International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 5, Issue 3, 2019 Special Edition: *Science, Applied Science, Teaching and Education*

Factors Predicting Timely Student Graduation in the Faculty of Science and Technology at Airlangga University

Siti Maghfirotul Ulyah, Marisa Rifada, Elly Ana

Department of Mathematics, Airlangga University, Indonesia

Email: maghfirotul.ulyah@fst.unair.ac.id

Abstract

The aim of this study is to explore the pattern of student's period of study by predicting it based on some variables related to students and other variables associated with the study period. The data in this work was from the Faculty of Science and Technology (FST) undergraduate students starting from 2008 - 2018 from 8 subjects. Those are Mathematics, Physics, Chemistry, Biology, Statistics, Information System, Biomedical Engineering, and Environmental Engineering. The attributes in this study consist of subject, gender, address, high school status, national exam score, admission method, subject selection order, parents' income, ELPT, and GPA. The dependent variable (study period) is divided as on-time and not on-time. The method used in prediction is the Decision Tree with C4.5 algorithm. The results of this study gives information that address and ELPT are not associated with the study period while the most dominant attribute for the prediction is GPA, followed by gender.

Keywords: student, prediction, timely graduation, study period, Decision Tree.

Introduction

Education is an important part forming the human development index (HDI). The higher the level of education per capita in a country, the higher the HDI of that country. One level of education that plays an important role in advancing education in Indonesia is higher education organised by universities, both public and private. Higher education includes diploma programs, undergraduate programs, master programs, doctoral programs, and professional programs, as well as specialist programs (Indonesian Government Regulation No. 66, 2010).

Airlangga University (UNAIR) is one of the top 5 universities in Indonesia which holds the status of autonomy as a public university (PTN-BH). To improve its quality, UNAIR has always been active in assessment activities by the National Accreditation Board of Higher Education (BAN-PT). Although currently UNAIR is getting the best score "A", the quality of higher education must be maintained and improved. BAN-PT measures the quality of a college based on several standards which include students and graduates (BAN-PT, 2014). This study will focus on the graduates since the accuracy of student graduation is very influential in the assessment of the quality of a college. For the undergraduate program, the study load charged according to government regulations, which is 144 credits, with a maximum study period of 14 semesters or 7

years. Students are said to graduate on time if their study period is less than or equal to 4 years.

According to Abu and El-Halees (2012), utilising educational data as learning attributes is one way that can be taken to increase the quality of higher education. Educational data usually includes student background data and student academic achievement. UNAIR has an integrated information system in the "cyber campus" which the data will be utilised for this study. This information can be analysed to draw conclusions and be used as material for policy making to improve the quality of a university.

The attributes for prediction include data on Grade Point Average (GPA), English Language Prediction Test (ELPT), high school status, national exam score in high school, subject selection order, admission path, parents' income, gender, and also regional origin (address). In its efforts to go to a World Class University, UNAIR also has a special admission program for students from Eastern Indonesia (Papua). Therefore, the student's origin is thought to be one of the factors in the classification of study period.

By looking at the number of students and the period of establishment of UNAIR, the educational data that will be collected can be ascertained as a large dataset. Therefore, the right method for this analysis is data mining. Based on Goela and Chanana (2012), data mining is an automatic search process for useful information in large data storage places. Supervised learning is an approach where we have data to be trained, and there are targeted variables. Therefore, the

purpose of this approach is to group data into existing data. One method included in supervised learning is classification. In classification, the object is assigned to one of several predefined categories (Tan et al., 2014). In this study, the methods used in classification are discriminant models based on rule-based reasoning (Decision Tree).

Classification is part of data mining which is most often studied in statistics. It is a task to map each attribute set to one of the predefined class labels by learning a target function (classification model) (Tan et al., 2014). The classification method that will be used in this study is Decision Tree. This study is expected to provide useful information to UNAIR in improving its quality. In addition, the results of this study are also expected to be used as material for determining policies in improving the quality of institutions.

The main objectives in this study are to obtain an overview of student profile data at the Faculty of Mathematics and Natural Sciences UNAIR and to identify the pattern of student's period of study using Decision Tree classifier.

Previous work on the pattern of graduation status of FMIPA Islamic University of Indonesia students was conducted by Kesumawati and Waikabu (2017), Kesumawati & Utari (2018) using Naïve Bayes method and Support Vector Machine (SVM). The results of the study concluded that the highest classification accuracy was obtained from the SVM model with Linear Kernel and the parameter C = 0.1 which was equal to 69.51%. The variables used in the prediction include

department, gender, GPA, city of residence, high school type, high school program, and parent occupation. Our study adopts the same scheme with Kesumawati and Utari (2018) to study about graduates. However, the object of the study is different (UNAIR). Moreover, some additional attributes are presented in our study such as ELPT, High school national exam, department selection order, admission path, and parents' income.

A study about the student aid was conducted by Glocker (2011). The study concluded that student aid recipients had shorter period of study than comparable students without it, such as students who are supported by the equal amount of parental or private transfers only. Additionally, on average, greater financial aid did not influence the duration of study. Therefore, our study will consider students who receive a scholarship upon the admission and also the income of their parents.

Gibert et al. (2010) in his research entitled "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation" explained that the supervised learning method that is appropriate to explain or predict qualitative variables is the discriminant method. Decision Tree is one of discriminant method with rule-based reasoning. Hssina (2014) did a comparative study about the algorithms of Decision Tree (ID3 and C4.5) and concluded that C4.5 algorithm had better performance in classification. Therefore, our study will apply C4.5 algorithm of Decision Tree method.

This paper is structured as follows. Section 1 presents the background of this study. Then, the literature review is explained in section 2. Section 3 is the method of the study and Section 4 presents the classification result. Finally, conclusion is given in the last section.

Materials and Method

Data

In this study, the data were FST graduates starting from 2008 to 2018 with 1957 observations. These data are the data of undergraduate students in each subject in FST (Mathematics, Physics, Chemistry, and Biology). The data will be divided into training data and testing data. Training data are data that will be used to form a classification model, while testing data are used to validate classification results. These secondary data are obtained from the database owned by the FST academic sub-division of UNAIR.

Research Variables

The research variables consist of dependent and independent variables. The variables in this study are shown in Table 1.

Table 1. The variables in the study

Variable		Category
Dependent variable	Study Period	On time
		Not on time
Independent variable (attributes)	Subject	Mathematics
		Statistics
		Information System
		Physics
		Biomedical Engineering
		Chemistry
		Biology
		Environmental Engineering
	Gender	Male
		Female
	Address	East Jawa
		Outside East Java
		Outside Java
	High School	Public
		Private
	High School National Exam Score	Excellent
		Very Good
		Good
		Fair
	Admission Method	Achievement
		National test
		Scholarship
		Independent
	Subject Selection Order	I
		II
		III/IV
	Parents' Income	Low
		Below average
		Average

Variable		Category
		Above average
		High
	ELPT	A2
		B1
		B2/C1
	GPA	Satisfactory
		Very satisfactory
		Cumlaude

The Method

The preliminary method in this study is obtaining the descriptive statistics of the variables and investigating the relation among variables. Then, the main analysis, which is classification, is conducted using Decision Tree method. For the Decision Tree, we use C4.5 algorithm to obtain the best hierarchical tree. After getting the classification results, the interpretation and reasoning will be presented.

Classification with Decision Tree Method

Decision Tree classifier is a simple and popular classification technique. It is a hierarchical structure that consists of nodes and directed edges. A Decision Tree has a root node, internal nodes, and many leaf nodes in which each of them is assigned a class label. The root node and internal nodes contain attribute test condition to separate records with different characteristics (Tan et al., 2014).

From a set of attributes, there are many Decision Trees that can be built. Therefore, to find the optimal tree, the C4.5 algorithm by Quinlan (1993) is used in this work. In this algorithm, cases with unspecified values are ignored while computing the information content, and the information gain for an attribute is multiplied by the fraction of cases for which the value of the attribute is known.

At each tree node, C4.5 algorithm chooses an attribute of the data in which most effectively splits the set of samples into subsets enriched in a class or the other. The normalised information gain (entropy difference) becomes the criterion. It is resulting from choosing an attribute to split the data. The decision can be made by selecting an attribute with the highest normalised information gain (Hssina, 2014).

Another issue of a classification model is over fitting. The performance of a bigger tree may not as good as a simple or smaller tree construction. Therefore, to avoid this problem, generally Decision Tree algorithms employ a "pruning" method meaning that they build a large tree and then eliminate some portion of it. In C4.5, the method of pruning is based on estimating the error rate of each sub tree and replacing the sub tree with a leaf node when the estimated error of the leaf is lower (Salzberg, 1994).

Results and Discussion

Descriptive Statistics

The descriptive statistics of the variables in this study are presented in Figures 1 to 7. Figure 1 depicts the proportion of the students who graduate in no more than 4 years (timely) and students whose study period is more than 4 years (untimely). Based on Figure 1, most undergraduate students of Faculty of Science and Technology (FST) finished their study timely (72.3%).

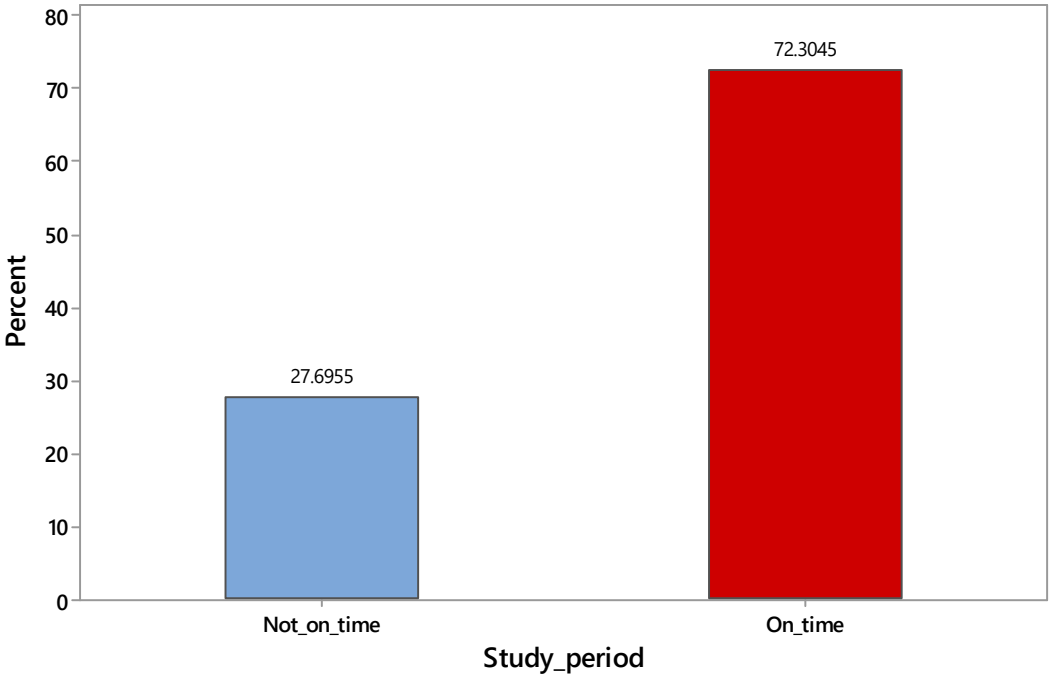


Figure 1. The bar chart of the dependent variable (study period)

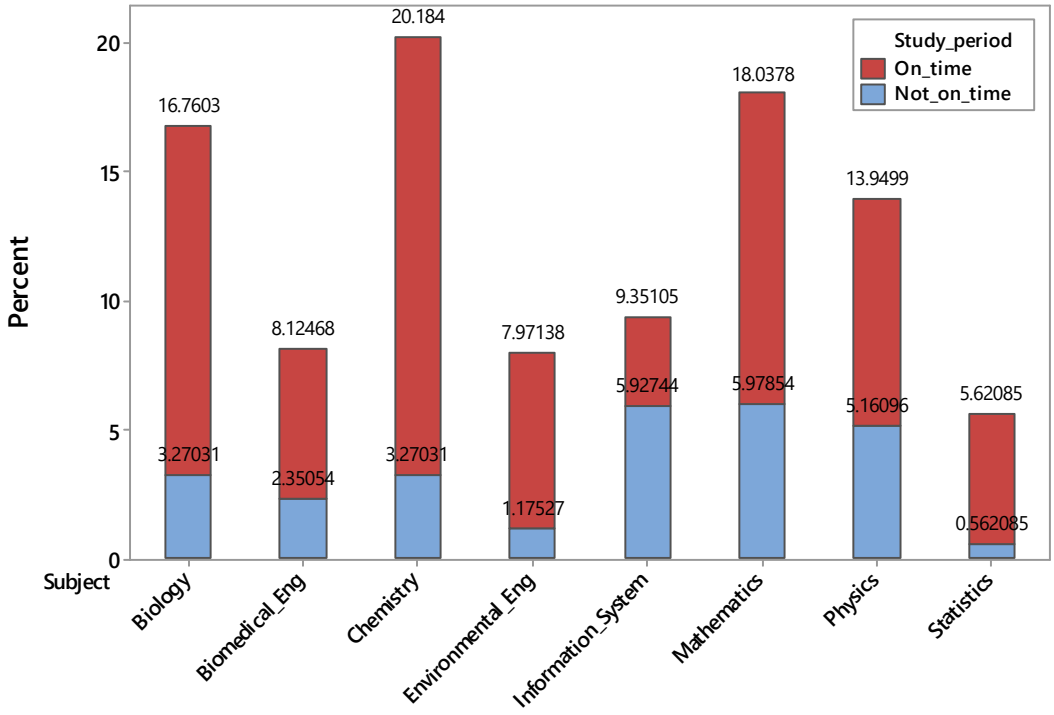
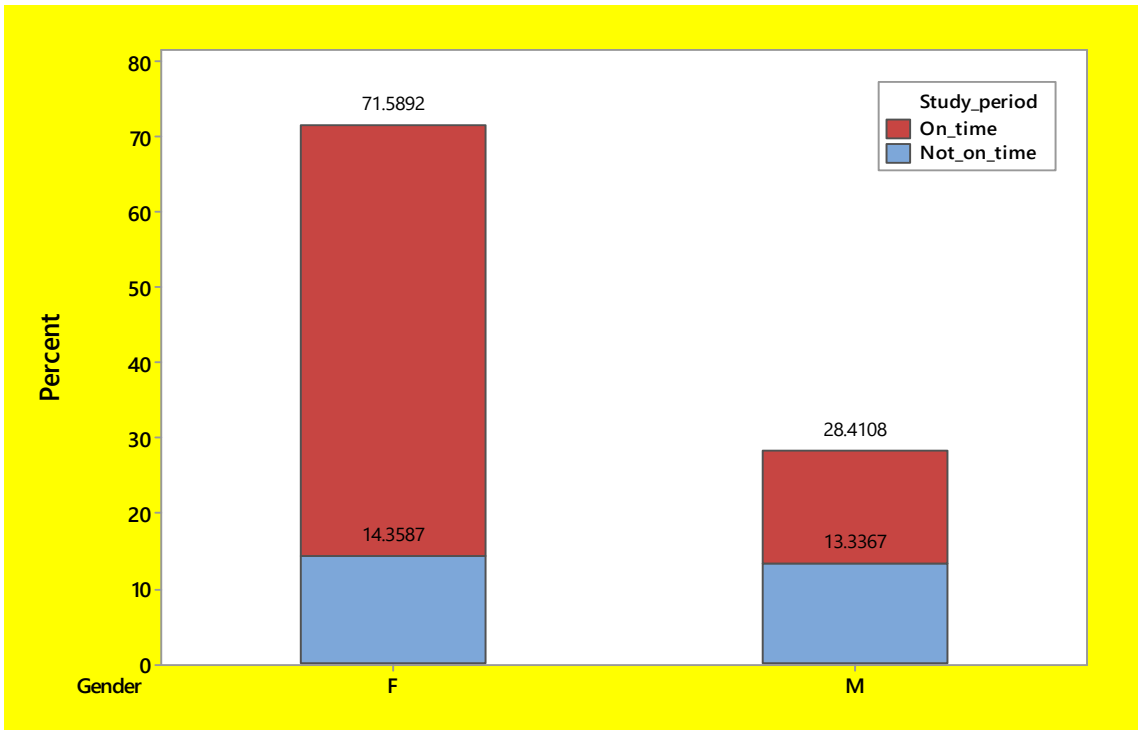


Figure 2. The bar chart of the subject

Figure 2 shows the subject based on the study period of FST student. The subject with highest proportion of on-time graduate is Statistics, followed by Environmental Engineering and Chemistry. On the other hand, Information system has highest percentage of late graduates, followed by Physics and Mathematics.



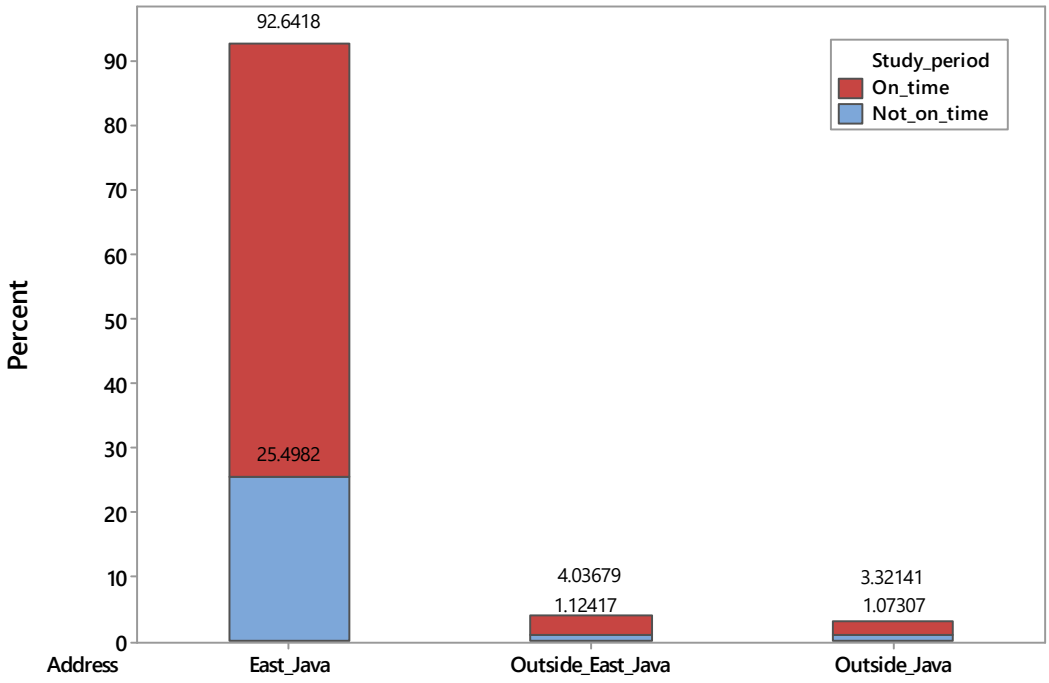
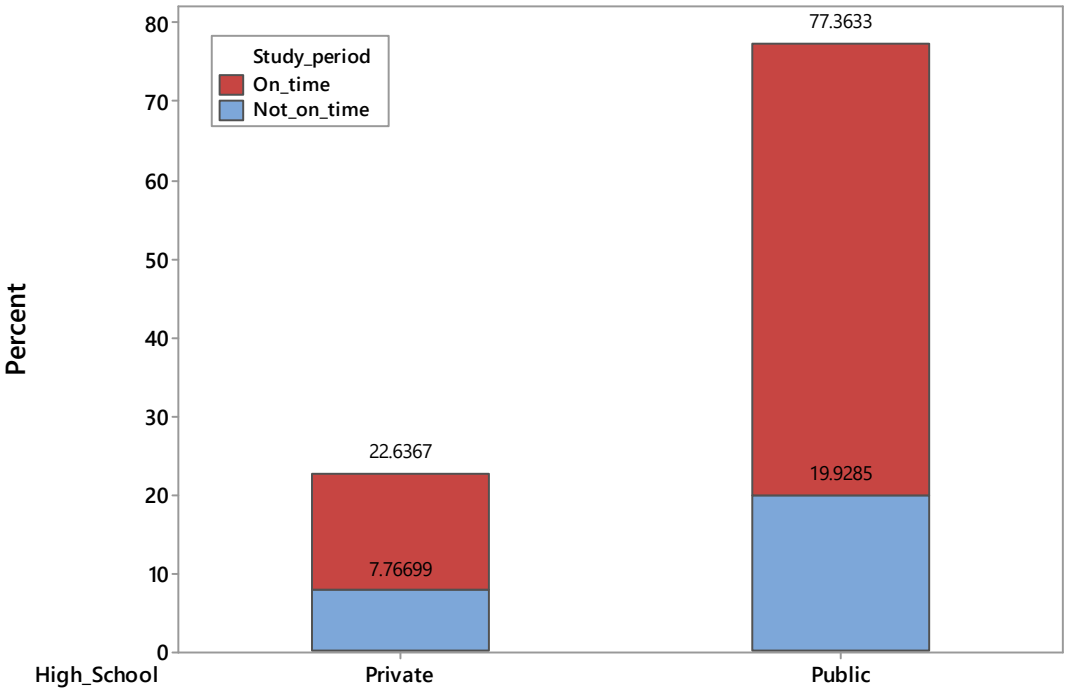


Figure 3. The bar charts of gender and address

Figure 3 gives information about the gender and address of FST graduates based on the period of study. Half of male students graduate on-time whereas female students tend to graduate on-time rather than not on-time. Moreover, students from East Java have a higher percentage of graduates who complete in no more than 8 semesters compared to those who come from outside East Java or Outside Java.



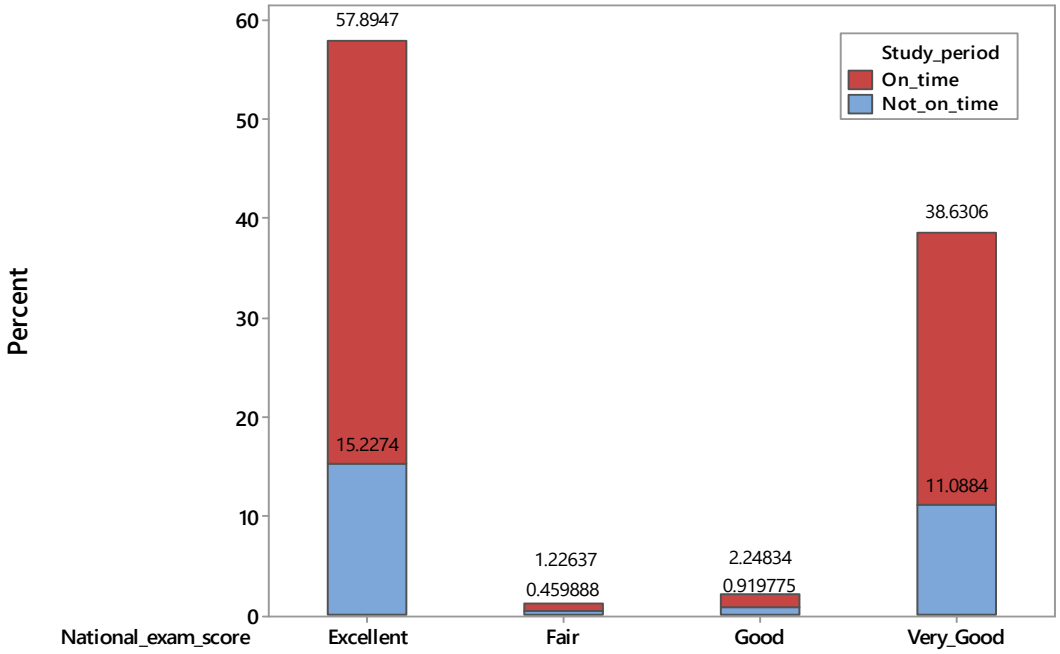


Figure 4. The bar charts of high school and national exam score

The status of the high school along with national exam score for high school student based on the period of study are given in Figure 4. Most of FST students come from public university. Furthermore, the percentage of on-time graduates from public university is greater than that of private university. However, the difference is quite small. Besides, students with excellent and very good national exam scores tend to come to FST and their percentage of graduating on-time is high.

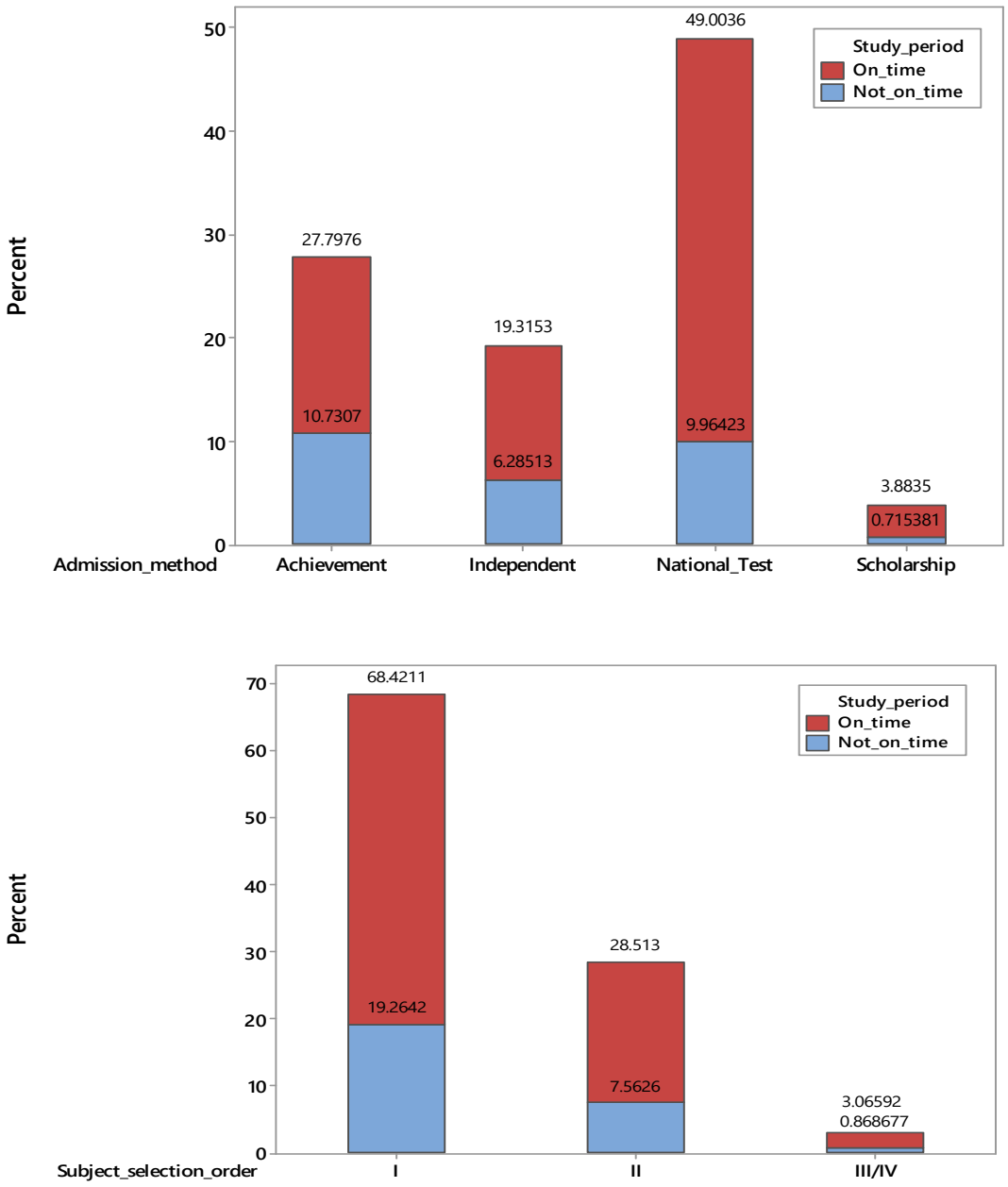


Figure 5. The bar charts of admission method and subject selection order

Figure 5 provides information on four admission methods and subject selection order. Most student participated in national tests to be admitted in FST and they have good capability to finish

their study on-time. Moreover, it is quite a few for students to place subjects in FST as their third or fourth priority. Many of them make their subject as their first priority and they have a commitment to graduate on-time.

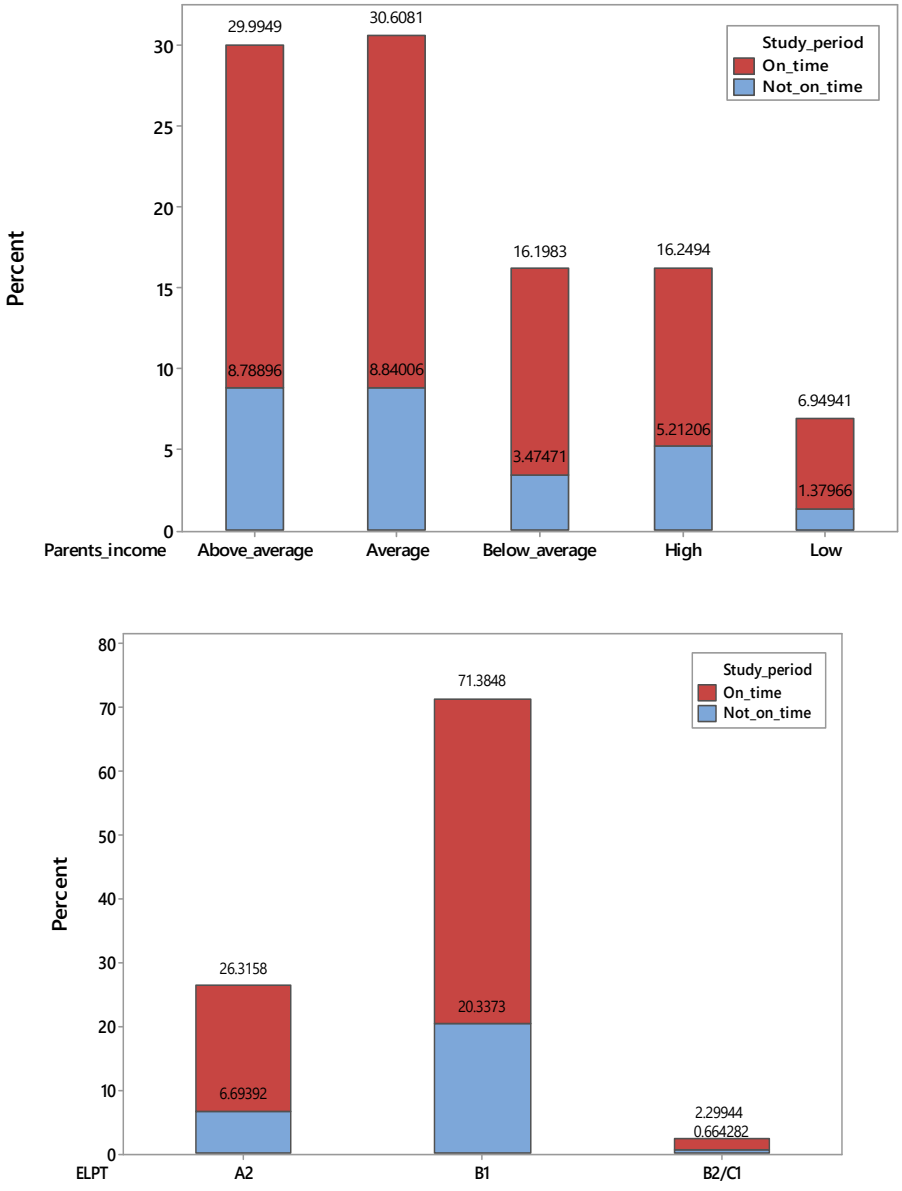


Figure 6. The bar charts of parents' income and ELPT

Beside the academic variable, this study tried to include parents' income as one of the attributes in the classification as given in Figure 6. Students come from low and below-average-income parents tend to perform better in the period of study than those who come from average, above-average, and high income family. Furthermore, based on ELPT score, most of graduates have good capability in English as many of them are in B1 and A2 grade. In all CEFR grade, the on-time graduates dominate the late graduates.

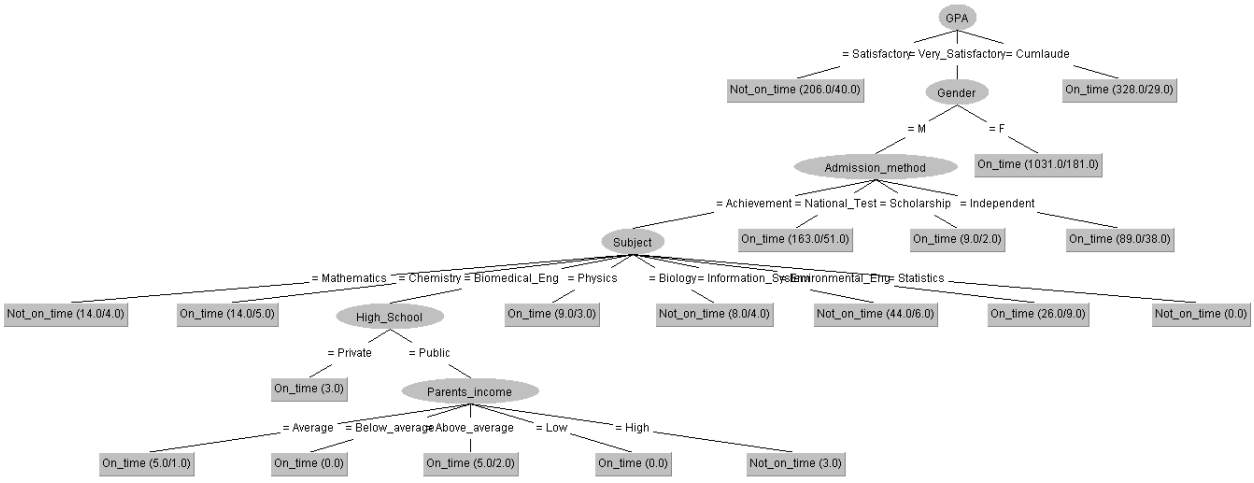


Figure 7. The tree visualisation of the Decision Tree method

GPA that categorises as satisfactory, very satisfactory, and cum laude is one of attribute in this study. Figure 7 depicts all of the cum laude graduates finished their study on-time while very satisfactory graduates tend to be on-time as well. In contrast, there is only small percentage for

satisfactory graduates who finish their study in 4 years or below.

The Test of Association between Dependent and Independent Variables

To examine the association between attributes (independent variables) and response (dependent variable), the Chi-square association test was conducted. The results are given in Table 2.

Table 2. The result of independence test between independent and dependent variables

No	Attribute	Pearson statistic	DF	P-value
1	Subject	200.862	7	0
2	Gender	136.429	1	0
3	**Address**	**0.718**	**2**	**0.698**
4	High School	12.517	1	0
5	National Exam Score	6.471	3	0.091
6	Admission Method	65.969	3	0
7	**Subject Selection Order**	**0.536**	**2**	**0.765**
8	Parents Income	14.573	4	0.006
9	**ELPT**	**1.784**	**2**	**0.41**
10	Graduate Predicate	353.746	2	0

According to Table 2, from the total of 10 attributes, three of them were not significantly associated with the study period at the 10% significance level. These attributes are Address, Subject Selection Order, and ELPT. Thus, based on parsimonious principle, those three attributes will not be included in the classification analysis.

The Classification Using Decision Tree Method

This subsection is the main analysis of this study. The classification based on Decision Tree classifier with C4.5 algorithm is conducted and the results are presented in Table 3.

Table 3. The result of classification of the method

Method	Training/Testing (%)	Accuracy (%)	RMSE	AUC
Decision Tree	30/70	78.4672	0.4068	0.686
	50/50	80.5726	0.3892	0.746
	70/30	78.5349	0.4006	0.743
	80/20	79.0281	0.3973	0.751
	90/10	79.0816	0.4036	0.731
	Cross-validation (10-fold)	79.6627	0.3939	0.741

Table 3 presents the various sample (training/testing) and 10-fold cross validation for each method along with the accuracy, RMSE, and area under ROC curve (AUC). Overall, the highest accuracy for the model is reached when the proportion of training/testing data are 50/50. Beside the accuracy, other goodness of fits are RMSE and AUC. The good model will have highest accuracy, smallest RMSE, and largest AUC.

One of the advantages of Decision Tree is the nice visualisation of the structural tree. It gives more explanation about the dominance or priority of the attributes in classification that will appear in the rule of the tree. Figure 8 shows the tree visualisation of 50/50 training/testing set of the

Decision Tree method. Based on Figure 8, the most important attribute is GPA. Students with cum laude predicate (GPA in the range of 3.5 to 4.0) tend to finish their study on-time whereas the students with a satisfactory predicate tend to be not on-time in finishing their study. On the other hand, the very-satisfactory students depend on other attributes, which is gender. Female students tend to have commitment to graduate on-time rather than male students. For very-satisfactory-male students, they depend on admission method. This rule will continue until the last attribute (parents' income), which is the least important attribute.

The students with scholarship to enter university tend to finish their study faster than those without the scholarship (independently funded by their parents or others). This result agrees with the finding of Glocker (2011). However, the students who come from wealthy parents with high income tend to have the longest duration of study of all.

The future work of this study will be the addition of the attribute that is expected to influence student graduation. Robertson et al. (2010) consider the student having a part-time job while studying at school. This attribute may be a good factor to be included in classification. Moreover, a work by Zavale et al. (2016) stated that graduation rates are not only affected by the academic aspects of students but also the aspect relating to the institution.

Conclusion

This study concludes that most of FST student graduates on-time with 72% proportion. Moreover, the Decision Tree with C4.5 algorithm has good classification performance with high classification accuracy (more than 75%). In addition, the most important variable to predict the study period of FST student is GPA followed by gender.

Acknowledgement

The authors thanks the Institute of Research and Innovation (LPI) of Airlangga University for funding this work.

References

Abu Tair, M. M., & Moh M., El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *International journal of information and communication technology research*, **2** (2), 140-146.

BAN PT - National Accreditation Board of Higher Education. (2008). Book VI Matrix Instrument Rating Program Accreditation.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters,* **27**, 861-874.

Gibert, K., Sànchez-Marrè, M., & Codina, V. (2010). Choosing the right data mining technique: classification of methods and intelligent recommendation. *Proceeding of 5th International Congress on Environmental Modelling and Software*, Canada.

Glocker, D. (2011). The effect of student aid on the duration of study. *Economic of education Review*, **30**(1), 177-190.

Goele, S., & Chanana, N. (2012). Data Mining Trend in Past, Current and Future. *International Journal of Computing & Business Research*.

Hssina, B., Merbouha, A., Ezzikouri, H. & Erritali, M. (2014). A comparative study of Decision Tree ID3 and C4.5. *International journal of advance computer science and applications*, **4**(2).

Indonesian Government Regulation No. 66 (2010). Management and Delivery of Education.

Kesumawati A., & Waikabu, D. (2017). Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine. *International journal of applied business and information systems,* **1**(2), 6-12.

Kesumawati A., & Utari, D. T. (2018). Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine. *AIP conference proceedings,* **2021**, 060005.

Larose, Daniel T., dan Larose, Chantal D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining Second Edition*. New Jersey: John Wiley & Sons Inc.

Pattekari, S., A. dan Parveen, A. (2012). Heart Disease Prediction System Using Naïve Bayes, *National Conference on Recent Advancements in Engineering*, **3**. ISBN: 978-93- 82062-40-0, 2012, (153-156. 40)

Anonymous, Government Rule of the republic of Indonesia No. 4 of 2014 about the establishment of higher education.

Quinlan, R. (1993). C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Salzberg, S. L. (1994). Book Review: C4.5: by J. Ross Quinlan. Inc., 1993. Programs for Machine Learning. *Machine Learning*, 16, 235-240.

Robertson, S. Canary, C. Orr, M. Herberg, P. Rutledge, A. N. (2010). *Journal of Professional Nursing*, **26**(2), 99–107.

Turban, E., J.E. Aronson dan T.P. Liang. (2005). *Decision Support System and Intelligent*

Systems - 7th ed. USA: Pearson Education, Inc.

Zavale, N. C. Santos, L. A. Manueld, L. Dias, M. C. L. Khan, M. Tostão, E. Mondjanaf, A. M. (2017). Decision-making in African universities demands rigorous data: Evidence from graduation rates at Eduardo Mondlane University in Mozambique. *International Journal of Educational Development*, **52**, 122–134.

Zhao, Y. (2013). *R and Data Mining: Examples and Case Studies*. Elsevier Inc