Xplore, 2013, *Vol. 2(1):e6(1-5)* © 2013 Departemen Statistika FMIPA IPB

Identification of Affecting Factors on the GPA of First Year Students at Bogor Agricultural University Using Random Forest

Sarah Putri*, Asep Saefuddin*, Bagus Sartono*
*Department of Statistics, Bogor Agricultural University

***Abstract*—Academically, the success of a student can be demonstrated by Grade Point Average (GPA). The performance of a student can be seen from the academic achievements, potential and motivation from themselves. Success in obtaining a high GPA can not be separated from the factors that affect the intellectual factor (the score of final examination in high school) and non-intellectual factors (gender, enrollment scheme to university, age when enrolled to university, senior high school status, etc). The data used for this research is the secondary data obtained from Student Affairs Directorate Bogor Agricultural University. Random forest method is an ensemble classifier using many decisions tree models. It can be used for classification or regression. This research is aimed to determine the size of the random forest and sample size of the explanatory variables that produces random forest with high prediction accuracy and stability that can identify the most important influential factors in student's academic achievements (GPA).**

I. Introduction

A. Background

Every new student of Bogor Agricultural University is required to follow the basic education program to learn basic general knowledge and skills for two semesters at the first year of study. In the first-year of study, students have to stay in university dormitory and early assessment of academic achievement starts from this period. The standard assessment of academic achievement is Grade Point Average (GPA).

Academic achievement is the result of lessons which is learned in school or college that is cognitive and usually determined through measurements and assessments. Achievements assessed during the education in colleges in the score of the subject assessment, assessment of the semester, assessment of the year end study and assessment of the end of the program study ([1]).

This study tried to investigate which factors influence the most important for first year students academic achievement. In teaching and learning process affected a number of factors that interact with each other in determine the process and learning outcomes. The success rate of students in the educational process is influenced by many factors. These factors can be grouped into intellectual factors, the ability of a person indicated by the intelligence and cleverness in thinking and non-intellectual factors, conditions from within and outside him or her self or surrounding environment, associated with a self in influencing ability to think and act ([2]). Random forest method is the analysis method used in this study because accuracy and variable importance information are provided with the result of random forest. Random forest is an ensemble classifier using many decisions tree models. It can be used for classification or regression

B. Objectives

The objectives of this research are to describe the student's characteristics of first year student Bogor Agricultural University batch 48 and to determine the size of the random forest and size of the explanatory variables that produce random forest with high prediction accuracy and stability that can identify the most important factors in student's academic achievements (GPA).

II. METHODOLOGY

A. Data Source

The data obtained in this study is a primary data from Student Affairs Directorate of Bogor Agricultural University, with total population of 3354 students. The response variable is the GPA of the first-year study, which is divided into two categories:

1 = GPA $\geq$ 3.00
2 = GPA $<$3.00

The set of the factor candidates consist of the following variables: gender, origin of student, enrollment scheme into Bogor Agricultural University, age when enrolled to Bogor Agricultural University, senior high school status, UAN SMA (total score of national final exam), parent's income, father's occupation, father's education level and mother's education level.

B. Methods

The methods used in this study are:

1) Collect and sort the data. From the data obtained, grouped the explanatory variables based on relevant literature.
2) Do an exploratory data analysis and identify a relationship between the student achievement and each of the explanatory variables.
3) Apply a random forest analysis to determine of effect of each explanatory variable. Previously, did a small simulation study to identify the best tuning parameters

for the random forest, which bootsrap sample size and the number of variables for node splitting.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. It is a substantial modification of bagging that builds a large collection of de-correlated trees, and then averages them. On many cases the performance of random forests is very similar to boosting, and they are simpler to train and tune. Random forest use of the strong law of large numbers shows that they always converge so overfitting is not a problem ([3]). The differences between random forest and bagging methods lie in the addition of random sub-setting stage before the formation of the tree ([4]). According to ([5]), the establishment of trees in random forest method does not use the whole total data set, but using roughly two thirds of the sample data set called the training data group and one third of it is used to calculate the value of one classification of tree establishment and estimates are significant variables, called a data group out of bag (OOB). ([6]) say random forest is used for classification. A random forest obtains a class vote from each tree, and then classifies using majority vote. The random forests algorithm is as the following:

a) Draw *n* bootstrap samples from the original data.
b) For each of the bootstrap samples, grow an unpruned classification tree, with the following modification at each node, rather than choosing the best split among all predictors with *m* (partially explanatory variables) $<$ *p* (total explanatory variables) and in addition, the inventors make the following recommendation that for classification, the default value for $m=\{\frac{\mathbf{1}}{\mathbf{2}}|\sqrt{\mathbf{p}}|, |\sqrt{\mathbf{p}}|, \mathbf{2}|\sqrt{\mathbf{p}}|\}$. This study does not follow the recommendation the default value for *m*. Values of m that is used in this study are 3, 4, 5 and 6. The formation of a single tree in a random forest algorithms use algorithms CART (Classification and Regression Tree) without pruning. CART is a nonparametric statistical methodology developed for classification analysis ([7]). CART will produce a tree regression if the variable responses numerically and generates the classification tree when the variable response was categorical. The formation of a single tree involves three things, namely the selection of the splits, the decisions when to declare a terminal node or to continue the splitting and the assignment of each terminal node to a class ([8]). The selection of the splits:
   At the selection of the splits stage, classifier has sought the most heterogeneous nodes. One technique used is the impurity using Gini index. Impurities using Gini index value at node *t*, *i(t)* can be written as follows:

$$i(t) = 1 - \sum_{j=1}^{q} \rho^2(j|t).$$

   Where *q* is class of response variable, $\rho(j|t)$ is the unit of observation opportunities in the *j*-th class of the node *t*.

$$\rho\,(j \mid t) = \frac{\pi_j \frac{N_j\ (t)}{N_j}}{\sum_{i=1}^{q} \pi_i \frac{N_i\ (t)}{N_i}}$$

   π_j is the proportion of objects that belong to class *j*, N_j is the number of observations to the classes *j*, and $N_j(t)$ is the number of units of observation are included in classes *j*, node *t*. Goodness of split is one of the evaluations of splitting by splits *s* at node *t* with the following formula:

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

   P_L is the proportion in observation t_L and P_R is the proportion in observation t_R. Splitting rules are defined by specifying a goodness of split function $\phi(s,t)$ defined for every $s \in S$ and node *t*. At every t, the split adopted is the split s^* which maximizes $\phi(s,t)$.

$$\triangle i(s*,t) = \max_{s \in \mathrm{S}} \triangle\ (s,t)$$

   The decisions when to declare a terminal node:
   Node *t* can be used as a terminal node if there is no significant heterogeneity reduction in sorting, have the minimum limit *n*, and limits on the number of levels or the maximum tree depth.
   The assignment of each terminal node to a class:
   Class label of the terminal node is determined by the rules of the greatest number (plurality).
c) Repeat Steps 1 and 2 for *k* times. Value of *k* is the numbers of single trees builded. Record the misclassification values for each *m* and *k* values tested. Values for *k* used in this study are 100, 500 and 1000. Analysis the misclassification rate uses explorative analysis.
d) Record the values of Mean Decrease Gini (MDG) on each *m* and *k* values tested.
   Mean Decrease Gini (MDG) is one measure used to see the importance of explanatory variables in the random forest method. At every split, one of the *m* -try variables is used to form the split and there is a resulting decrease in the Gini. The sum of all decreases in the forest due to a given variable, normalized by the number of

trees, forms the Gini measure. This measure is not as reliable as the margin measure above but it is automatically computed in every run of random forests ([5]).
The formula for Mean Decrease Gini (Sandri dan Zuccolotto 2006):

$$MDG_h = \frac{1}{k}\sum_t [d(h,t)]I(h,t)]$$

where *n* is the number of trees built, $d(h,t)$ is a large decrease in Gini index for the explanatory variables X_h at node *t* and $I(h,t)$ = 1 when X_h sorting node *t* and 0 in others

e) Interpretation of the results of the random forest output.

III. RESULT AND DISCUSSION

A. *Exploratory Data*

The number of students Bogor Agricultural University batch 48 are 3354 students. From the total of students in the first year study identified that 52.27% of the students achieve the GPA $\geq$ 3.00 and the rest 47.73% achieve GPA $<$ 3.00 (Figure 1). This suggests that a proportion of the number of students that achieve the GPA $\geq$ 3.00 and the students GPA $<$ 3.00 more or less the same. Most of the students that achieve the GPA $\geq$ 3.00 are female and from Java.

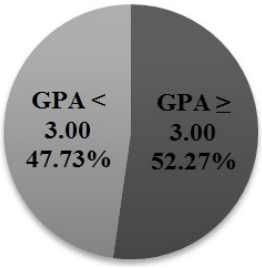

Figure 1. Distribution of student based on GPA

Number of female students in Bogor Agricultural University exceeds the male. In Figure 2(a) shows that 59.33% students are female and 40.67% students are male. Majority of the student are from Java. Bogor Agricultural University is also located in Java. There are 79.84% student from Java and 20.16% students from outside of Java (Figure 2(b)). Bogor Agricultural University had 6 schemes of enrollment in 2011. There were SNMPTN Undangan (National selection of college entrance by selection of score in Senior High School), SNMPTN UTUL (National selection of college entrance by written test), jalur prestasi (achievement), jalur beasiswa (scholarship), mahasiswa asing (international students) and UTMI (written test in Bogor Agricultural University). In Figure 3 shows that 67.89% of the student are from SNMPTN Undangan, as 17.50% of the student are

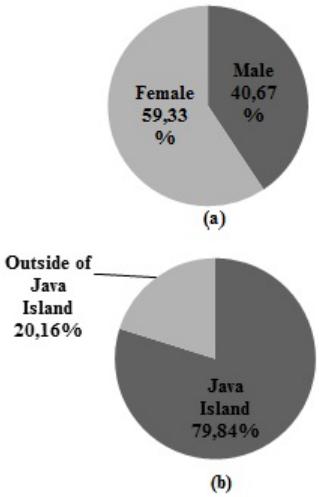

Figure 2. Distribution of student: (a) based on gender, (b) based on origin

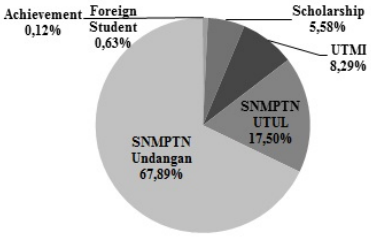

Figure 3. Distribution of student based on enrollment to Bogor Agricultural University

from SNMPTN UTUL, from UTMI there are 8.29% of the student, as 5.58% of the student are from jalur beasiswa, it 0.63% of the student that from mahasiswa asing and 0.12% of them that from jalur prestasi (Figure 3).

Variable age is not considered as a categorical variable. The age of students when they are starting to study in Bogor Agricultural University has mean and median of 18 years and 1 month. The youngest student at this batch has the age of 15 years and 8 months, while the oldest is 23 years and 1 month (Table I).

Table I
STATISTICS VALUES BASED ON AGE (YY.MM) OF THE STUDENTS

Statistics	Value(yy. mm)
Mean	18. 01
Median	18. 01
Minimum	15. 08
Maximum	23. 01

Senior high school status of the student is grouped as public school, full private school, subsidized school and foreign school. Figure 5 shows that most of students are from public school and just a few of student that from subsidized private school. Table 2 shows the statistics values

based on score of final exam and based on parental income. Variable total score of final exam in senior high school (UAN SMA) is also not considered as categorical data. Mean of the student's score of final exam is 51.17. Median of the student's score of final exam is 51.52. The minimum score is 33.00 and the maximum score is 58.45. Variable of parent's income is also not considered as categorical data. Table II shows that mean of the parental income are Rp. 4.432.410,00.-. Median of the parental income are Rp. 3.500.000,00-. The minimum of the parental income are Rp. 100.000,00.- and the maximum of the parental income are Rp. 90.000.000,00-.

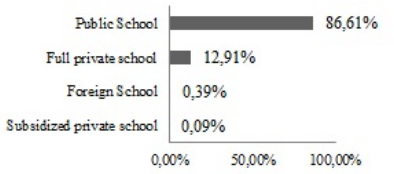

Figure 4. Distribution of student based on senior high school status

Table II
ADD CAPTION

Statistics	Values	Statistics	Values (Rp)
Mean	51.17	Mean	4432410
Median	51.52	Median	3500000
Minimum	33.00	Minimum	100000
Maximum	58.45	Maximum	90000000

Father's occupations are grouped as 5 categories. There are Labor/ Farmer/ Fisherman, civil servant, private employee/self-employed, professional and other. Figure 6 shows that most of the father's occupations are private employee/self-employed (45.29%), civil servant (30.11%), labor/farmer/fisherman (18.22%), other (5.40%) and professional (0.98%).

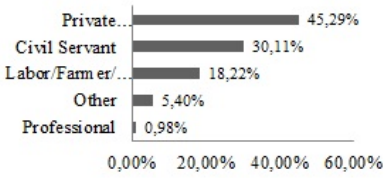

Figure 5. Distribution of student based on father's occupation

Father's education level and mother's education level grouped as 9 categories. There are not completed primary school, primary school, junior high school, senior high school, diploma, bachelor, undergraduate, master and doctor. Figure 6 shows father's education level and Figure 7 shows mother's education level. From the Figure 6 shows that the most father's education level of the student are senior high school (36.37%) and only a few of father's education level of the student are doctor (2.00%). Figure 7 shows that the most mother's education level of the student are senior high school (40.55%) and only a few of mother's education level are doctor (0.51%).

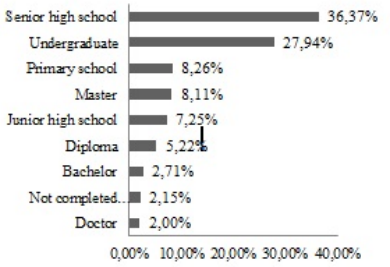

Figure 6. Distribution of student based on father's education level

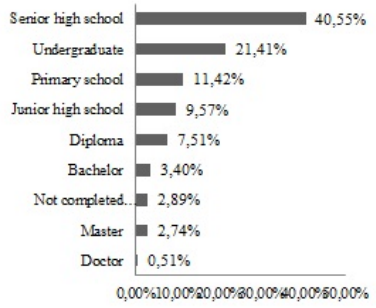

Figure 7. Distribution of student based on mothers education level

B. Random Forest Analysis

In this study, random forest analysis was used to to determine the size of the random forest and sample size of the explanatory variables that produces random forest with high prediction accuracy and stability that can identify the most important influential factors in student's academic achievements (GPA). Random forest prediction accuracy is measured from the level misclassification rate. High accuracy and stability can be obtained if the sample size explanatory variable *(m)* and the size of the random forest *(k)* are determined by the right value (Breiman01). A preliminary study was conducted to identify the best possible values of m and k The simulation done by implementing a random forest for every combination of all parameters. The combination which gave low and consistant misclassification rate would be considered as the good one. The values of m that used in this study are 3, 4, 5 and 6. The values of k that used in this study are 100, 500 and 1000.

Random forest is in the optimum condition when at certain m and k value produces the smallest value on misclassification rate. The changes of misclassification rate

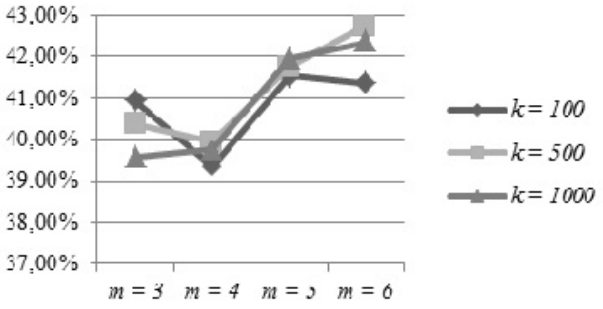

Figure 8. Random forest curve misclassification rate for m and k tested

depend of different *m* value shows on Figure 8. The lowest misclassification rate always reached on $m = 4$.

Figure 8 shows that $m = 4$ is the m optimal. It also shows that the *m* optimal is already unknown even when the value of *k* is small. The comparison of misclassification rate for every *m* and *k* tested shows in Table III.

Table III
COMPARISON OF MISCLASSIFICATION RATE RANDOM FOREST OF *m* AND *k* TESTED

m	*k*	Misclassification Rate
3	100	40.95%
	500	40.36%
	1000	39.56%
4	100	39.36%
	500	39.96%
	1000	39.76%
5	100	41.55%
	500	41.75%
	1000	41.95%
6	100	41.35%
	500	42.74%
	1000	42.35%

Table III shows the results of 12 times the formation of random forest. The smallest misclassification rate is on the $m = 4$ and $k = 100$. The misclassification rate is 39.36%. From the Table 3, known that the optimum random forest when are $m = 4$ and $k = 100$. Beside misclassification rate, the correct classification rate, sensitivity and specificity also known from random forest.

Optimum random forest trees used to get the best identifier variables. Determination of variable importance is by ranking criteria used Mean Decrease Gini (MDG) value of random forest when $m = 4$ and $k = 100$. Figure 9 shows the rank of variable importance. The most influential variable of the GPA of student at Bogor Agricultural University in the first year study is the enrollment scheme when enter Bogor Agricultural University. Gender and final score of exam in senior high school are the second and the third position of variable importance. The Mean Decrease Gini (MDG) of variable enrollment, gender and final score of exam have a slight difference value and as the stable variable.

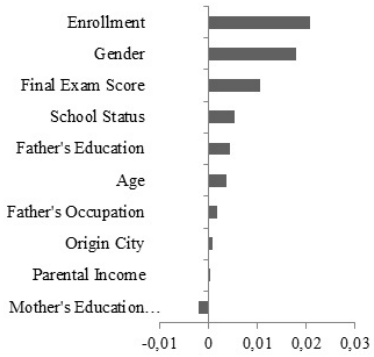

Figure 9. Mean Decrease Gini (MDG) values for m = 4 and k = 100

REFERENCES

[1] Institut Pertanian Bogor, *Panduan Program Pendidikan Sarjana)*, Bogor (ID): IPB Press, 2011.

[2] Ariwibowo MS, *Pengaruh Lingkungan Belajar terhadap Prestasi Belajar Mahasiswa PPKn Angkatan 2008/2009 Universitas Ahmad Dahlan semester Ganjil Tahun Akademik 2010/2011*, Jurnal Citizenship, (1)2, 2012.

[3] Breimen L, *Random Forests. Machine Learning*, 45:5-32, 2001.

[4] Sartono B, Syafitri UD, *Ensemble Tree: an Alternative toward Simple Classification & Regression Tree*, Forum Statistika dan Komputasi 15(1):1-7, 2010.

[5] Breiman L, Cutler A, *Manual on Setting Up, Using, and Understanding Random Forest V4.0.)*, JIAB [Online] [2013, May 15] tersedia pada : http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf, 2003.

[6] Hastie TJ, Tibshirani RJ, and Friedman JH, *The Elements of Statistical Learning: Data-mining, Inference and Prediction.Second Edition*, New York (US): Springer-Verlag, 2008.

[7] Sutton CD, *Classification and Regression Trees, Bagging, and Boosting*, Handbook of Statistics 24:303- 329, 2005.

[8] Breimen L, Friedman JH, Olshen RH, and Stone CJ, *Classification and Regression Trees*, New York: Chapman & Hall, 1984.