TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 1764 – 1770

A Critical Assessment of Balanced Class Distribution Problems: The Case of Predict Student Dropout

Siti Mutrofin[#], Raden Venantius Hari Ginardi[*1], Chastine Fatichah[*2], Arrie Kurniawardhani[+]

[#] *Department of Information Systems, University Pesantren Tinggi Darul Ulum, Kompleks Ponpes Darul Ulum Peterongan, Jombang, 61481, Indonesia*

E-mail: sitimutrofin@ft.unipdu.ac.id

**Department of Informatics, Institut Teknologi Sepuluh Nopember, Jl. Teknik Kimia Sukolilo, Surabaya, 60111, Indonesia*

E-mail: [1]hari@its.ac.id;[2]chastine@cs.its.ac.id

[+] *Indonesia Department of Informatics, Universitas Islam Indonesia, Jl. Kaliurang Km. 14.5, Yogyakarta, 55584, Indonesia*

E-mail: arrie.kurniawardhani@uii.ac.id

Article Info
Volume 81
Page Number: 1764 – 1770
Publication Issue:
November-December 2019

Abstract
The general objective of this study is to help universities find the most influence factors which causes students drop out. The specific objective is to find the precise algorithm to predict dropout student in balanced class distribution case. Dataset was obtained from academic information system of a University in East Java, Indonesia. Data taken between 2009-2015 consists of 32 attributes, 425 data, and 2 classes. Type of data attributes are nominal and numerical. The results of this study state that the most influence factors which causes students to drop out are lecture programme; number of courses; credit amount in semester 3; credit amount in semester 6; credit amount in semester 9; Grade Point Average (GPA) in semester 2; GPA in semester 3; GPA in semester 4; and GPA in semester 6. Random Forest algorithm with gain ratio criteria parameter and shuffled sample method has the best performance, namely 99.29%, 99.47%, 9.09%, 99.28%, 0.71%, and 0.999 for accuracy, precision, recall, f-measure, classification error, and Area Under Curve (AUC), respectively. While the worst performance algorithm is Decision Tree with linear sampling method and information gain criteria, namely 83.19%, 83.47%, 86.32%, 84.87%, 16.81%, and AUC 0.3 for accuracy, precision, recall, f-measure, classification error, and AUC, respectively.

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *09 December 2019*

Keywords: *Balanced class distribution; Classification; Decision Tree; Dropout; Educational Data Mining; Random Forest.*

I. INTRODUCTION

Universities have many problems. One of them is the student dropout. This problem is experienced by many universities in all countries. There is no exception for Universities in East Java, Indonesia which have the highest dropout rate in 2017 [1]. Students dropout occur in each semester, and the amount is not small. The high level of students who dropout can affect the image of Universities in society.

Some reasons that influence students to stop study in university include academic matters, financial difficulities, motivational problems, personal considerations, dissastifaction with the college, military service, and a full time job [2]. Academic matters are related to bad grades, boredom, changes in career goals, and disability. This study will focus on the academic matters relating to grades supplemented by gender, age, lecture programme, year of batch, regional origin, last education background, etc. These data was chosen to utilize data that are already had by the university in academic information system (AIS). All this time, data in AIS have been ignored and continue to grow without being utilized. If the data is explored more deeply, there will be many knowledge that can be utilized by the University in many ways, such as minimizing the number of students drop out.

November-December 2019
ISSN: 0193-4120 Page No. 1764 – 1770

Data in AIS have several challenges including data redundancy, missing values, outliers, imbalanced class distribution, etc. Method that is commonly used to overcome complex problems in the education field is Educational Data Mining (EDM) [3][4]. This method will be used to explore data in the education field in order to obtain important knowledge that is still hidden. One of them is finding students who potentially stop studying so that they can be handled immediately at the beginning of the semester.

Research related to EDM that has been conducted was focused on predicting the freshman student attrition in imbalanced class distribution [5]. Other studies predict drop out students in the second semester and sixth semester using Artificial Neural Network (ANN) algorithms [6]. In this study students in each semester will be analyzed, because in fact in each semester there are several students who drop out. The pattern of students in each year of study have different characteristics.

This study has a balanced class distribution because the number of data in each class, graduating students class and dropout students classes, has almost the same amount namely 55% and 45%, respectively. For imbalanced class distribution, the amount of data in each class is quite far with a minimum ratio of around 65%:35% [7]. Many studies are interested in imbalanced class distribution problems [8][9]. So that this study proposes a study about prediction of students that have the potential to quit college by paying attention to Grade Point Average (GPA) and the number of Credit Units (SKS) in each semester using some of the most appropriate classification algorithms for balanced class distribution.

The general objective of this study is to find the most influence factors which causes students to drop out so and can predict early drop out students. In the hope that the university will have the opportunity to make a better improvement/policy so that the number of students dropping out is reduced without reducing the existing quality standards. The specific objective is to find the most precise classification algorithm in balanced class distribution problem to identify students who have potential to stop study in college.

II. Material and Method

This section will explain the stages of this research. The stages consist of data description, data analysis, classification algorithms, validation models, and evaluation models.

A. Data

The data in this study were obtained from academic information system (AIS) owned by a university in East Java, Indonesia in Information Systems, Informatics Engineering, Computer Science, or allied major. These majors were chosen because the percentage of students who dropout in these majors was higher than it in other majors. The data used are all student data from 2009 batch to 2015 batch. Data consists of 32 attributes with varied data types namely nominal and numerical. The number of data is 425 with a comparison of the amount of data in the Dropout class and the Graduated class is 45%:55%. It means that the distribution of this data in each class is balanced.

Data has a missing value problem. In Table 1, if the data in the Missing Value column is "yes" it means no data value is stored, and vice versa. The missing value problem is overcome by imputation by giving a value of "zero (0)" for numeric types, and giving an "unknown" value for nominal data types.

The data description is shown in Table 1. Gender were chosen as attribute because in general people tend to consider that major related to informatic was dominated by men. So that it was necessary to prove whether men have more potential to quit college or conversely they are more tough to complete their studies. Lecture programs are grouped based on the implementation of lectures, whether extension, regular, or transfer. In Indonesia, in general, regular students only focus on learning and joining organizations, while most extension students not only focus on studying but also focus on their family and work because they are married.

TABLE I
SUMMARY OF ATTRIBUTES USED

Attributes	Data Types	*Missing Value*
Gender	Binomial	No
Lecture Programme	Polynomials	No
Age when register	Integer	Yes
Origin	Polynomials	Yes
Year of graduation	Polynomials	Yes
Type of last education	Polynomials	Yes
Last education status	Polynomials	Yes
Major of last education	Polynomials	Yes
Marital status	Polynomials	Yes
Year of batch	Polynomials	No
History of leave of absence	Polynomials	No
Number of courses	Integer	No
Credit unit semester 1	Integer	Yes
Credit unit semester 2	Integer	Yes
Credit unit semester 3	Integer	Yes
Credit unit semester 4	Integer	Yes
Credit unit semester 5	Integer	Yes
Credit unit semester 6	Integer	Yes
Credit unit semester 7	Integer	Yes
Credit unit semester 8	Integer	Yes
Credit unit semester 9	Integer	Yes
Credit unit semester 10	Integer	Yes
GPA semester 1	Real	Yes
GPA semester 2	Real	Yes
GPA semester 3	Real	Yes
GPA semester 4	Real	Yes
GPA semester 5	Real	Yes
GPA semester 6	Real	Yes
GPA semester 7	Real	Yes
GPA semester 8	Real	Yes
GPA semester 9	Real	Yes
GPA semester 10	Real	Yes

So it needs to be investigated who can survive whether regular students or extension students to complete the study. Age also being observed to find out whether students who go on to college immediately after graduating from previous education will be more enthusiastic in completing their studies at university or conversely they are not focused on studying because they try to find better universities. Origin of the region was chosen to find out whether students who came from the city were better because they have supportive facilities to develop Science and Technology or conversely

November-December 2019
ISSN: 0193-4120 Page No. 1764 – 1770

students from villages are more able to survive because they are far from promiscuity and focus on fighting for their dreams in improving their lives. Years of graduation are observed to find out which students have greater potential to graduate whether fresh graduate students (from the previous level of education) or students who have applied for leave or students who have experienced work. Type of last education is observed to find out whether vocational or public or religious school graduates that is more potential to finish study. Last education statusis observed to find out whether students graduating from public schools are more likely to survive than them from private schools. Major of last education is observed to find out whether science major would be more survive than social humanities major. Married status is observed to find out whether marriage affects them to survive or not. Year of batch is observed to determine whether the quality of education influences students to stop studying, because each year the quality of lectures is different. In private universities, the frequency of lecturer alteration is very high. Student quality is influenced by the quality and enthusiasm of the lecturer. If the lecturer has the motivation to educate optimally and encourage students to get the achievement, students who stop studying will be reduced. For example, students in 2009-2012 had a very high percentage of dropping out of college because at that time, many lecturers had not yet obtained a master's degree. the majority of lecturers is young lecturers. They have not yet familiar with research activities and not know much about how to motivate students so students can be actively involved in competition, etc. The history of leave of absence is observed to find out whether it become the characterizes of students who intend to drop out of school, by taking leave first and will never be active again. The number of courses, credits, and GPA were observed to determine whether these become the characterizes of students who drop out.

B. Classification Algorithm

The classification algorithm used is popular classification algorithm and able to handle many type of data, such as numerical, binomial, and polynomial. These algorithms include Gradient Boosting (GB), Random Forests (RF), Logistic Regression (LOG), Naive Bayes (NB), Deep Learning (DL), Decision Tree (DT), and k-Nearest Neighbor (kNN). These algorithms were chosen because they proved to have good or bad performance in cases of imbalanced class distribution. This study observes whether the algorithm will provide the same performance in the case of balanced class distribution.

Algorithm 1. Gradient Boosting Algorithm[11]

Inputs:

- Input data $(x, y)^N i = 1$
- number of iterations M
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

1: initialize $\hat{f}_0$ with a constant
2: **for** $t = 1$ to M **do**
3: compute the negative gradient $g_t(x)$
4: fit a new base-learner function $h(x, \theta_t)$
5: find the best gradient descent step-size ρ_t:

$$\rho_t = arg\ min_\rho \sum_{i=1}^{N} \Psi\left[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t\right]$$

6: update the function estimate:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho h(x, \theta_t$$

The basic idea of the Gradient Boosting algorithm is a proposal from Freund and Schapire in the case of classification. This idea is further complemented by a number of analyzes by Breiman that make fundamental observations that AdaBoost Freund and Schapire are actually gradient-descent type algorithms in function spaces which were later developed by Friedman [10].

Gradient Boosting algorithm is an ensemble algorithm that based on supervised learning which focuses on regression and classification case in Data Mining [11]. Gradient Boosting algorithm developed by Friedman is presented in Algorithm 1. x=(x_1,…,x_d) is an attribute while y is a label [11][12]. The Boosting Gradient Algorithm on imbalanced class distribution problems has a very good performance, but Random Forests performs better [8]. Gradient Boosting is built by building the trees one by one which new trees will improve the previous tree performance.

Algorithm 2. Random ForestsAlgorithm[11][17]

To generate *c* classifiers:
for*i* = 1 to *c* **do**
Randomly sample the training data *D* with replacement to produce D_i
Create a root node, N_i containing D_i
Call BuildTree(N_i)
end for
BuildTree(*N*):
if *N* contains instances of only one class **then**
return
else
Randomly select *x*% of the possible splitting features in *N*
Select the feature *F* with the highest information gain to split on
Create *f* child nodes of *N* , N_1 ,..., N_f , where *F* has *f* possible values (F_1 , … , F_f)
for i = 1 to f **do**
Set the contents of N_i to D_i , where D_i is all instances in *N* that match
F_i
Call BuildTree(N_i)
end for
end if

Random Forests algorithm proposed by Breiman[13]. Just like Gradient Boosting, Random Forests is an ensemble algorithm that can be applied to regression and classification problems. Random Forests works by building a random set of trees that are useful for creating a model from training data [14]. Random Forests on imbalanced class distribution problems have very good performance [8]. Random Forests algorithm developed by Breiman is presented in Algorithm 2.

Logistic Regression in this study is used for classification and also used to analyze the most influence attributes which

November-December 2019
ISSN: 0193-4120 Page No. 1764 – 1770

causes students to drop out. Logistic Regression algorithm on imbalanced class distribution problems has a pretty good performance [8]. Logistic Regression method is very popular in statistics, computer science, mathematics, etc[15]. Logistic Regression method is one method based on statistics that describes the relationship between attribute (x) and class (y) [16]. The classes of data In this study are Graduation and Exit or Dropout class. Logistic Regression has a better performance than C4.5, QDA, Lin LS-SVM, and kNN for imbalanced class distribution issues [8].

Naïve Bayes is often used for imbalanced class distribution problems, but has less good results compared to Deep Learning [18]. The advantages of Naïve Bayes compared to Deep Learning in terms of computing time is fast. Recently, Deep Learning has become a fairly popular algorithm. Deep Learning is not something new, because Deep Learning is the development of Artificial Neural Network by giving more layers, so that it requires quite a lot of computing time. Deep Learning is very suitable for large and high dimensions data [18].

C. Validation Model

The validation model used in this study is 10-fold cross-validation with all types of sampling, namely Linear sampling, Shuffled sampling, and Stratified sampling. 10-fold cross-validation means that data will be divided into ten equal parts, which each part will be used as training data and test data alternately. The illustration of 10-fold cross-validation can be seen in Table 2.

TABLE II
VALIDATION MODEL OF 10-FOLD CROSS-VALIDATION [9]

***n*-validation**	**Dataset's partition**									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Linear sampling is data without randomization. Then data is divided directly into 10 equal parts. This allows unbalanced class distribution to occur between training data and testing data. For example, class A is data with a sequence of 1 to 50, class B is data with a sequence of 51 to 100, and class C is data with a sequence of 101 to 150, as in Iris data [19]. Shuffled sampling is data that is randomized first. Then the data is divided into 10 equal parts, but without paying attention to class distribution in training data and testing data. This allows unbalanced class distribution to occur, although the percentage of imbalanced class distribution is smaller than Linear Sampling's. Stratified sampling is data that is randomized first. Then the data is divided into 10 equal parts by considering the class distribution between training data and testing data. So that testing data and training data are balanced class distribution.

D. Evaluation Model

TABLE III
AUC VALUE, MEANING AND SYMBOL [9]

AUC	**Meaning**
0,9 - 1	Excellent classification
0,8 – 0,9	Good classification
0,7 – 0,8	Fair classification
0,6 – 0,7	Poor classification
< 0,6	Failure

Evaluation model used in this study is Area Under Curve (AUC) and confusion matrix can be seen in Table 3. AUC is used to know the performance of classification algorithm, is algorithm, whether the algorithm includes good classification or not. Confusion matrix used to get Accuracy (A), Precision (P), and Recall (R) values, because AUC is not enough, especially for the case of balanced class distribution and imbalanced class distribution [5].

III. RESULTS AND DISCUSSION

To see the algorithm performance that the most suitable for this data, 93 tests were carried out. The first testing scenario was conducted with 21 kNN trials based on sample types and the number of k = 1, 5, 10, 20, 25, 50, and 100. The optimal k obtained is 1 for all types of samples. kNN has good accuracy and precision in small k value, while kNN will have good recall and classification performance when the value of k is large. This applies to the case of balanced class distribution shown in the Shuffled sampling and Stratified sampling. kNN has poor performance when the k value is high for all types of samples. The testing results of kNN can be seen in Table 4.

TABLE IV
THE TESTING RESULTS OF KNN

Sample	***k***	***A* (%)**	***P* (%)**	***R* (%)**	**AUC**
Linear	**1**	97,18	**97,84**	97,01	**0,4**
	5	97,18	97,44	97,44	0,198
	10	**97,4**	97,05	98,29	**0,098**
	20	97,17	96,64	98,29	**0,098**
	25	97,17	96,64	98,29	**0,098**
	50	**96,23**	96,58	**96,58**	0,099
	100	96,7	**95,45**	**98,72**	**0,098**
Shuffled	**1**	**98,1**	**98,78**	97,87	**0,5**
	5	97,64	97,62	98,25	0,984
	10	97,64	97,62	98,25	0,987
	20	97,18	96,84	98,25	0,987
	25	97,18	96,84	98,25	0,987
	50	96,7	96,86	**97,46**	0,991
	100	**96,47**	**95,03**	**98,7**	**0,994**
Stratified	**1**	**98,11**	**98,71**	97,84	**0,5**
	5	97,64	97,49	98,28	0,986
	10	97,64	97,49	98,28	0,986
	20	97,17	96,66	98,28	0,987
	25	97,17	96,66	98,28	0,988
	50	96,69	96,62	**97,41**	0,991
	100	**96,46**	**95,06**	**98,71**	**0,99**

The second testing scenario was conducted with 24 Gradient Boosting trials with three sampling methods, two discrete probability distribution methods namely Bernoulli and Multinomial, four trials based on the number of trees namely 10, 25, 50, and 100. Table 5 shows that the number of trees influences Gradient Boosting performance for all types

TEST Engineering & Management

November-December 2019
ISSN: 0193-4120 Page No. 1764 – 1770

of samples and Bernoulli and Multinomial distributions. When the average number of trees is 25 or 50, it has the best performance whereas when the number of trees is 100 has the worst performance. In general, Multinomial distribution has the best performance, while Bernoulli has the worst performance, especially when it applied in balanced class distribution problem. Balanced class distribution data using Bernoulli and Multinomial distributions still have the best performance with 50 trees. The testing results of Gradient Boosting can be seen in Table 5.

The third testing scenario was conducted with 3 Naïve Bayes trials with three sampling methods. The testing results show that Shuffled sampling has the best performance, while Stratified sampling has the worst performance. This shows that Naïve Bayes is not suitable for balanced class distribution problem, although the difference is not too significant for each sampling method. The testing results of Naïve Bayes can be seen in Table 6.

TABLE V
THE TESTING RESULTS OF KNN

Sample	Distribusi	*Tree*	*A* (%)	*P* (%)	*R* (%)	AUC
Linear	Bernoulli	10	92,52	97,2	88,89	0,1
		25	93,91	96,85	91,88	0,1
		50	93	98,57	88,46	0,1
		100	92,29	99,51	86,32	0,1
	Multinomial	10	92,52	97,2	88,89	0,1
		25	95,32	96,52	94,87	0,1
		50	93	98,57	88,46	0,1
		100	92,53	99,51	86,75	0,1
Shuffled	Bernoulli	10	97,17	97,75	97,37	0,995
		25	97,64	97,75	98,23	0,999
		50	97,41	99,1	95,96	0,999
		100	96,93	100	94,42	0,999
	Multinomial	10	97,4	98,07	97,37	0,995
		25	97,64	97,75	98,23	0,999
		50	97,65	99,1	96,43	0,999
		100	96,93	99,47	94,8	0,999
Stratified	Bernoulli	10	98,11	98,71	97,84	0,996
		25	97,88	97,12	99,15	0,997
		50	98,58	99,58	97,84	0,999
		100	97,17	99,58	95,27	0,999
	Multinomial	10	98,11	98,71	97,84	0,997
		25	97,64	97,1	98,71	0,997
		50	98,58	99,58	97,84	0,999
		100	97,17	99,58	95,27	0,999

TABLE VI
THE TESTING RESULTS OF NAÏVE BAYES

Sample	A (%)	P (%)	R (%)	AUC
Linear	95,77	96,55	95,73	0,198
Shuffled	95,76	96,79	95,79	0,97
Stratified	95,75	96,55	95,72	0,972

The fourth testing scenario was conducted with 3 Logistic Regression trials with three sampling methods. The testing results show that Shuffled sampling has the best performance, while Linear sampling and Stratified sampling has bad performance. This shows that Logistic Regression is not suitable for imbalanced class distribution problem, although the difference is not too significant for each sampling method. The testing results of Linear sampling can be seen in Table 7.

TABLE VII
THE TESTING RESULTS OF LOGISTIC REGRESSION

Sample	*A* (%)	*P* (%)	*R* (%)	AUC
Linear	98,59	98,31	99,15	0,1
Shuffled	99,05	99,47	98,65	0,99
Stratified	98,36	98,77	98,3	0,99

The fifth testing scenario was conducted with 6 Deep Learning trials with three sampling methods and two discrete probability distribution methods namely Bernoulli and Multinomial. The testing results show that Shuffled sampling with Multinomial discrete probability distribution method has the best performance, while Linear sampling has the worst performance. This shows that Deep Learning is not suitable for imbalanced class distribution problem, although the difference is not too significant for each sampling method. The testing results of Deep Learning can be seen in Table 8.

TABLE VIII
THE TESTING RESULTS OF DEEP LEARNING

Sample	Distribusi	*A* (%)	*P* (%)	*R* (%)	AUC
Linear	Bernoulli	97,19	96,64	98,29	0,1
	Multinomial	97,18	96,64	98,29	0,1
Shuffled	Bernoulli	98,59	98,93	98,63	0,995
	Multinomial	99,06	99,18	99,02	0,994
Stratified	Bernoulli	97,64	97,55	98,26	0,996
	Multinomial	97,64	97,93	97,86	0,996

The sixth testing scenario was conducted with 24 Random Forest trials with three sampling methods, four methods of object separation criteria, two voting strategy methods namely Confidence Vote and Majority Vote. The testing results show that Shuffled sampling with Gain Ratio criteria has the best performance, while Linear sampling with Information Gain criteria has the worst performance. Two voting strategy methods namely Confidence Vote and Majority Vote have no difference. This shows that Random Forest is not suitable for imbalanced class distribution problem. The testing results of Random Forest can be seen in Table 9.

TABLE IX
THE TESTING RESULTS OF RANDOM FOREST

Sample	Kriteria	*Vote*	*A* (%)	*P* (%)	*R* (%)	AUC
Linear	Gain Ratio	Convidence	97,41	97,45	97,86	0,1
		Majority	97,41	97,45	97,86	0,1
	Information Gain	Convidence	96,94	97,42	97,01	0,1
		Majority	96,94	97,42	97,01	0,1
	Gini Index	Convidence	97,41	97,85	97,44	0,1
		Majority	97,41	97,85	97,44	0,1
	Accuration	Convidence	96,94	97,02	97,44	0,1
		Majority	96,94	97,02	97,44	0,1
Shuffled	Gain Ratio	Convidence	99,05	99,1	99,09	0,999
		Majority	99,29	99,47	99,09	0,999
	Information Gain	Convidence	98,12	98,76	97,77	0,998
		Majority	98,12	98,76	97,77	0,998
	Gini Index	Convidence	98,35	98,76	98,25	0,999
		Majority	98,35	98,76	98,25	0,999
	Accuration	Convidence	98,12	98,3	98,25	0,998
		Majority	98,12	98,3	98,25	0,998
Stratified	Gain Ratio	Convidence	98,83	99,2	98,71	0,998
		Majority	98,83	99,2	98,71	0,998
	Information Gain	Convidence	98,12	98,37	98,3	0,998
		Majority	98,12	98,37	98,3	0,998
	Gini Index	Convidence	98,35	98,37	98,71	0,998
		Majority	98,59	98,37	99,15	0,998
	Accuration	Convidence	98,35	98,37	98,71	0,998
		Majority	98,35	98,37	98,71	0,998

The seventh testing scenario was conducted with 12 Decision Tree trials with three sampling methods and four methods of object separation criteria. The testing results show that Stratified sampling has the best performance, while Linear sampling has the worst performance. Gain Ratio criteria are very suitable for imbalanced class distribution, while for the Gini Index and Accuration it is very suitable for balanced class distribution. This shows that Decision Tree is suitable for balanced class distribution problem [8]. The testing results of Decision Tree can be seen in Table 10.

The algorithm that has the best performance in balanced class distribution case is Random Forest and Gradient Boosting as shown in Table 11 with blue fonts. While the algorithm that has the worst performance in balanced class distribution case is Naïve Bayes shown in Table 11 with red fonts. The algorithm that has the best performance in imbalanced class distribution case is Logistic Regression as shown in Table 11 with blue fonts. While the algorithm that has the worst performance in imbalanced class distribution case is Decision Tree shown in Table 11 with red fonts.

TABLE X
THE TESTING RESULTS OF DECISION TREE

Sample	Kriteria	*A* (%)	*P* (%)	*R* (%)	AUC
Linear	Gain Ratio	92,53	97,64	88,46	0,2
	Information Gain	83,19	83,47	86,32	0,3
	Gini Index	91,59	96,26	88,03	0,25
	Accuration	91,12	95,37	88,03	0,198
Shuffled	Gain Ratio	96,23	96,86	96,51	0,722
	Information Gain	97,41	98,46	96,89	0,873
	Gini Index	96,48	96,8	96,8	0,914
	Accuration	95,77	96,36	96,01	0,954
Stratified	Gain Ratio	96,93	98,33	96,14	0,916
	Information Gain	97,17	97,51	97,41	0,812
	Gini Index	97,87	97,92	98,28	0,926
	Accuration	97,41	98,33	97,01	0,973

TABLE XI
THE TESTING RESULTS BASED ON THE TYPE OF SAMPLE SELECTION

Metode	*Sample*	*Acccuracy* (%)	*Precision* (%)	*Recall* (%)	AUC
kNN	Linear	97.18	97.84	97.01	0.4
	Shuffled	98.1	98.78	97.87	0.5
	Stratified	98.11	98.71	97.84	0.5
GB	Linear	95.32	96.52	94.87	0.1
	Shuffled	97.65	99.1	96.43	0.999
	Stratified	98.58	99.58	97.84	0.999
NB	Linear	95.77	96.55	95.73	0.198
	Shuffled	95.76	96.79	95.79	0.97
	Stratified	95.75	96.55	95.72	0.972
LOG	Linear	98.59	98.31	99.15	0.1
	Shuffled	99.05	99.47	98.65	0.99
	Stratified	98.36	98.77	98.3	0.99
DL	Linear	97.19	96.64	98.29	0.1
	Shuffled	99.06	99.18	99.02	0.994
	Stratified	97.64	97.55	98.26	0.996
RF	Linear	97.41	97.45	97.86	0.1
	Shuffled	99.29	99.47	99.09	0.999
	Stratified	98.83	99.2	98.71	0.998
DT	Linear	91.59	96.26	88.03	0.25
	Shuffled	96.48	96.8	96.8	0.914
	Stratified	97.87	97.92	98.28	0.926

The most influence attributes which causes students to stop study in university are obtained from the results of analysis using Logistic Regression with the Backward Wald method in SPSS software. The analysis results show that the most influential attributes based on a significant value of less than an alpha value of 10% were the lecture program (Regular, Extension, and Transfer) with a comparison between those who drop put and those who passed each program is 24,18%:75,82%, 57,5%:42,5%, dan 9,1%:90,9%; Number of courses, the more number of courses taken, the less likely students drop out; credit amount in semester 3, semester 6, and semester 9; GPA in semester 2, semester 3, semester 4, and semester 6.

IV. CONCLUSIONS

Based on experiment results in this study, Random Forests and Gradient Boosting algorithm have the best performance for balanced class distribution case, as shown in Table 11. Random Forests and Gradient Boosting algorithm achive the best result when using Shuffled sampling dan Stratified sampling. Random Forests and Gradient Boosting algorithm are also suitable for imbalanced class distribution case [8]. Naïve Bayes algorithm have the worst performance for balanced class distribution case. Naïve Bayes algorithm get the worst performance when using Linear sampling. It means in that experiment data distribution of each class between training and testing data is imbalanced, because in this study Dropu Out class is at the 191 first orders and the rest is Graduating class. This confirms Brown's research that Decision Tree is not suitable for imbalanced class distribution problems [8]. The imbalanced class distribution problem can be solved using Logistic Regression algorithm.

The University can give instructions to Academic Advisor Lecturers to pay more attention to and direct their students in each semester. In addition, Academic Advisors are expected to monitor their students in terms of the number of courses taken, Grade Point Average, the credit amount taken, and selected lecture programme in order to minimize the number of dropping out students.

ACKNOWLEDGMENT

Appreciation and thanks to the Directorate of Research and Community Service; Director General of Strengthening Research and Development; Ministry of Research, Technology and Higher Education; who funded the Inter-Higher Education Collaboration Research in 2018 with the title "Educational Data Mining Based on Classification to Analyze Students Who Potentially Stop Studying at University in Imbalanced Class Distribution problem".

REFERENCES

[1] P. I. D. Kemenristekdikti, "Statistik Pendidikan Tinggi Tahun 2017," Kemenristekdikti, Jakarta, 2017.

[2] R. Ricky and I. Nurarya, "Identitas korporat dan pengaruhnya terhadap daya tahan studi melalui reputasi dan kepuasan mahasiswa angkatan

2015 Institut Bisnis dan Informatika Kwik Kian Gie," *Jurnal Ekonomi Perusahaan*, vol. 22, no. 2, pp. 160-188, 2015.

[3] S. Pal, "Mining Educational Data to Reduce Dropout Rates of Engineering Students," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 2, pp. 1-7, 2012.

[4] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601-618, 2010.

[5] D. Thammasiri, D. Delen, P. Meesad and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Systems with Applications*, vol. 41, no. 2, pp. 321-330, 2014.

[6] M. M. Hidayat, R. H. Ginardi and D. Purwitasari, "Analisis Prediksi DO Mahasiswa dalam Educational Data Mining menggunakan Jaringan Syaraf Tiruan," *Jurnal IPTEK*, vol. 17, no. 2, pp. 109-119, 2013.

[7] H. Li and J. Sun, "Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples–Evidence from the Chinese hotel industry," *Tourism Management*, vol. 33, no. 3, pp. 622-634, 2012.

[8] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, p. 3446–3453, 2012.

[9] R. S. Wahono, N. S. Herman and S. Ahmad, "A comparison framework of classification models for software defect prediction," *Advanced Science Letters*, vol. 20, no. 10-12, pp. 1945-1950, 2014.

[10] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[11] G. Biau, B. Cadre and L. Rouvière, "Accelerated Gradient Boosting," *ARXIV*, pp. hal-01723843, 2018.

[12] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, pp. 1-21, 2013.

[13] N. Sirikulviriya and S. Sinthupinyo, "Integration of rules from a random forest," in *In International Conference on Information and Electronics Engineering*, Singapore, 2011.

[14] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[15] I. D. Id, A. Astried and E. Mahdiyah, "Deteksi dan Perbandingan Kinerja Algoritma Random Forest dan Boosted C5.0," *Jurnal Teknologi dan Sistem Informasi*, 2017.

[16] N. Mohd and Y. Yahya, "A Data Mining Approach for Prediction of Students' Depression Using Logistic Regression And Artificial Neural Network," in *The 12th International Conference on Ubiquitous Information Management and Communication*, Langkawi, 2018.

[17] R. O. Samosir, Y. Wilandari and H. Yasin, "Perbandingan Metode Klasifikasi Regresi Logistik Biner Dan Radial Basis Function Network Pada Berat Bayi Lahir Rendah (Studi Kasus: Puskesmas Pamenang Kota Jambi)," *Jurnal Gaussian*, vol. 4, no. 4, pp. 997-1005, 2015.

[18] A. A. Amri, A. R. Ismail and A. A. Zarir, "Comparative Performance of Deep Learning and Machine Learning Algorithms on Imbalanced Handwritten Data," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, pp. 258-264, 2018.

[19] R. Fisher, "Iris Data Set," UCI Machine Learning Repository, Irvine, 1936.