Student Academic Evaluation using Naïve Bayes Classifier Algorithm

1st Haviluddin
Faculty of Computer Science and Information Technology
Universitas Mulawarman
Samarinda, Indonesia
haviluddin@unmul.ac.id

2nd Nataniel Dengen
Faculty of Computer Science and Information Technology
Universitas Mulawarman
Samarinda, Indonesia
ndengen@gmail.com

3rd Edy Budiman
Faculty of Computer Science and Information Technology
Universitas Mulawarman
Samarinda, Indonesia
edy.budiman@fkti.unmul.ac.id

4th Masna Wati
Faculty of Computer Science and Information Technology
Universitas Mulawarman
Samarinda, Indonesia
masnawati@fkti.unmul.ac.id

5th Ummul Hairah
Faculty of Computer Science and Information Technology
Universitas Mulawarman
Samarinda, Indonesia
ummulhairah@gmail.com

***Abstract*—One of the department tasks is to predict study duration-time of each student in order to anticipate dropout (DO), which causes the department performance to be poorly. Consequently, study duration-time of each student is indispensable. Furthermore, the evaluation showing whether the student will pass or fail would benefit the student/instructor and act as a guide for future recommendations/evaluations on performance. An in-depth study on the student academic evaluation techniques by using Naïve Bayes Classifier (NBC) has been implemented. The dataset with specific parameters among others age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and progress GPA (pGPA) and Total GPA (tGPA) of undergraduate program from semester 1-4 with three times graduation criteria (i.e., fast, on, and delay times) have been described and analyzed. The experimental results indicated that accuracy algorithm (AC) of 76.79% with true positive rate (TP) of 44.62% by using quality training data of 80% and 90% have a good performance accuracy value.**

Keywords—naïve bayes classifier, confusion matrix, student academic evaluation

I. INTRODUCTION

Currently, Indonesia has 24,539 department or study program which these fields are listed on Ministry of Research Technology and Higher Education Republic of Indonesia databases. One of the department tasks is to predict study duration-time of each student in order to anticipate dropout (DO), which causes the department performance to be poorly. Consequently, study duration-time of each student is indispensable. Furthermore, machine learning (ML) technique might be used in order to process the large database. This technique is considered appropriate in order to explore the information that contained in a large database, especially academic database [1].

Numerous machine learning (ML) methods can be used in classification [2]–[4], prediction [5]–[7], cluster [8], [9], etc., and still increasing interest in the research of data mining. In this article, the Naïve Bayes Classifier (NBC) method for predicting student academic performance has been utilized. Along with the development of machine learning, many researchers are trying to make classifications using the NBC technique in the field of education. [2] have implemented Naïve Bayes in learning data classification in order to Find Meaningful Pattern (FMP) for the students monitored at college level of the University of Rajasthan (UOR). The data set were contained 42 categorical/ nominal attributes i.e., father/mother occupation, caste, sub-caste, 10th-12th percentage, medium of education etc. The results showed that NBC model can be used for the future scheduling of student selection criteria at college level. [10] have explored C4.5 decision tree and k-means algorithms in order to find classification rules between student academic performance and master program. The data were collected from students attending the second and the third year of Informatics and Information and Communication Technologies branches at Faculty of Natural Science. The results confirmed that both techniques suggest helping students to focus on the area they are interested in. But, C4.5 decision tree and k-means algorithms were not used in this work. Then, [11] have been explored Naive Bayes, 1-NN and WINNOW algorithms in order to predict a student's performance. The dataset was collected from the 2 years academic of informatics course of the Hellenic Open University (HOU). The results indicated that these algorithms were the most appropriate to be used for the construction a software support tool. [12] have implemented Naïve Bayes algorithm for identified the slow learners and the performance of students from some private and government universities. The Naive Bayes algorithm was explored using Power BI application. The results showed that the student performance on the basis of different attributes like mobile phones, computer at home, and net access, board of student etc. and comparison of all the result in the form of chart have been able to identified. [13] also explored tree Naïve Bayes algorithm for student learning style of 46 undergraduate bioinformatics students for 7 weeks on genomic technology topics via Moodle. The results indicated that the tree augmented naive has higher precision than the Bayesian network.

Therefore, the purpose of this study is to scrutinize Naïve Bayes classifier (NBC) algorithm in order to student academic learning evaluate performance. It is expected that NBC model analysis result might be used in order to support academic decisions maker, especially head of department. Therefore, all students could improve and increase the learning process. This rest of paper is consisting of four sections. Section 1 is the motivation to do the writing of the article. Next, the NBC methodology and techniques is discussed in Section 2. Section 3 presents the experimental results and discussion, and finally Section 4 describes the research summaries and conclusion.

II. METHODOLOGY

A. Naïve Bayes Classifier (NBC) Algorithm

Naïve Bayes classifier (NBC) algorithm is a probability classifier that apply Bayes theorem with assuming a high independent. NBC is discovered by Thomas Bayes, the British scientist. In general, NBC theoretical is to predict future opportunities based on the experience in the past. In other words, all attributes will contribute in decision-making with equally weight attributes. Then, each attribute is independent of each other. The basic idea of Bayes' rule is a hypothesis (H) can be expected based on the evidence (E) that have been observed with (1) an initial probability [H or P (H)] is the probability of a hypothesis before evidence (*prior probability*) was observed, and (2) a final probability [H or P (H | E)] is the probability of a hypothesis after evidence (*posterior probability*) was observed. Furthermore, NBC general formula in (1).

$$P(H|E) = \frac{P(E|H) \text{ x } P(H)}{P(E)} \quad (1)$$

The NBC algorithm consists of two stages: learning and classify. First, learning the final probability P (Y|X) for each combination of X and Y have been prepared. Then, second, maximize P (Y|X) value for X classification based on Y value. The NBC formulation for classification can be seen in Eq. 2 and 3.

$$P(X|Y = y) = \prod_{i=1}^{11} P(X_i|Y = y) \quad (2)$$

With posterior probability

$$\prod_{i=1}^{11} P(X_i|Y = y)P(X|Y = y) = \alpha \text{ x } \frac{P(Y = y)}{P(y)} x \quad (3)$$

Where, *Y* is class output label; *y* is class values (quick, on-time, delay graduate); $P(X_1|Y = y)$ is sex probability value; $P(X_2|Y = y)$ is age probability value; $P(X_3|Y = y)$ is status probability value; $P(X_4|Y = y)$ is department probability value; $P(X_5|Y = y)$ is intake model probability value; $P(X_6|Y = y)$ is GPA Semester 1 probability value; $P(X_7|Y = y)$ is GPA Semester 2 probability value; $P(X_8|Y = y)$ is GPA Semester 3 probability value; $P(X_9|Y = y)$ is GPA Semester 4 probability value; and $P(X_{10}|Y = y)$ is activeness organization probability value. The flowchart of NBC can be seen in Fig. 1.

B. Datasets

In this study, the student dataset including of biographical, academic portfolios, course duration, and student participation in the organization's activities. The data were collected from 2014-2017 (279 samples data). Furthermore, the dataset contains including progress GPA (pGPA) and Total GPA (tGPA). Where, pGPA and tGPA are calculated from course subjects values. Before training, all datasets will be normalization by using cleaning, integration and transformation, Fig. 2. Furthermore, performance of NBC algorithm by using confusion matrix has been applied. Then, Rapid Miner Studio 7.3 software has been utilized in the process of calculation and modeling.

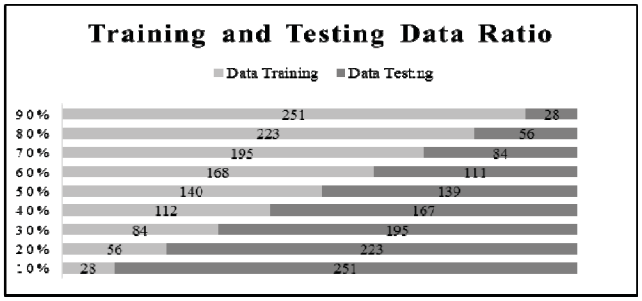

Fig. 1. Distribution of Training Data

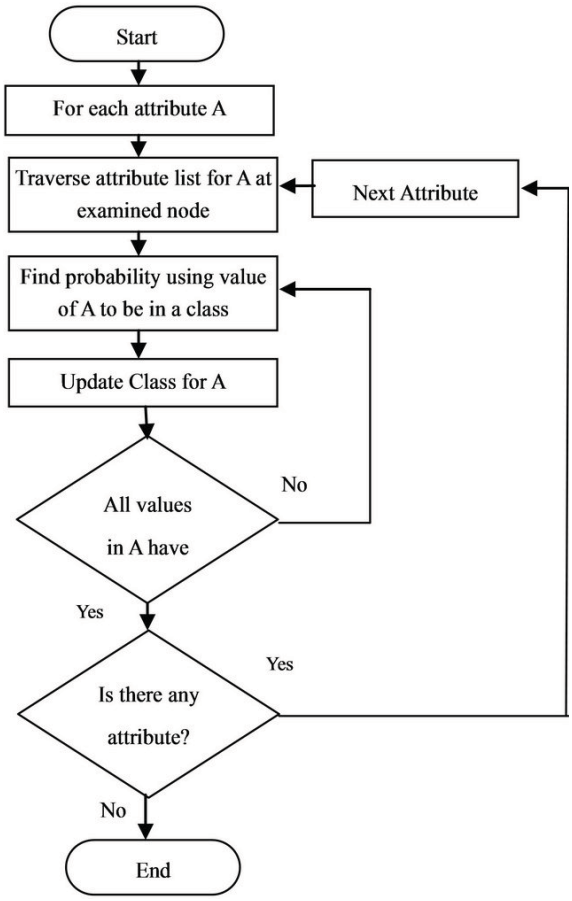

Fig. 2. NBC Flowchart [11]

C. Performance of Evaluation

In this study, confusion matrix (CM) for evaluation of NBC model has been used. Where, CM is a matrix of prediction that will be compared with the original class of input, Table I. In other words, the matrix contains the actual value information and predictions on the classification [14]. Then, the accuracy (AC) can be seen in Eq. 4.

$$C = \frac{a + b + c}{N} \quad (4)$$

TABLE I. CONFUSION MATRIX 2 CLASS

		Predicted	
		Negative	Positive
Actual	*Negative*	*a*	*b*
	Positive	*c*	*d*

III. RESULTS AND DISCUSSION

In this experiment, NBC for student academic performance was explored. Based on predetermined rules, nine training and testing classes' dataset have been established. In this experiment, the dataset among others students' academic performance evaluation variables including age, place of birth, gender, high school status (public or private), department in high school, organization activeness, age at the start of high school level, and pGPA and tGPA from semester 1-4. Furthermore, 10% to 90% of training data has been explored. Meanwhile, in order to get the best accuracy, CM as a performance of NBC algorithm by using three times criteria (i.e., *fast*, *on*, and *delay*) has been utilized, Table II and Fig. 3.

TABLE II. TRAINING AND TESTING DATASET

Confusion Matrix	Training Data		
	Fast-Time	On-Time	Delay-Time
10%	115	27	8
	19	11	46
	6	10	9
20%	98	13	2
	14	13	23
	13	16	31
30%	86	12	2
	10	15	8
	13	10	39
40%	74	8	4
	7	9	2
	13	15	36
50%	63	6	3
	3	7	0
	12	13	32
60%	49	7	1
	2	4	1
	11	10	26
70%	38	6	3
	2	5	0
	7	5	18
80%	26	3	1
	0	4	0
	5	4	13
90%	12	3	0
	0	0	0
	4	2	7

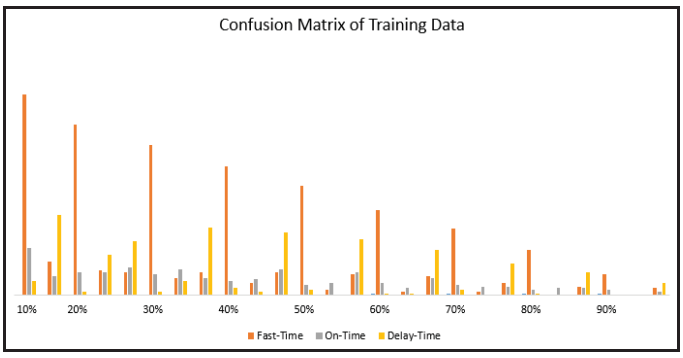

Fig. 3. Confusion Matrix of Training Data Plot

Based on experiment, confusion matrix of NBC algorithm shows that 76.79% algorithm accuracy (AC) with 44.62% true positive rate (TP) of 90% quality training data have best accuracy value. Means, that the best accuracy of NBC algorithm by using 80% and 90% of training data ratio, Table III and Fig. 4.

TABLE III. CONFUSION MATRIX NBC ALGORITHM

Algorithm Evaluation		Algorithm Accuracy (AC)	True Positive Rate (TP)
Training Data Ratio	10%	53.78%	42,38%
	20%	63,68%	54,80%
	30%	71,79%	64,79%
	40%	70,83%	64,10%
	50%	73,38%	71,21%
	60%	71,17%	66,14%
	70%	72,62%	70,76%
	80%	76,79%	81,92%
	90%	67,86%	44,62%

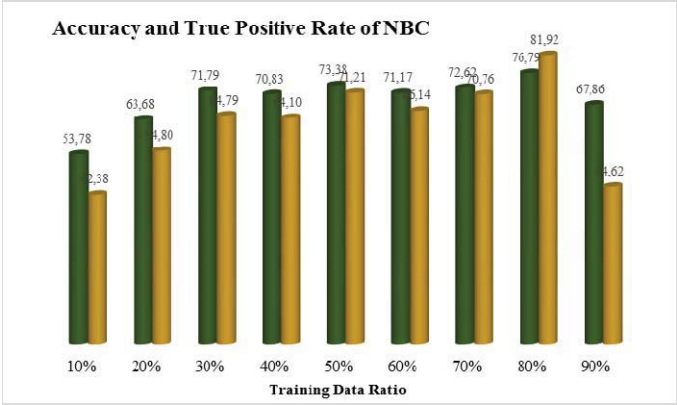

Fig. 4. Graphic of Confusion Matrix NBC algorithm

IV. CONCLUSION

In this paper, the analysis using NBC technique in order to achieve the model of students' academic performance have been implemented at the Faculty CSIT, Universitas Mulawarman. This study confirmed that NBC algorithm have an accuracy better on evaluate students' academic performance. In other words, NBC algorithm might be utilized as an alternative model in student academic evaluation performance. Therefore, other machine learning algorithms in order to get the better accuracy performance is a future works.

REFERENCES

[1] S. Roy and A. Garg, "Analyzing performance of students by using data mining techniques a literature survey," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, 2017, pp. 130–133.

[2] A. Dangi and S. Srivastava, "Educational data Classification using Selective Naïve Bayes for Quota categorization," in *Innovation and Technology in Education (MITE) in 2014 IEEE International Conference on MOOC*, 2014.

[3] E. Budiman, Haviluddin, N. Dengan, A. H. Kridalaksana, M. Wati, and Purnawansyah, "Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation," in *Lecture Notes in Electrical Engineering*, 2018, vol. 488, pp. 380–389.

[4] A. P. A. Harwati and F. A. Wulandari, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)," in *The 2014 International Conference on Agro-industry (ICoA): Competitive and sustainable Agroindustry for Human Welfare*, 2015.

[5] H. Haviluddin and R. Alfred, "Performance of Modeling Time Series Using Nonlinear Autoregressive with eXogenous input (NARX) in the Network Traffic Forecasting," *Proceeding IEEE*, pp. 164–168, 2016.

[6] T. Mahboob, S. Irfan, and A. Karamat, "A machine learning approach for Student Assessment in E-Learning Using Quinlan's C4.5, Naïve Bayes and Random Forest Algorithms," 2016.

[7] M. Pandey and S. Taruna, "Towards the integration of multiple classifier pertaining to the Student's performance prediction," *Perspect. Sci.*, vol. 8, pp. 364–366, 2016.

[8] Purnawansyah and Haviluddin, "K-Means clustering implementation in network traffic activities," in *Proceedings - CYBERNETICSCOM 2016: International Conference on Computational Intelligence and Cybernetics*, 2017.

[9] M. Karim and R. M. Rahman, "Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *J. Softw. Eng. Appl.*, vol. 6, pp. 196–206, 2013.

[10] A. Ktona, D. Xhaja, and I. Ninka, "Extracting Relationships Between Students'Academic Performance and Their Area of Interest Using Data Mining Techniques," in *2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks*, 2014.

[11] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education," *Knowledge-Based Syst.*, vol. 23, pp. 529–535, 2010.

[12] M. Tripathi, A. K. Agarwal, P. G. Scholar, and S. P. KNIT, "Probabilistic Determination of Student Performance using Naive Bayes Classification Algorithm," *Int. J. Eng. Sci.*, vol. 7, no. 8, pp. 14749–14752, 2017.

[13] L. X. Li and S. S. A. Rahman, "Students' learning style detection using tree augmented naive Bayes," *R. Soc. open Sci.*, vol. 5, no. 7, p. 172108, 2018.

[14] F. Gorunescu, *Data Mining (Intelligent Systems Reference Library Volume 12)*. Craiova: Springer, 2011.