Advances in Social Science, Education and Humanities Research, volume 474
Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)

Classifying Student's Duration of Study in Faculty of Science and Technology UNAIR Using Naïve Bayes and Neural Network Classifiers

Siti Maghfirotul Ulyah[1,*], Marisa Rifada[1], Elly Ana[1]

[1]*Department of Mathematics, Airlangga University, Indonesia*
[*]*Corresponding author: maghfirotul.ulyah@fst.unair.ac.id*

ABSTRACT
Timely graduation is one of the essential criteria for a university in the accreditation program. The objective of this study is to predict the duration of study based on several factors related to students. The data in this study were the data of Faculty of Science and Technology (FST) graduates for 11 years (2008-2018) but limited to the undergraduate degree. The department in FST includes Mathematics, Statistics, Information System, Chemistry, Biology, Physics, Biomedical Engineering, and Environmental Sciences and Technology. The attributes in this work are department, address, gender, high school status, high school national exam score, admission program, department selection order, parents' income, GPA and ELPT. The dependent variable, study duration, is divided into two categories, which are a timely graduate (less or equal to 4 years) and untimely graduate (more than 4 years). The classification methods in predicting the period of study are Naïve Bayes and Neural Network. In this study, various percentages of training data and testing data will be compared. The results reveal that Naïve Bayes outperforms Neural Network in classification accuracy, even in the smaller sample.and their difference is statistically significant.
Keywords: *graduate, duration of study, classification, neural network, Naïve Bayes.*

1. INTRODUCTION

Education has an essential role in the composition of the human development index (HDI). In general, the higher the level of education per capita of a country, the higher HDI of the country. One level of education that plays an important role in advancing education in Indonesia is higher education organized by universities, both public and private universities. Higher education consists of a diploma program, undergraduate program, post-graduate program (master and doctoral programs), and professional programs, as well as specialist programs (Indonesian Government Regulation No. 66, 2010).

From thousands of universities in Indonesia, Airlangga University (UNAIR) is one of the best universities which holds autonomy as a public university called "PTN-BH". In improving the quality, UNAIR has always participated in some assessment activities held by the National Accreditation Board of Higher Education (BAN-PT). UNAIR is currently holding the best score of accreditation (A). However, the quality of UNAIR still has to be maintained and improved. In examining a university, BAN-PT measures the quality of a university based on several standards, including students and graduates [1]. The work in this paper will focus on the graduates since the timely student graduation is very influential in the assessment of the quality of a college. For an undergraduate program, the credit is charged according to government regulations, which is 144 credits, with a maximum study period of 14 semesters or 7 years. Students are graduating on time if their study period is less than or equal to 4 years.

The study by Abu and El-Halees [2] stated that utilizing educational data as learning attributes is one to improve the quality of higher education. Educational data usually consist of student background and student academic achievement. To access this data, UNAIR has an integrated information system in the "cyber campus". These data will be analyzed to create useful information that can be used as material for policymaking to improve the quality of a university.

In this work, several attributes are considered for prediction. We have more attributes than the previous work by Kesumawati and Utari [3]. The attributes include the data of Grade Point Average (GPA), English Language Prediction Test (ELPT), national exam score in high school, high school status, subject selection order, admission program, parents' income, gender, and also regional origin (address). In the efforts to go to World Class University, UNAIR applied a special admission program for students from Eastern Indonesia (Papua). Thus, the student's origin is considered as one of the factors in the classification of the study period.

Considering the number of students and the period of establishment of UNAIR, the educational data that will be collected are a large dataset. Thus, data mining is the proper method to deal with that large dataset. Data mining is an automatic learning process for useful information in large data storage places [4]. The supervised learning approach is a method in data mining where one has data to be trained, and there are targeted variables. A method in supervised learning

is the classification where the object is assigned to one of some predefined categories [5]. In this study, the methods used in classification are probabilistic (Naïve Bayes) and Artificial Neural Network (ANN) Classifiers.

The main objectives in this study are to have an overview of graduates profile at the Faculty of Science and Technology UNAIR and to identify the pattern of student's period of study using Naïve Bayes and Neural Network classifiers. Additionally, the difference in the classification of the two methods will be examined.

Previous studies on the pattern of graduation status of FMIPA Islamic University of Indonesia students was studied by Kesumawati and Waikabu [6] and Kesumawati & Utari [3] by using Naïve Bayes method and Support Vector Machine (SVM). The results concluded that the greatest classification accuracy was obtained from the SVM model which was equal to 69.51%. The attributes used in the prediction are department, gender, GPA, city of residence, high school type, high school program, and parent occupation. Our object of study is similar to Kesumawati and Utari [3] that study about graduates but in a different place (UNAIR) and additional method (ANN). Besides, some additional attributes are introduced in our studies, such as ELPT, high school national exam score, department selection order, admission program, and parents' income.

A study about the student aid was conducted by Glocker [7] where concluded that student aid recipients tend to have a shorter time to graduate than comparable students without it, such as students supported by the equal amount of parental or private transfers only. In addition, on average, higher financial aid did not influence the duration of the study. Thus, our study has intention to consider students who receive a scholarship upon the admission and also the income of the parents.

Moreover, Gibert et al. [8] in his work entitled "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation" revealed that the appropriate supervised learning method to explain or predict qualitative variables is the discriminant method. Naïve Bayes is the member of the discriminant method. Naïve Bayes is a combination of prior models and iterative refinements based on future data (Bayesian learning).

This paper is structured as follows. Section 1 presents the background and the objective of this study. Then, the next section shows the materials method of the study. Section 3 presents the classification result and the conclusion is given in the last section.

2. MATERIALS AND METHODS

2.1. Data

The data in this study were the profile of FST graduates from 2008 to 2018 with 1957 observations. One should note that UNAIR has three times in a year of graduation. Those observations are the sum of all graduation time. These data are the data of undergraduate students in each subject in FST (Mathematics, Statistics, Physics, Biomedical Engineering, Chemistry, Environmental Sciences and Technology, and Biology). The data will be divided into training data and testing data. Training data are data that will be used in classification modelling, while testing data are used to validate classification results. These secondary data were obtained from the database of academic sub-division of FST UNAIR.

2.2. Research Variables

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

Table 1. The research variables

Variable	Attribute	Category
Dependent variable	Study Duration	On-time
		Not on time
Independent variable	Subject (department)	Mathematics
		Statistics
		Information System
		Physics
		Biomedical Engineering
		Chemistry
		Biology
		Environmental Sciences and Technology
	Gender	Male
		Female
	Address	East Jawa
		Outside East Java
		Outside Java
	High School	Public
		Private
	National Exam Score	Excellent
		Very Good
		Good
		Fair

Variable	Attribute	Category
	Admission Method	Achievement
		National test
		Scholarship
		Independent
	Subject Selection Order	I
		II
		III/IV
	Parents' Income	Low
		Below average
		Average
		Above average
		High
	ELPT	A2
		B1
		B2/C1
	GPA	Satisfactory
		Very satisfactory
		Cumlaude

2.3. The Method

The preliminary method in this work is obtaining the descriptive statistics of the variables and investigating the relationships among variables. Since the scale of all variables is nominal and ordinal, the Chi-square independence test will be conducted. After that, the main analysis (classification) is presented using Naïve Bayes and Neural Network methods then make comparison among the two approaches. After obtaining the classification results, the interpretation and reasoning will be given.

Naïve Bayes Classifier

Naïve Bayes algorithm can be used to predict the probability of membership of a class in classification problem. It is based on the Bayes theorem which has the same classification ability as neural network and the performance is quite good with large data [3]. Naïve Bayes is based on the simplifying assumption that attribute values are conditionally independent if given output value. This means, given the output value, the probability of observing together is a product of probability. Pattekari and Parveen [9] reveal that Naïve Bayes has advantages in terms of the amount of training data. This method only requires a small amount of training data to be able to do parameter estimates in classification. In many cases in the real world, Naïve Bayes tends to produce optimal results.

Based on the Bayes Theorem, the probability of $E = (x_1, x_2, \ldots, x_n)$ to be class c is

$$p(c \mid E) = \frac{p(E \mid c)\, p(c)}{p(E)}. \tag{1}$$

Suppose there are two classes namely positive (+) and negative (-). E will be classified as class C = + if and only if

$$f_b(E) = \frac{p(C = + \mid E)}{p(C = + \mid E)} \geq 1, \tag{2}$$

where $f_b(E)$ is the Bayesian classifier. Assume that given an output value, all attributes are independent. In other words:

$$p(E \mid c) = p(x_1, x_2, \ldots, x_n \mid c) = \prod_{i=1}^{n} p(x_i \mid c). \tag{3}$$

Then, the resulting classifier is

$$f_{nb}(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^{n} \frac{p(x_i \mid C = +)}{p(x_i \mid C = -)}. \tag{4}$$

The function $f_{nb}(E)$ is referred to as a naïve Bayesian classifier or more simply naïve Bayes. Naïve Bayes is the simplest form of Bayesian network where all independent attributes are given the value of the class variable (output). This is called conditional independence [10].

Artificial Neural Network Classifier

ANN is a system resembling neurophysiology that consists of a collection of simple nonlinear computing elements whose inputs and outputs are tied together to form a network. The general structure of ANN consists of input layer, hidden layer, and output layer in which all of them are the "black box" that cannot be interpreted. In each layer, there are many nodes as showed in Figure 1. ANN Model is an assembly of inter-connected nodes and weighted links. The output node sums up each of its input value according to the weights of its links. Then, the output node is compared to some threshold *t.*

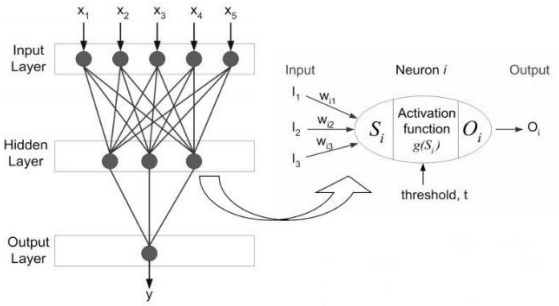


Figure 1. The general structure of ANN

There is two kinds of ANN model (Supervised and Unsupervised Learning). For classification, the supervised learning will be conducted. In a supervised learning algorithm, one obtains the training samples from the domain area and tries to learn the relationship between input variables and output variables [5].

Regarding the number of hidden-layer nodes, basically, more hidden-layer nodes makes it reaches the convergence slower. However, it may result in smaller error value, especially the training sample's error. A rule for determining the number of hidden-layer nodes is $(N_O + N_I)/2$ where N_I is the number of input-layer nodes and N_O is the number of output-layer nodes. This rule will be applied in this study.

3. RESULTS AND DISCUSSION

3.1. The Data Overview

The descriptive statistics of the variables were given in Figures 2 to 8. Figure 2 depicts the percentage of the student who graduates not more than 4 years (timely) and student whose study period of more than 4 years (untimely). It reveals that most undergraduate student of Faculty of Science and Technology (FST) finished their study on-time (72.3%).

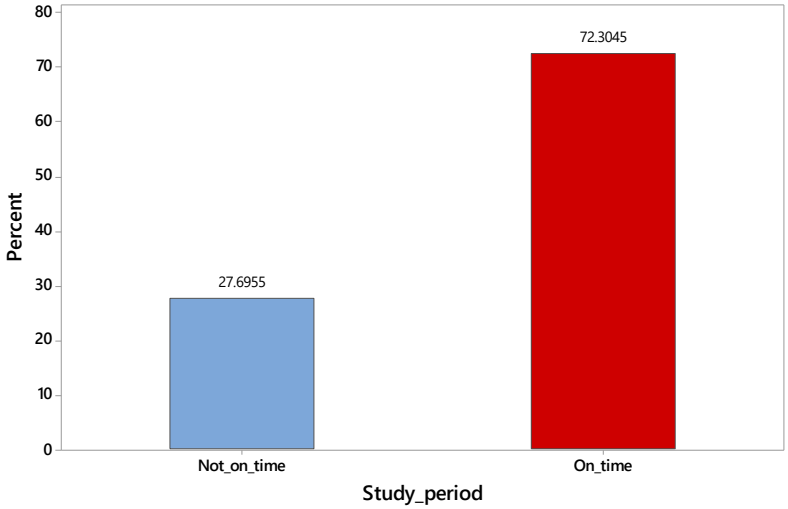


Figure 2. The bar chart of the study period (duration)

Figure 3 depicts the subject based on the study duration of FST student. The subject with the highest proportion of on-time graduate in Statistics, followed by Environmental Sciences and Technology and Chemistry. On the other hand, Information system has the highest percentage of untimely graduate, followed by Physics and Mathematics.

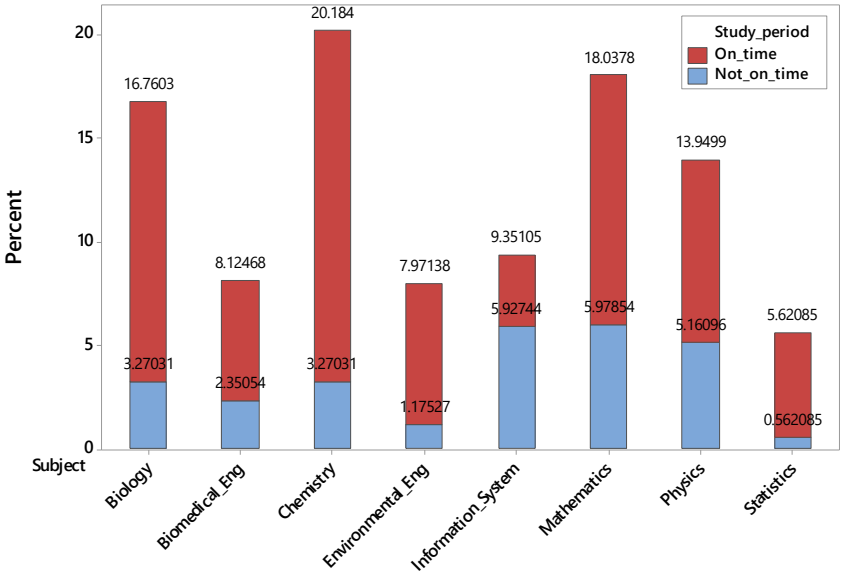


Figure 3. The bar chart of the subject of study

Figure 4 provides information on the gender and regional origin of FST graduates based on the duration of the study. About half of male students graduate on-time while female students tend to have timely graduation rather than untimely. Besides, students coming from East Java have a higher percentage of not-more-than-8-semesters graduation compared to others.

Based on Figure 5, FST students are mostly from public high school. Furthermore, the percentage of on-time graduates from public high school is larger than that of private high school. However, their difference is small. In addition, students having excellent and very good national exam score tend to come to FST and their percentage of graduating on-time is great.

Figure 6 gives information about four admission programs and subject selection order when applying to the desired university. Most students had participated in a national test to be admitted in FST and they have good performance to finish their study on-time. Furthermore, only a few students who place subjects in FST as their third or fourth priority. They mostly make their current subject as their first priority and have a commitment to graduate on-time.

Besides the academic variables, this study also include parents' income as one of the attributes in classification. Students from low and below-average-income parents tend to perform better in terms of the duration of study than those who come from higher income (average, above-average, and high income) family. Moreover, according to ELPT score, most of the graduates have good ability in English (many of them are in B1 and A2 grade). In all CEFR grade, the timely graduates dominate the untimely graduates.

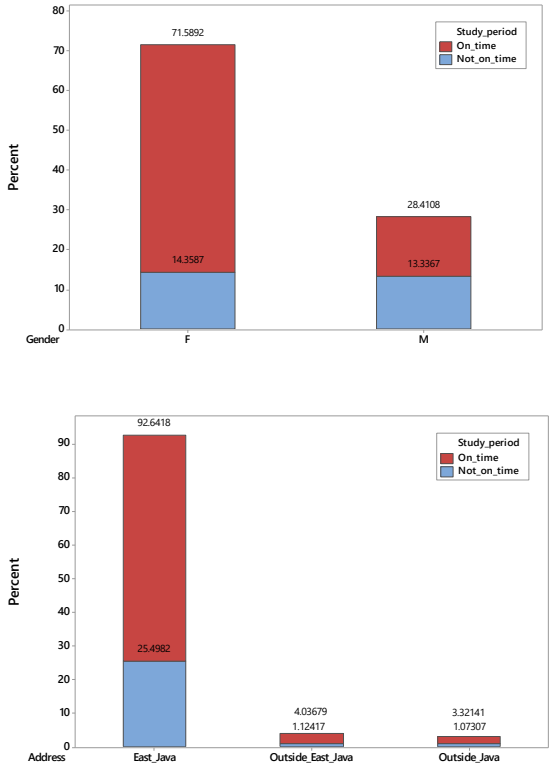


Figure 4. The bar chart of gender and address

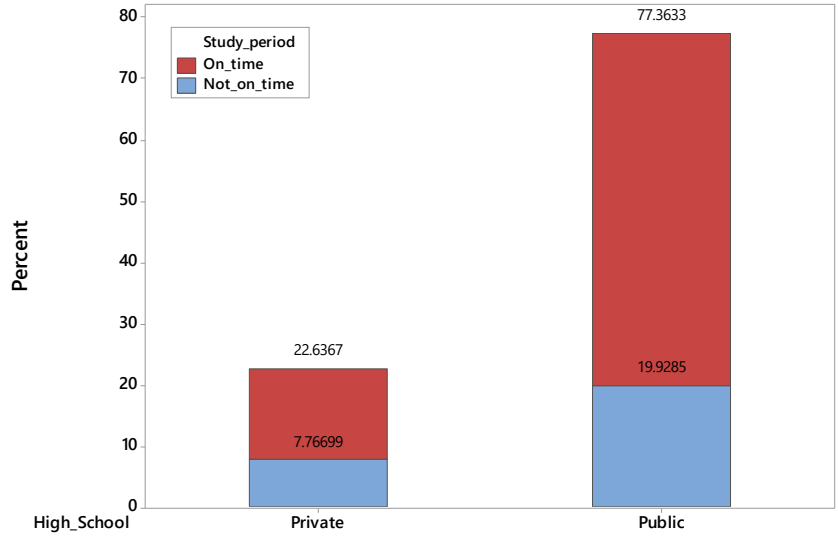
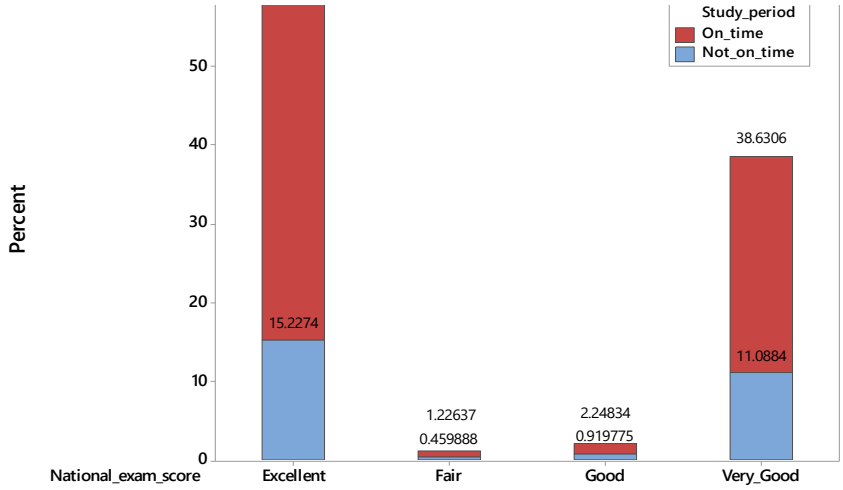


Figure 5. The bar chart of high school and national exam score



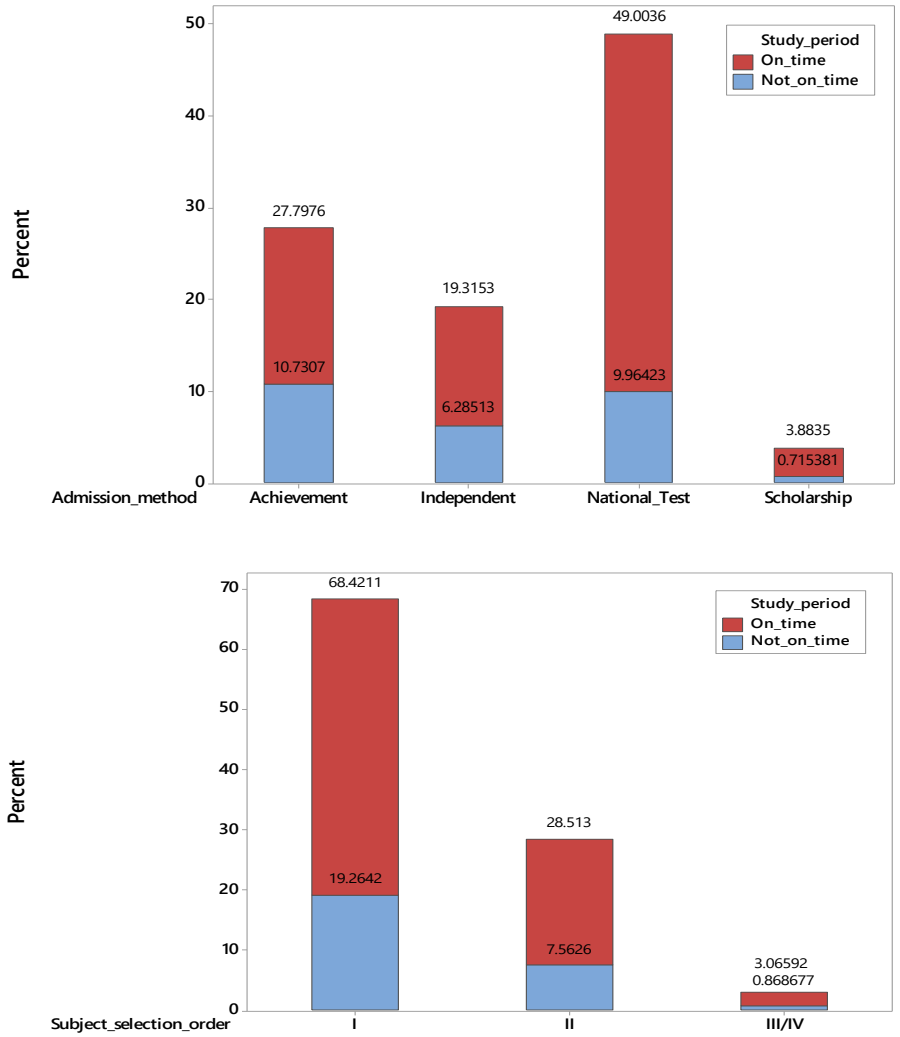
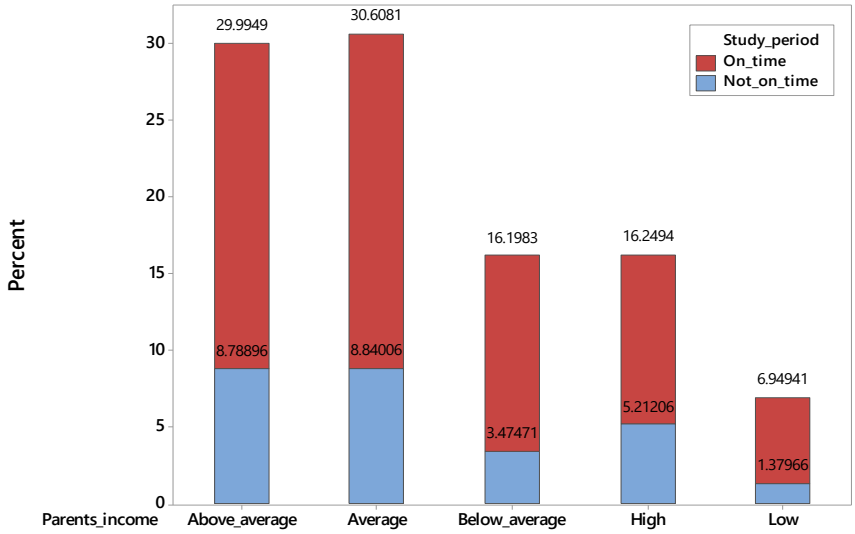


Figure 6. The bar chart of admission method and subject selection order



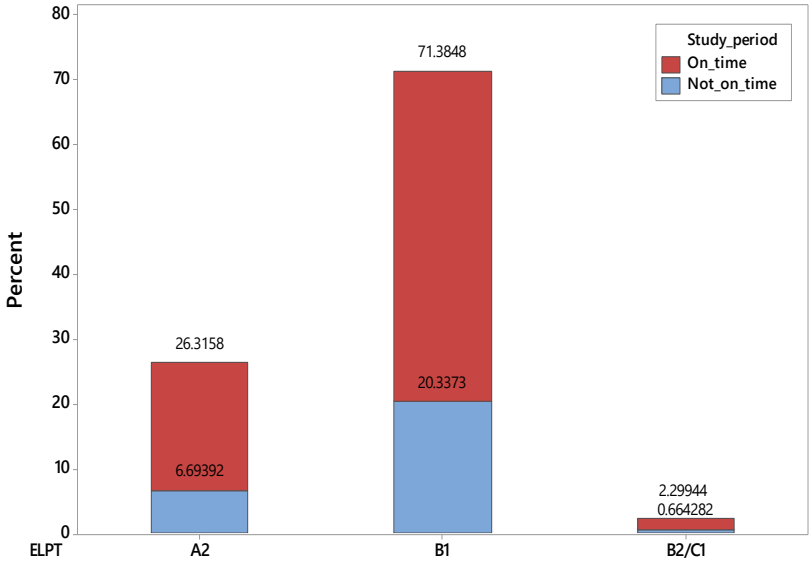


Figure 7. The bar chart of parents' income and ELPT

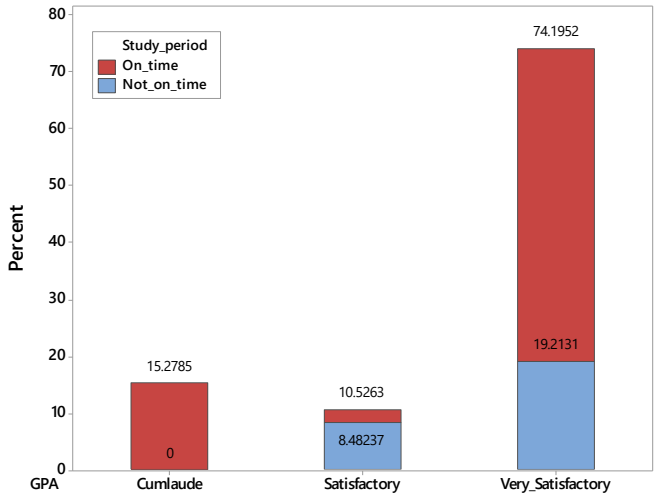


Figure 8. The bar chart of GPA

GPA consisting of satisfactory, very satisfactory, and cum laude is one of the attributes in this work. Figure 8 presents all of cum laude graduates finished their study on-time whereas very satisfactory graduates mostly be on-time as well. On the other hand, there is quite a small percentage for satisfactory graduates finishing their study in 4 years or less.

3.2. The Chi-Square Independence Test

To investigate 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 an independence test between attributes and response

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	GPA	353.746	2	0

According to Table 2, there are three out of ten attributes were not significantly associated with the study duration at the 10% significance level. These attributes are Address (regional origin), Subject Selection Order, and ELPT (written in bold). Thus, to get parsimonious model, these three attributes were not be included in the classification analysis.

3.3. The Classification Using Neural Network and Naïve Bayes Methods

The classification results based on Neural Network and Naïve Bayes methods are given in Table 3. The various percentage of training and testing data was given to show the performance of each method in various sample size.

Table 3. The result of classification

Method	Training/Testing (%)	Accuracy (%)	RMSE	ROC Area
Neural Network (Backpropagation)	30/70	76.3504	0.456	0.743
	50/50	75.1534	0.4507	0.772
	70/30	**77.3946**	**0.438**	**0.762**
	80/20	76.3203	0.447	0.747
	90/10	76.7263	0.4248	0.782
	Cross-validation (10-fold)	77.0408	0.4308	0.769
Naïve Bayes	30/70	79.4161	0.3861	0.796
	50/50	**81.5951**	**0.3718**	**0.817**
	70/30	79.8978	0.379	0.806
	80/20	80.5627	0.3791	0.812
	90/10	80.1020	0.3823	0.802
	Cross-validation (10-fold)	80.4292	0.3781	0.816

Table 3 shows the accuracy, RMSE, and area under ROC curve (AUC) of each methods based on various sample proportion (training/testing) and 10-fold cross validation. In general, the highest accuracy for the two models is reached when the proportion of training/testing data are 50/50. Beside the accuracy, other goodness of fits are RMSE and AUC. The best model will have highest accuracy and AUC, and smallest RMSE. According to these criteria, one can see that Naïve Bayes has better performance in classifying the study period of FST students than ANN, even in the smaller sample.

A reason for this is the common issue for ANN, which is over fitting. Also, making decisions only based on point prediction may result incorrect classifications with spuriously high confidence [11][12]. In this study, the number of hidden-layer nodes are defined as the sum of the number of input-layer nodes and output-layer nodes divided by two. The number of hidden layers of ANN may influence the classification results and have a risk to overfitting. Overfitting may result in ANN model that are more complex than necessary.

3.4. The Comparison between Neural Network and Naïve Bayes Classifiers

The performance of the two classifiers has been compared in Section 4.3 and Naïve Bayes classifier had better performance than ANN. However, one wants to know whether the difference in classification result between those two is statistically significant. Here the comparison method is adopted from Tan [5]. The results are presented in Table 4.

Table 4. The comparison between the two classifiers

Cross-validation (*k*-fold)	Error (%)		$\lvert d_i \rvert$	$(d_i - \bar{d})^2$
	ANN	Naïve Bayes		
2	23.1477	19.673	0.0347	0.0007458
3	23.4032	19.162	0.0424	0.0012232
4	23.812	19.4686	0.0434	0.0012957
5	22.0746	19.5197	0.0255	0.0003280
6	21.9213	19.2642	0.0266	0.0003661
7	23.4032	19.3153	0.0409	0.0011183
8	22.4323	19.673	0.0276	0.0004062
9	23.1477	19.673	0.0347	0.0007458
10	25.1405	19.5708	0.0557	0.0023290
			$\bar{d}$ = 0.0368	$\sum$ = 0.0085581

Based on Table 4, we can compute the statistic $\hat{\sigma}_d^2$, which is the variance of the difference, as follows:

$$\hat{\sigma}_d^2 = \frac{\sum_{i=2}^{k} (d_i - \bar{d})^2}{k(k-1)} = 0.00009509 . \quad (4)$$

Then, the 95% confidence interval is

$$(1-\alpha)\%CI = \bar{d} \pm t_{(1-\alpha,k-1)} \hat{\sigma}_d$$

$$95\%CI = \bar{d} \pm t_{(95\%,10-1)} \hat{\sigma}_d = 0.0074 \pm (2.26)(0.00975) = (0.0148, 0.0589)$$

. (5)

According to the calculation above, the variance of the difference is not included in the 95% confidence interval. Therefore, it can be concluded that the difference between Naïve Bayes and ANN classifiers is statistically significant at 5% level.

4. CONCLUSION

This study concludes that Naïve Bayes has superb classification performance. Comparing to ANN, the classification accuracy of Naïve Bayes is higher than that of ANN. Naïve Bayes method outperforms ANN even in a smaller sample. Moreover, the result between the two classifiers has a statistically significant difference at 5% level of significance. Future work may determine the optimal number of hidden layer and hidden node of ANN by implementing a genetic algorithm proposed by Stathakis [13] to enhance the prediction and minimize the classification error.

ACKNOWLEDGMENT

The authors thank the institute of research and innovation (LPI) Airlangga University for the funding of this work.

REFERENCES

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

[2] Abu Tair, M. 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.

[3] 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.

[4] Goela, S., Chanana, N. (2012). Data Mining Trend in Past, Current and Future. International Journal of Computing & Business Research.

[5] Tan, P. N. (2018). Introduction to Data Mining. Pearson Education India.

[6] 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.

[7] Glocker, D. (2011). The Effect of Student Aid on the Duration of Study. Economic of Education Review. 30(1): 177-190.

[8] Gibert, K., Sànchez-Marrè, M., Codina, V. (2010). Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation.

[9] Pattekari, S., A. 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)

[10] Zhang, H. (2004). The optimality of naive Bayes. 1(2): 3.

[11] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q. (2017). On Calibration of Modern Neural networks. In: International Conference on Machine Learning. 1321–1330.

[12] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., Hinton, G. (2017). Regularizing nNeural Networks by Penalizing Confident Output Distributions, arXiv preprint arXiv:1701.06548.

[13] Stathakis, D. (2009). How Many Hidden Layers and Nodes?. International Journal of Remote Sensing. 30(8). 2133-2147.