189–203, 2018.
[18] T. Raj and P. G. Nair, "Comparative analysis of classifiers for predicting academic performance," Journal of Machine Learning Applications, vol. 10, no. 2, pp. 95–104, 2019.
[19] Singh and V. Sharma, "Random oversampling techniques for improving Random Forest accuracy," International Journal of Computer Applications, vol. 183, no. 36, pp. 21–28, 2022.
[20] B. Fernandez, A. G. Ruz, and D. B. Huertas, "Random forest and oversampling techniques to improve classification performance," Knowledge-Based Systems, vol. 110, pp. 106–113, 2016.
[21] M. Kuhn and K. Johnson, Applied Predictive Modeling, Springer, 2013.
[22] Zhang and I. Y. Song, "Categorical-Boosted Trees for highly imbalanced datasets," in Proceedings of the 2018 ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), London, 2018, pp. 1077–1086.
[23] Rokach and O. Maimon, Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing, 2008.
[24] Patel, R. Bhatt, and P. Shah, "Machine learning approaches for academic performance prediction," Applied Artificial Intelligence Journal, vol. 29, no. 8, pp. 763–777, 2020.
[25] R. Johnson and J. K. Miller, "Using XGBoost to predict graduation time," Data Science in Education Review, vol. 5, no. 2, pp. 202–212, 2021.
[26] S. K. Verma and M. K. Sharma, "Improving Random Forest performance on imbalanced educational datasets," Machine Learning in Education Journal, vol. 12, no. 3, pp. 155–170, 2021.
[27] K. Kumar and S. Ahuja, "Using ensemble methods to predict academic success," International Journal of Educational Data Mining, vol. 8, no. 4, pp. 95–110, 2022.