BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1419–1428
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1419-1428

STUDY TIME CLASSIFICATION OF MATHEMATICS AND INFORMATION TECHNOLOGY DEPARTMENT OF KALIMANTAN INSTITUTE OF TECHNOLOGY USING NAÏVE BAYES ALGORITHM

Fatrysia Wikarya Sucipto[1], Ramadhan Paninggalih[2*], Indira Anggriani[3]

[1,3]*Mathematics Study Program, Department of Mathematics and Technology Information, Kalimantan Institute of Technology*
[2]*Informatics Study Program, Department of Mathematics and Technology Information, Kalimantan Institute of Technology*
Soekarno Hatta Street Km.15, Balikpapan, 76127, Indonesia

*Corresponding author's e-mail: *ramadhanpaninggalih@lecturer.itk.ac.id*

Article History:

Received: 10th February 2023
Revised: 15th July 2023
Accepted: 25th July 2023

Keywords:

Accuracy;
Classification;
College quality;
F1-score;
Naïve Bayes

ABSTRACT

Institut Teknologi Kalimantan (ITK) is one of the state universities in Indonesia which has 5 majors, one of them is the Department of Mathematics and Information Technology (JMTI). JMTI has six study programs, and only three study programs have graduates, namely Mathematics, Information Systems, and Informatics. Every year the number of new students continues to grow, but this is not proportional to the number of graduates, because some students study for more than 8 semesters. Because of this, the quality of study programs being poor. In this research, a model was built that could classify student study timeliness, using the naïve Bayes algorithm. The data used is data from JMTI student graduates from the 2013 to 2019 batch. The 2013 to 2018 batch data will be training data and validation data, while the 2019 batch data will be testing data. This research compare accuracy and F1-score naïve Bayes algorithm without correlation and with correlation. The best model obtained from training data is a model with variables that have gone through a correlation test, namely 70:30, 80:20, and 90:10. The attributes selected after the correlation test, namely, IP Tahap Bersama, GPA, Final GPA, Length of Study (Semester), dan Graduation GPA (Category), yield results for accuracy and an F1-score of 1.

How to cite this article:

F. W. Sucipto, R. Paninggalih and I. Anggriani., "STUDY TIME CLASSIFICATION OF MATHEMATICS AND INFORMATION TECHNOLOGY DEPARTMENT OF KALIMANTAN INSTITUTE OF TECHNOLOGY USING NAÏVE BAYES ALGORITHM," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1419-1428, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

One of the most important things in life is education because education can be a human foundation for improving the quality of life [1]. Therefore, education is one of the rights that must be obtained by all Indonesian people without exception. One of the places to get an education in college. The quality of tertiary institutions, in this case, the study programs in Indonesia, is measured based on the accreditation held by the Independent Accreditation Agency (LAM). Graduates and Students are one of the nine-main standards seen in terms of accreditation. Explicitly regarding graduate and student assessment standards, the part surveyed is the framework for enrolling new students and graduates (average GPA and length of study). Institut Teknologi Kalimantan (ITK) is a state university that focuses on the field of innovative technology to address current industrial world problems [2]. ITK has five departments, one of which is the Department of Mathematics and Information Technology (JMTI). JMTI itself has six study programs including Mathematics, Information Systems, Informatics, Actuarial Science, Statistics, and Digital Business. Of the six study programs at JMTI ITK, only three study programs have graduated because the Actuarial Science and Statistics study programs were only inaugurated in 2020, as well as Digital Business in 2021 [3]. The increase in the number of students in the Mathematics, Informatics, and Information Systems study programs is not always proportional to the number of graduates each year. Therefore, we need a system that can classify student study time. This system can later be used in classifying when students study whether they can graduate on time or not.

The naïve Bayes algorithm is used for prediction problem solving graduation student education Mathematics of UIN Raden Intan Lampung. The results of the discussion are obtained by using RapidMiner 5.3 with data training as much as 51 data obtained accuracy of 74.67%. testing data as many as 184 students as test data the result was that students who will pass on time by 42 students or around 22.8% of the total inappropriate data testing and students time as many as 142 students or approximate 77.2% [4]. The naïve Bayes algorithm is used for solving data classification problems aid recipients. Result of discussion obtained naïve Bayes and generated rules have levels accuracy of 90%, using 90% training data and 10% testing data [5]. The naïve Bayes algorithm is used to solve major class XI (IPA and IPS) class classification problems. The results of the discussion obtained the level of accuracy using naïve Bayes of 81.82%, and by using K-Nearest Neighbor of 92.73% [6]. Based on these researches then classification of student study time can be classified using the naïve Bayes algorithm.

The dataset in this research split into three parts, namely training data, data validation, and data testing. The training data and validation data used in this research are graduation data from students of the Mathematics, Informatics, and Information Systems study programs at the Institut Teknologi Kalimantan batch 2013-2018, and data testing uses data from the batch 2019. One of the considerations for students to graduate on time or not is their academic history. Factors that exist in the academic history, there are Graduation Grade Point Average (IPK), Joint Preparation Stage Grade Point Average (IP TPB), and length of study. The attributes that will be used are academic history factors and adding study programs taken (Mathematics, Information Systems, and Informatics), hometown, and gender. Classification will be carried out with these attributes to get good classification results. Based on these problems, the naïve Bayes algorithm will be used to classify the timeliness of students majoring in Mathematics and Information Technology graduating.

2. RESEARCH METHODS

2.1 Dataset Description

This research uses graduation data from JMTI ITK student batch 2013-2019. The input and output variables are described in **Table 1** and **Table 2**, and the graphs of the number of students according to the attributes are described in **Figure 1(a)** to **Figure 1(g)**.

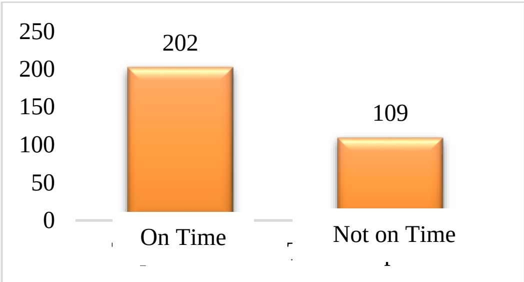
Table 1. Input Variable

Attribute	Type	Description
Study Program	Category	Study Programs
City of Birth	Category	City of Birth
Gender	Category	Gender
GPA	Numeric	Grade Point Average
Joint Preparation Stage GPA	Numeric	Joint Preparation Stage Grade Point Average

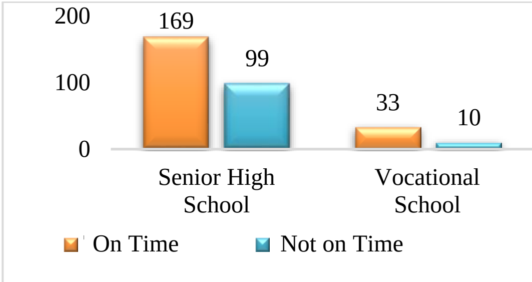
Attribute	Type	Description
Length of Study	Numeric	Length of Study (Semester)
Graduation Predicates	Category	Graduation Predicates
School Types	Category	School Types

Table 2. Output Variable

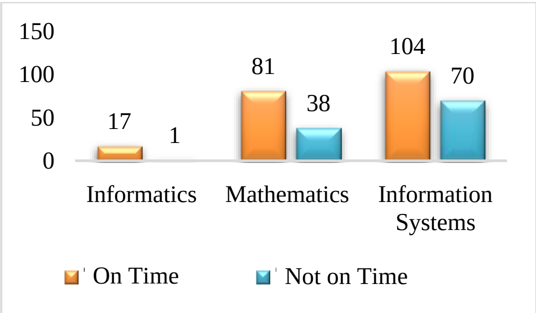
Attribute	Class
Study Timeliness	On time; Not on time



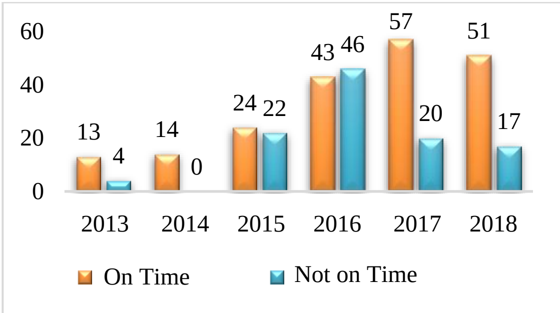
(a)



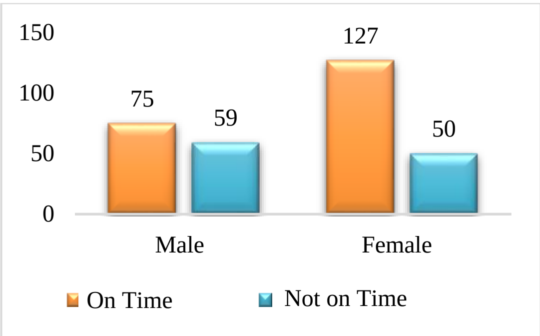
(b)



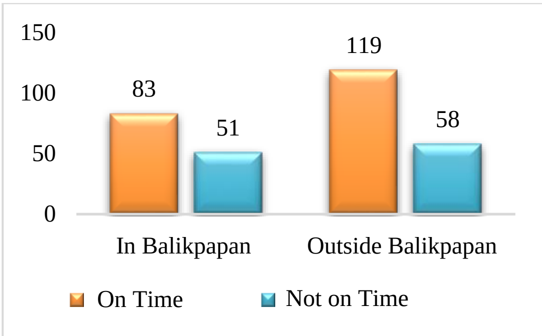
(c)



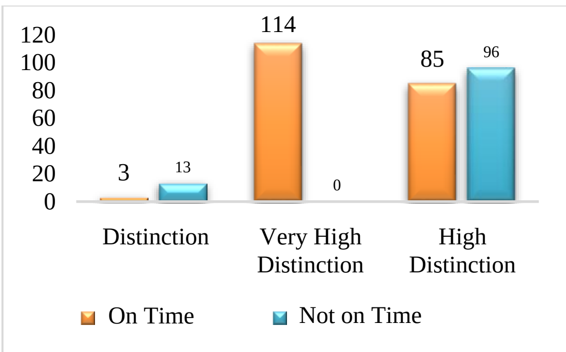
(d)



(e)



(f)



(g)

Figure 1. Bar Chart about attribute input variables with study timeliness attribute, (a) study timeliness bar chart, (b) school types bar chart, (c) study programs bar chart, (d) batch bar chart, (e) gender bar chart, (f) city of birth bar chart, (g) graduation predicates bar chart.

In **Figure 1(a)** there are 202 students graduate on time and 109 students graduate not on time in JMTI. **Figures 1(b)** based on high school types, in the senior high school, there are 169 students graduate on time and 99 graduate not on time. In the vocational school, there are 33 students graduate on time and 10 students graduate not on time. **Figures 1(c)** based on study programs, in informatics there are 17 students graduate on time and 1 student graduate not on time. In mathematics, there are 81 students graduate on time and 38 students graduate not on time. In information systems, there are 104 students graduate on time and 70 students graduate not on time. **Figures 1(d)** based on batch year, in 2013, 2014, 2014, 2015, 2016, 2017, and 2018 respectively number of students graduate on time and not on time are 13, 4, 14, 0, 24, 22, 43, 46, 57, 20, 51 and 17. In 2016, the students graduate not on time is more than the students graduate on time. **Figures 1(e)** based on gender, in male there are 75 students graduate on time and 59 students graduate not on time. In female, there are 127 students graduate on time and 50 students graduate not on time. **Figures 1(f)** based on city of birth, in Balikpapan there are 83 students graduate on time and 51 students graduate not on time. Outside Balikpapan, there are 119 students graduate on time and 58 students graduate not on time. s **Figures 1(g)** based on graduation predicates, in distinction there are 3 students graduate on time and 13 students graduate not on time. In very high distinction, there are 114 students graduate on time. In high distinction, there are 85 students graduate on time and 96 students graduate not on time. Then on the graduation predicate attribute there are 3 categories, two of the categories namely Satisfying, and Very Satisfying have more students graduating not on time.

2.2 Pearson Correlation

Correlation is a number that shows the relationship between two (or more) quantitative variables. Pearson Correlation measures the extent to which two variables are linearly correlated. Pearson Correlation allows for easy comparisons between multiple pairs of variables. This analysis is based on the assumption of a straight-line relationship between quantitative variables. The result of the correlation is the correlation coefficient (r) whose value is between $-1 \leq r \leq 1$ [7]. The formula for determining the correlation coefficient (r) between the dependent variable Y and the independent variable X is given in **Equation (1)** with n amount of data [8]. **Equation (1)** is referred to as the Pearson correlation coefficient equation.

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} \tag{1}$$

The correlation coefficient can be divided into several categories as shown in **Table 3**.

Table 3. Correlation Categories

Correlation coefficient	Correlation Categories
$\lvert 0-0{,}20 \rvert$	Very Weak
$\lvert 0{,}21-0{,}40 \rvert$	Weak
$\lvert 0{,}41-0{,}60 \rvert$	Enough
$\lvert 0{,}61-0{,}80 \rvert$	Strong
$\lvert 0{,}81-1{,}00 \rvert$	Very Strong

2.3 The Naïve Bayes Algorithm

Naïve Bayes is a classification method that utilizes probability and statistics echoed by British scientist Thomas Bayes. Naïve Bayes predicts future opportunities based on past events. The assumption used in the naïve Bayes algorithm is the conditional simplification of attribute values when added to the output value. The output value, the probability of seeing together is the result of individual probabilities.

The naïve Bayes algorithm uses the Bayes theorem principle, which calculates the probability of an event based on certain conditions, using **Equation (2)** [9]:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}, \tag{2}$$

where X is the unknown class data, C is the specific class, $P(C|X)$ is the probability of hypothesis C based on the condition of X, $P(C)$ is the probability of the hypothesis of C, $P(X|C)$ is the probability of X based on the condition on hypothesis C, and $P(X)$ is probability X. Classification with continuous or numerical data can be used with the Gaussian Density formula in **Equation (3)**:

$$P(X_i = x_i | C = c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{\frac{-(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \tag{3}$$

where σ is the standard deviation, μ is the average or mean, $i = 1,2, \dots, n$, $j = 1,2, \dots, m$ with n is the number of data and m is the number of classes.

Maximizing $P(X_i|C_j) \cdot P(C_j)$ to obtain the class C want to classify is done by multiplying $P(X_i|C_j)$ and $P(C_j)$ for all possible classifications in **Equation (4)**:

$$P(C_j | X_1, X_2, \dots, X_n) \propto P(C_j) \prod_{i=1}^{n} P(X_i|C_j) \tag{4}$$

$\boldsymbol{C_j}$ is the number of classes. The class assigned to class $\boldsymbol{C_j}$ is the maximum probability of $\boldsymbol{P(X_i|C_j) \cdot P(C_j)}$.

2.4 Confusion Matrix

Evaluating data objects that will later be included for certain classes from various existing classes is called classification **[10]**. The system of this classification is expected to be able to classify data sets accurately. Even so, inaccuracies in the classification process cannot be avoided. Therefore, performance measurement is needed for this classification system. In general, the confusion matrix is used to measure classification performance. The confusion matrix is a table that records the results of classification performance. The confusion matrix for the two classes (binary) can be seen in **Table 4**.

Table 1. Confusion matrix for Classification of Two Classes

		Prediction Class	
Actual Class		**Negative**	**Positive**
	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

True Positive (TP) is the number of classifications of predictions that state that students graduate on time and the actual event is that students graduate on time. True Negative (TN) is the number of classifications of predictions that state that students graduate not on time and the actual event is that students graduate not on time. False Positive (FP) is the number of classifications of predictions that state that students graduate on time but the actual event is that students graduate not on time. False Negative (FN) is the number of classifications of predictions that state that students graduate not on time while in the actual event, students graduate on time.

The accuracy value of the model is the amount of data that is correctly classified divided by the total amount of data, as in **Equation (5)**:

$$\text{Accuracy} = \frac{\text{The predicted amount of data is correct}}{\text{Number of predictions made}} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \tag{5}$$

Good accuracy in machine learning is subjective. But what is often thought, is that if more than 70% of the model has good performance. An accuracy of between 70%-90% is not only ideal but realistic **[11]**, **[12]**. As for **Table 5**, there are basic rules when understanding accuracy scores.

Table 2. Accuracy Categories

Accuracy Value	Accuracy Categories
$> 90\%$	Very Good
$70\% - 90\%$	Good
$60\% - 70\%$	Enough
$< 60\%$	Bad

Precision is the ratio of the True Positive (TP) to the overall positive predicted result (TP+FP). The equation for calculating precision can be seen in **Equation (6)**:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

Recall (Sensitivity) is a comparison of True Positive (TP) with all true positive data (TP+FN), as in **Equation (7)**:

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

Specificity is the ratio of True Negative (TN) to all negative data (TN+FP), as in **Equation (8)**:

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

Accuracy may not be a good measure if the data set is imbalanced (both negative and positive classes have large amounts of data). In **Table 6**, three levels of degrees of imbalanced data are given **[13]**, **[14]**, **[15]**.

Table 3. Degrees of Imbalanced Data

Degrees of Imbalanced Data	Minority class proportion
Low	$20 - 40\%$ of dataset
Average	$1 - 20\%$ of dataset
Extreme	$< 1\%$ of dataset

To measure the accuracy of classification with imbalanced data, precision and recall values are needed. F1-score is the harmonic average between precision and recall. The F1-score value can be calculated by **Equation (9)**:

$$F1 - score = 2 \times \frac{\text{Precision} \times Recall}{\text{Precision} + Recall} \tag{9}$$

3. RESULTS AND DISCUSSION

3.1 Preprocessing Data

The data preprocessing stage is carried out to prepare data before entering the modeling stage. The first is cleaning the data, the next step is the data selection process, namely selecting the input variables and target variables to be analyzed. Before calculating the correlation of each attribute with the decision attribute, the category-type attributes are labeled first. Attributes in the labeled data include study program, city of birth, gender, graduation title, and type of school. The results of data labeling can be seen in **Table 7** to **Table 11** as follows.

Table 7. Labelling of Study Program Attributes

Study Program	Code of Study Program
Informatics	0
Mathematics	1
System Informatics	2

Table 8. Labelling of City of Birth Attributes

City of Birth	Code of City of Birth
In Balikpapan	0
Outside Balikpapan	1

Table 9. Labelling of Gender Attributes

Gender	Code of Gender
L	0
P	1

Table 10. Labelling of Graduation GPA (Category)

Graduation GPA (Category)	Code of Graduation GPA (Category)

Distinction	0
Very High Distinction	1
High Distinction	2

Table 11. Labelling of School Types

School Types	Code of School Types
SMA	0
SMK	1

Table 7 show the labelling of study programs attributes. Informatics, mathematics, and system informatics respectively label by number 0, 1, and 2. **Table 8** show the labelling of city of birth attributes. In Balikpapan and outside Balikpapan respectively label by number 0, 1. **Table 9** show the labelling of gender attributes. Male (L) and female (P) respectively label by number 0, 1. **Table 10** show the labelling of graduation GPA (Category) attributes. Distinction, very high distinction, and high distinction respectively label by number 0, 1, and 2. **Table 11** show the labelling of school types attributes. Senior high school and vocational school respectively label by number 0, 1.

This research, will compare the model of naïve Bayes algorithm using all attributes and the model of naïve Bayes algorithm using selection attributes use Pearson correlation. This research selected 4 best attributes form the absolute value of Pearson correlation. The absolute values of Pearson correlation are often employed to measure the strength of the relationship between two variables regardless of its direction (positive or negative). The results of the correlation of each attribute to the decision attribute using **Equation (1)** can be seen in **Table 12**.

Table 12. Pearson Correlation Between All Attributes and Study Timeliness Attributes

Attributes	Correlation with Study Timeliness Attributes
Student Number	0.041443
Study Program	0.159626
Year Class	−0.053548
Gender	−0.163792
Joint Preparation Stage GPA	−0.327561
Final GPA	−0.473960
Length of Study (Semester)	0.800957
City of Birth (Category)	−0.054918
School Type (Category)	−0.099001
Graduation GPA (Category)	0.285926

In **Table 12**, it can be seen that the 4 best attributes by correlation attributes with an absolute value, namely IP TPB variable of -0.327561, Final GPA (IPK) of -0.473960, Length of Study (Semester) of 0.80095, and Graduation GPA (Category) of 0.285926. Following are the pieces of data along with the variables used in this study.

3.2 Naïve Bayes Calculations

After doing various kinds of experiments, the distribution of the proportion of training data and data validation using Google collabs. Accuracy values and F1-scores for each proportion of data distribution are obtained as shown in **Table 13**.

Table 13. Comparison of Accuracy and F1-Score Model Naive Bayes

Data Training: Data Validation	Without Correlation		With Correlation	
	Accuracy	F1-Score	Accuracy	F1-Score
70: 30	0.71	0.42	1.00	1.00
80: 20	0.68	0.41	1.00	1.00
90: 10	0.72	0.42	1.00	1.00

The similarity in values between accuracy and F1-score indicates that your model has consistent performance across all classes. This means the model does not merely pick the majority class to predict correctly while ignoring the others. It can be seen in **Table 13** that the accuracy and F1-Score are better when using the selected attributes after the correlation test. This provides information that the classification accuracy of JMTI ITK students graduating by using additional attributes besides the length of study (semester), namely Joint Stage IP (IP TPB), Final GPA (IPK), and Graduation GPA (Category), produces the same results as the classification that has been carried out using the length of study (semester). Therefore, testing on data testing will use the attributes selected after conducting a correlation test, namely Joint Stage IP (IP TPB), Final GPA (IPK), Length of Study (Semester), and Graduation GPA (Category) will be carried out in one of the proportions.

3.3 Applied In Data Testing

Testing data that has been cleaned and deleted some of its attributes so that it has the same attributes as the training data. After that, applied testing data using the best existing model obtained previously to classify the timeliness of students from the batch of 2019 graduating. Attributes that will be used are Joint Stage IP (IP TPB), Final GPA (IPK), Length of Study (Semester), and Graduation GPA (Category), with a data proportion of 70:30, and it is also assumed that the length of study of all students is 7 semesters. The results of the testing data testing can be seen in **Table 14**.

Table 14. Slice of Model Test Results on Data Testing

NIM	IP TPB	IPK	Length of Study (Semester)	Graduation GPA (Category)	Study Timeliness
02191001	2.76	3.03	7	2	Not On Time
02191002	2.58	2.2	7	0	Not On Time
02191003	3.61	3.34	7	2	On Time
10191001	3.36	3.63	7	1	On Time
10191002	3.17	3.53	7	1	On Time
10191008	2.63	2.72	7	0	Not On Time
11191001	2.84	3.25	7	2	On Time
11191002	3.63	3.5	7	2	On Time
11191012	2.71	3	7	0	Not On Time
11191022	2.91	3.03	7	2	Not On Time

In the prediction using data in batch 2019, the results showed that 157 students were classified as graduating on time, and 21 students were classified as not graduating on time.

4. CONCLUSIONS

Factors that influence the timeliness of graduation of JMTI ITK students are IP TPB, Final GPA (IPK), and Graduation GPA (Category), which are attributes in the model that have passed the results of the correlation test, with overall data proportions of 70:30, 80:20, and 90:10 results in accuracy and a perfect F1-Score, namely 1. This is because grades during lectures affect student study time. A bad score (D or E) in a course, makes students need to repeat the course in the next 2 semesters. In addition, students cannot take courses that have graduation requirements in courses that the student repeats, this can hamper the student's study time to graduate.

REFERENCES

[1] W. N. Putri, "Pengaruh Media Pembelajaran Terhadap Motivasi Belajar Bahasa Arab Siswa Madrasah Tsanawiyah," *Lisania,* 2017.

[2] itk.ac.id, "Tentang ITK," 2022. [Online]. Available: https://itk.ac.id/home/tentang-itk/. [Accessed 21 September 2022].

[3] Wikipedia, "Institut Teknologi Kalimantan," 2022. [Online]. Available: https://id.wikipedia.org/wiki/Institut_Teknologi_Kalimantan. [Accessed 19 September 2022].

[4] D. Heryana, DATA MINING UNTUK MEMPREDIKSI KELULUSAN MAHASISWA PENDIDIKAN MATEMATIKA UIN RADEN INTAN LAMPUNG MENGGUNAKAN NAIVE BAYES, Lampung: UNIVERSITAS ISLAM NEGERI RADEN INTAN, 2019.

[5] N. M. M. Alfitri and E. Mashamy, "Klasifikasi Data Penduduk Untuk Menerima Bantuan Pangan Non Tunai Menggunakan Algoritma Naïve Bayes," *Jurnal Riset Komputer,* vol. 9, pp. 1035-1043, 2022.

[6] M. Y. Putra and D. I. Putri, "PEMANFAATAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI JURUSAN SISWA KELAS XI," *Jurnal Teknokompak,* vol. 16, pp. 176-187, 2022.

[7] N. Gogtay and U. Thatte, "Principles of Correlation Analysis," *Journal of The Association of Physicians of India,* vol. 65, pp. 78-81, 2017.

[8] D. Edelmann, T. F. Móri and G. J. Székely, "Onrelationships between the Pearson and the distance correlation coefficients," *Statistics and Probability Letters,* vol. 169, 2021.

[9] Bustami, "Penerapan Algoritma Naïve Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," *Jurnal Informatika,* pp. 884-898, 2014.

[10] A. Luque, A. Carrasco, . A. Martín and A. d. l. Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition,* vol. 91, pp. 216-231, 2019.

[11] K. Barkved, "How To Know if Your Machine Learning Model Has Good Performance," 9 March 2022. [Online]. Available: https://www.obviously.ai/post/machine-learning-model-performance.

[12] Harikrishnan, "Confusion Matrix, Accuracy, Precision, Recall, F1 Score," 2019. [Online]. Available: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd. [Accessed 24 September 2022].

[13] developers.google.com, "Machine Learning," 2022. [Online]. Available:https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data [Accessed 24 September 2022].

[14] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence,* vol. 5, pp. 221-232, 2016.

[15] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk and F. Herrera, Learning from Imbalanced Data Sets, Springer Cham, 2018.