Journal of Information Technology and Computer Science
Volume 4, Number 1, June 2019, pp. 15-25
Journal Homepage: www.jitecs.ub.ac.id

Intelligent Computing System to Predict Vocational High School Student Learning Achievement Using Naive Bayes Algorithm

Admaja Dwi Herlambang[1], Satrio Hadi Wijoyo[2], Aditya Rachmadi[3]

[1,2,3]Faculty of Computer Science, University of Brawijaya,
Veteran Road, 8, Malang, East Java, Indonesia

[1]herlambang@ub.ac.id, [2]satriohadi@ub.ac.id, [3]rachmadi.aditya@ub.ac.id

Received 14 October 2018; accepted 06 May 2019

Abstract. The success of the learning process in Vocational High School Information and Communication Technology Clusters (ICT Vocational School) can be shown from how much student achievement is achieved. Learning achievement is a complex variable, meaning that efforts to improve learning achievement are influenced by several factors. Sometimes students can fail or drop out. So that student learning achievements need to be grouped based on internal and external factors. The grouping process can use the Naive Bayes algorithm because the algorithm uses the Bayes theory by assuming all factors are not interconnected. The success of learning can be defined by how advanced the student's learning achievement. Achievement in learning can be affected by multitude factors, this makes measuring success in learning a complex activities. Factors that affect learning success of students, these factors can be grouped in two category, internal and external factors. The sample data used is the master data and academic data of the Information and Communication Technology (ICT) clusters of the 2014-2017 students who have already passed. This data has gender, age, mother's education, father's education, mother's occupation, father's job, reasons for choosing school, travel time to school, duration of study, failure in previous classes, school support, health, absenteeism, first year exam scores, second year exam value, third year value. Failure in learning can make student desperate, some of them even dropped out of school because of it. So that student learning achievements in the Vocational High School with ICT major need an intelligent computing system that could predict the student learning achievement. The system used fifteen achievement indicators and Naïve Bayes algorithm in data processing. Testing on student achievement data produces the conclusion that is the highest intelligent accuracy values in 83% with lowest accuracy value in 68% based on Naïve Bayes algorithm processing. The result of mining process using Naïve Bayes algorithm can be used to classify the 3rd year student achievement to five categories. These categories are Very Good, Good, Fair, Poor, and Failed. The system testing result showed that this intelligent computing system function was fitted with Vocational High School's system requirement, system design, and system implementation.

Keywords: *intelligent computing system, prediction algorithm, vocational high school, learning achievement, naïve bayes*

1. Introduction

Education is a conscious and planned effort to realize learning process so that the student will actively increase their potential to have religious spiritual strength, self-

control, personality, intelligence, noble character, and skills needed by themselves, society, also their nation and their country [1]. The success of learning can be defined by how advanced the student's learning achievement. Achievement in learning can be affected by multitude factors, this makes measuring success in learning a complex activities. Factors that affect learning success of students, these factors can be grouped in two category, internal and external factors [2].

Every students have varied learning difficulties level, and each of them has experienced learning difficulties. Failure in learning can make student desperate, some of them even dropped out of school because of it. Student's failure in finishing every subject can be caused by internal and external factors. Failure in learning can also be affected by teacher's limited ability in analyzing the factors that influence student behavior to predict student's success rate in solving schools subject problems. By overseeing this condition, research in predicting student's success rates during study period by looking at factors that can affect student's learning behavior is needed.

Nowadays, there are various classification algorithm used to predict student success rate in learning. One of them is Naïve Bayes classification algorithm. This algorithm use overall probability, namely document's probability against the category (prior). Document will be categorized based on its maximum probability (posterior). In other words, this algorithm assumes that presence or absence of certain features of the class is not related to the presence or absence of other features [3]. Because Naïve Bayes is a simple classification algorithm that applies Bayes theorem by assuming each feature are not related to each other.

A research has conducted to classify the abilities of secondary school students in Portugal using data mining. The research classifies students in Portugal who take subject such as mathematics and Portuguese by determining the factors that can affect student achievement in learning [4]. The study uses three different method of data mining to predict student's ability with different accuracy values, that is Naïve Predictor (60,50%-78,50%), Decision Tree (62,90%-76,10%), and Random Forest (33,50%-36,70%). Conclusion of this research define that Naïve algorithm have the best accuracy compared to other algorithms.

A similar research using classification algorithm in data mining to discover student learning retention using data mining technique. Student learning retention will be an indicator for measuring academic performance, it can also be used as a base for decision making by school management authority [5]. The study uses three classification algorithm, namely Naïve Bayes, Support Vector Machine and Decision Three. From the usage of those three algorithm, the most accurate are Naïve Bayes algorithm with 89,50% accuracy values, in the second place is Support Vector Machine with 83,50% accuracy values, and the last is Decision Tree with 81,30% accuracy values. Based on the description of current problems and the support of previous study related to the use of Naïve Bayes algorithm to predict student learning achievement, this research will propose a decision support system that aims to predict student learning achievement using Naïve Bayes Algorithm.

1.1 Learning Achievement

Learning achievement is the result that achieved by someone in their learning effort and stated in the report card. Achievement also could be defined as degree of perfection achieved by someone in thinking, feeling and doing [6]. Learning achievement is perfect if it composes of three aspect, namely cognitive, affective and

psychomotor, on the contrary learning achievement is not satisfactory if someone has not been able to meet the target in those three aspect. Based on its definition, it can be summarized that learning achievement is the level of humanity student in accepting, rejecting and assessing information obtained in the teaching and learning processes. A person's learning achievement is in accordance with the level of success in studying the subject which is expressed in report card after experiencing teaching and learning processes for each field of study. Student's learning achievement can be discovered after an evaluation. The result of evaluation can show the peak and the pitfall of a student's achievement in learning.

Learning is an effort to teach the student. While studying is an activity that produce new skill or abilities that are permanent in students. By looking learning and studying as a system, there are factors that affect learning and studying [7]. These factors can be described as internal and external factors. Internal factors are factors that come from within a person and can affect individual learning outcomes. This internal factors include physiological and psychological factors. Physiological factor is a factor that represent individual physical condition. While psychological factor is a person psychological state that can affect learning process. Some of the main psychological factor that influence the learning process are student intelligence, motivation, interests, talent, and attitudes. The factors that influence the result of studies on students are not only based on internal factors exist in the student but also from external factors that support the students. Some of the external factors that can affect student learning are gender, age, study time, failure in the previous subject or class, school support and family support.

1.2 Naïve Bayes Algorithm

Naïve Bayes is a simple classification that implement Bayes theorem with the assumption each of its feature are not related to each other. Bayes algorithm used all of its probability, namely document's probability against the category (defined as prior). The document/text will be categorized based on maximum probability (defined as posterior). In other word this method assumes that presence or absence of certain features of the class is not related to the presence or absence of other features [3]. Equation 1 and 2 define the formula of classification using Naïve Bayes Method. V_{NB} is output from the result of Naïve Bayes classification. $P(a_i|v_j)$ is ratio between n_c/n, whereas n_c is the amount of data training for $v = v_j$ and $a = a_i$. n is the total probability of output.

$$P(a_1, a_2, a_3, \dots, a_n|V_j) = \prod_{i=1}^{n} P(a_i|v_j) \quad (1)$$

$$V_{NB} = \arg\max_{v_j \in V} P(V_j) \prod_{i=1}^{n} P(a_i|v_j) \quad (2)$$

2. Research Method

In this research, the methodology of research or research stage will be represented with Figure 1. This research have four main phases. First phase is requirement analysis followed by system design, system implementation, and system testing. System implementation has processing phase. Processing phase has activities in accordance with the steps in data mining. The steps are data cleaning, data integration, selection of data, data transformation, and creating dataset. Dataset in this study are used as a training and testing data.

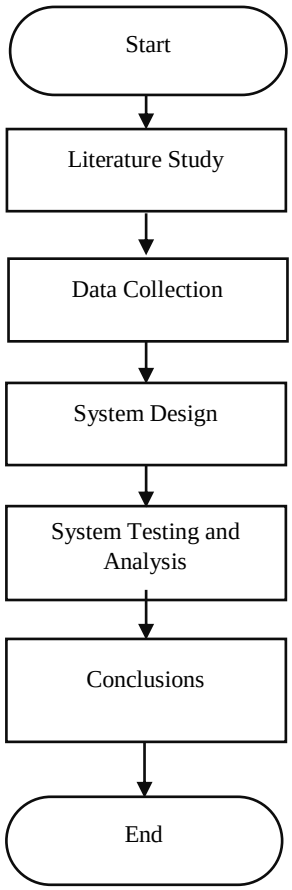

Figure 1. the methodology of research stage

The data used as a process in mining and testing is in the form of master data sample and academic data for the 2014-2017 generation of Information and Communication Technology (ICT) Vocational High School students who have passed. In this study, data from 3 Vocational High Schools in Malang City were used. The 3 Vocational Schools are Malang State Vocational High School 12, Malang State Vocational High School 3, and Singosari State Vocational High School 2. Each Vocational School is taken from students who have graduated between 100-150 data. So that the data to be collected is 450 data. The data have attributes like gender, age, mother education level, father education level, mother occupation, father occupation, reason for choosing school, time to go to school, time of study, failure in the previous class, school support, health, presence in school time, first year grade score, and second year grade score. This fifteen attributes are referenced from previous research and validated by Vocational High School teachers to adjust the attributes used in this study.

Target data is a data in the form of a sample of academic data from ICT Vocational High School students of the 2014-2017 generation. This data represent a third year exam score. The third year exam value data is used because students are declared achievers if they can graduate school and get a very good predicate.This exam score data has target

class obtained from Vocational High School teachers that can be defined as an experts. Classes or outputs of learning achievement are represented in classification form from the best to worst in accordance to the level of learning outcomes, these classification level are Very good, Good, Fair, Poor and Failed.

After dataset needed for this research is obtained, the following step is design stage. This stage are done by carrying out all of the design processes for the Vocational High School student learning achievement classification system. The design carried out includes the process of requirement analysis and designing algorithm. Requirement analysis is done to get the features and data needed by the systems, while designing algorithm is done to design the algorithm needed for the implementation of Naïve Bayes algorithm in the system.

This intelligent computing system has three main process, that are (1) training data retrieval from database, (2) Naïve Bayes algorithm processing, and (3) classification result. Training data retrieval is a training data retrieval process that used as calculation base using Naïve Bayes Algorithm method. Training data retrieval is very influential on system accuracy. Because the training data will be processed to calculate the probability rule in the Naïve Bayes algorithm. The amount of training data is influential to the performance of system. Naïve Bayes algorithm process is a simple classification that implement Bayes theorem with the assumption each of its feature are not related to each other. The outcomes from Naïve Bayes algorithm will be defined as reference for classification stage. The result of classification represent highest value in the calculation and it will determine which data will be entered in which class.

3. System Design and Implementation

3.1. System Design

The design of the system for predicting student achievement did not yet exist because school authority still did it manually. In this system design stage there are 60 amount of student data that are used as data training using Naïve Bayes algorithm. Student achievement prediction system is shown in Figure 2.

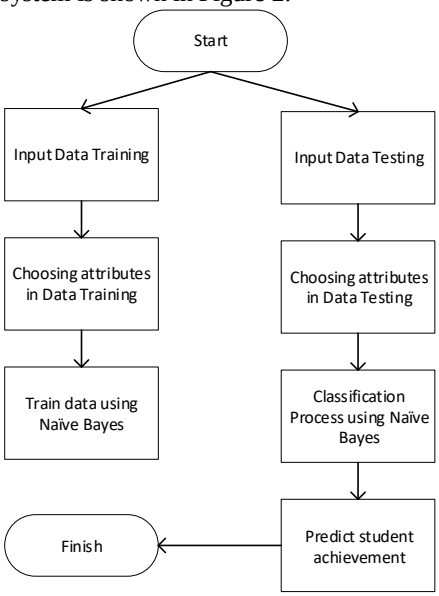

Figure 2. General description of the system to predict student achievement

Use Case Diagram is used to represent functional requirement for a system [Pressman]. Use Case for the student achievement prediction system has 5 function, these are processing student achievement, import data, export data, accessing data training based on amount of data, and accessing data testing. This 5 use case are derived from requirements that have been defined from problems.

First problem is there should be a function to connect calculation from Naïve Bayes Classification using 15 factors as indicators. Requirement to answer this problem is, 15 variables will be translated as a variable for classification using Naïve Bayes, the result are categories used as a decision support. From this requirement emerge use case Processing Student Achievement.

Besides having ability to produce classification the system must also be able to facilitate the work of teacher, it means if the data used are numerous then doing calculation one by one is inefficient. From this problem a need arise that can store many data simultaneously and data is stored in easy to read format like excel so it can be used in the future. From this requirement arise use case Export Data. The data that have been saved can be used as training data using system feature described in use case Import Data.

Use Case Accessing Data Training Based on Amount of Data and Accessing Data Testing is derived from requirement that teacher can access data training and data testing when using the System. Class diagram is a diagram that describe object in a system and its relation with other object.

3.2. System Implementation

This intelligent computing system interface is used by users to interact with software. The page interface is divided into 3 pages: the login page interface, the main page interface, and the user profile interface. The login page interface display shown in Figure 3 is the start page when the user opens the system, on this page there is a sign in button. The user must enter a username and password to sign in to the main page interface.

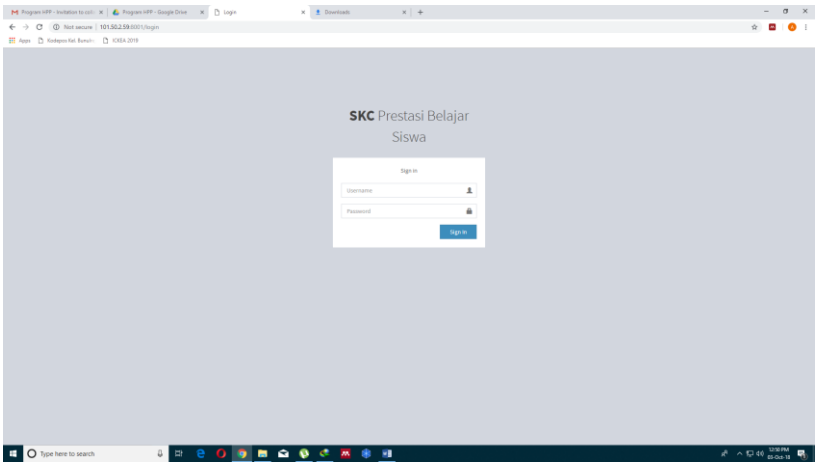

Figure 3. Login Page Interface

The main page interface display shown in Figure 4 is the second page when the user opens the system, on this page there are four buttons or menus namely Training, Testing, Testing, and Manage Data. In addition to these four buttons, users can enter student data to test intelligent computing systems.

The user profile interface shown in Figure 5 is the third page when the user opens the system, on this page there is a profile update button. The profile update button is used to change the user name, user username, and user password.

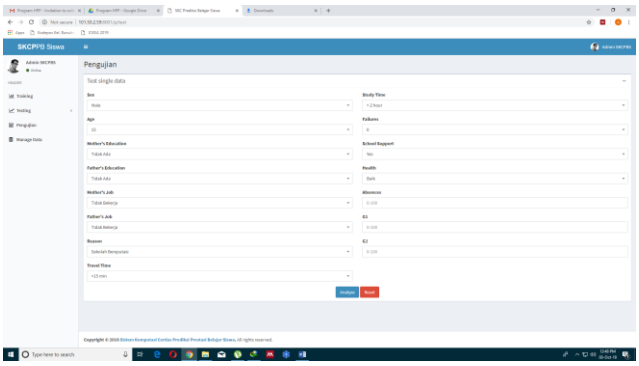

Figure 4. Main Page Interface

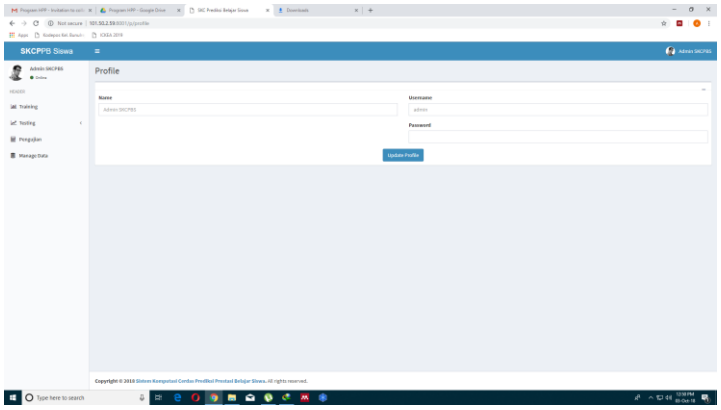

Figure 5. Profile Page Interface

4. Result and Analysis

Accuracy Testing

Evaluation is the activity of measuring the success of implementation result by comparing system performance with the desired result. In this system. The standard used in evaluating is by measuring system performance, namely accuracy. Accuracy is closeness a number as result of calculation to the actual number (true value or reference value). In this study accuracy testing performed to define performance of system in providing predictive conclusion. The accuracy calculation are done using formula (3).

$$Accuracy\ Value = \frac{Amount\ of\ data\ that\ is\ accurate}{Amount\ of\ all\ data} \times 100\% \quad (3)$$

Testing Scenario

Testing had been done as many as 4 times using the same training and testing data. First data testing uses 50 data with the number of different result of exam in each class at third year student. The second data testing uses 100 data with the number of each grade for the third year of exam is different. The third data testing uses 200 data with the number of each grade for the third year exam is different. The fourth data testing uses 300 data with the number of each grade for the third year exam is different. Training data uses data that numbered about 60 student data sample. Amount of training data is kept constant in each

experiment to test if the accuracy of the system prediction affected by the number of data used as data training.

The prediction result of 1st Experiment

Data training uses 60 data with 50 data in data testing. The process after importing data training, in application is shown in Figure 6.

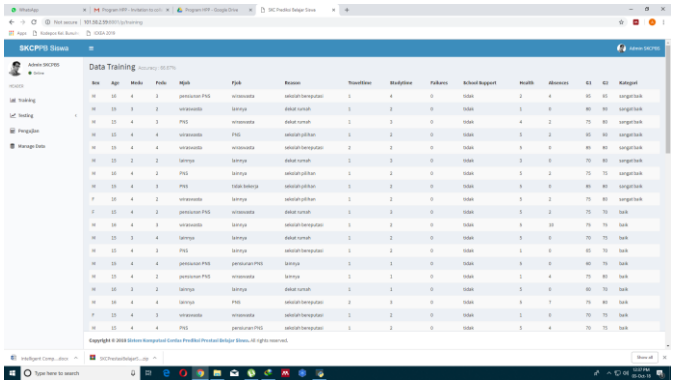

Figure 6. Data Training with 60 Data

The calculation in Figure 6 is used as a basis for making probability table that will be used in the data testing classification process. The result of application using 50 data testing are shown in Figure 7.

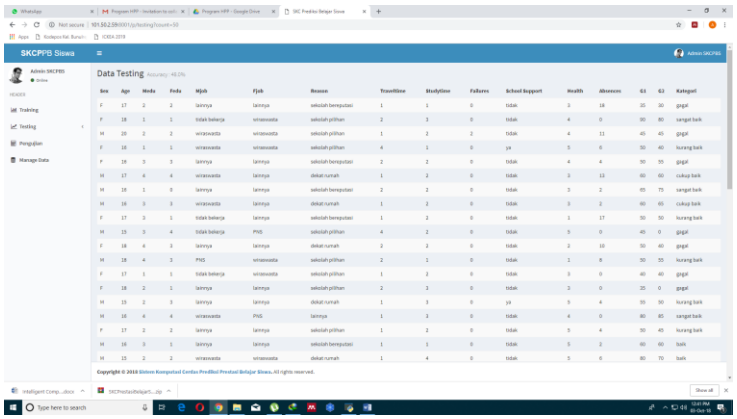

Figure 7. Data Testing with 50 Data

From the result of the 1st experiment, the accuracy value is 68%. A total of 34 data is in accordance with the target data. While 16 data from the experiment is not in accordance with target data.

The prediction result of 2nd Experiment

Data training used in 2nd experiment is 60 data with the amount of data testing at 100 data. The result in Figure 8 represent 100 data testing after experimentation based on calculation of application.

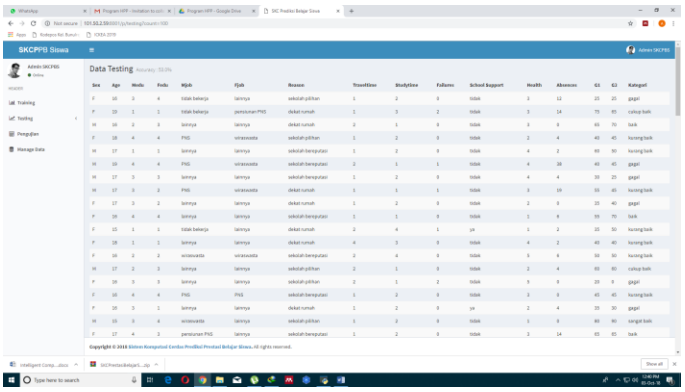

Figure 8. Data Testing with 100 Data

From the 2nd experiment, the accuracy value is 83%. A total of 83 data is in accordance with the target data. While 17 data from the experiment is not in accordance with target data.

The prediction result of 3rd Experiment

The 3rd experiment uses 60 data for data training with the amount data for testing data at 200 data. The result in Figure 9 represent the result of 200 data after experimentation based on calculation of application.

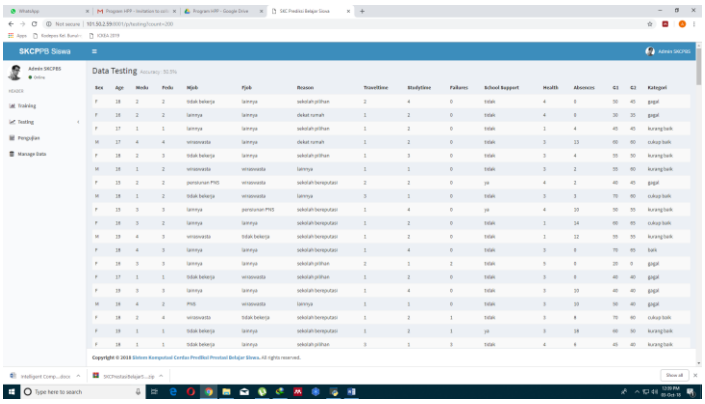

Figure 9. Data Testing with 200 Data

From the 3rd experiment, the accuracy value is 76%. A total of 152 data is in accordance with the target data. While 48 data from the experiment is not in accordance with the target data.

The prediction result of 4th Experiment

The 4th experiment uses 60 data for data training with the amount data for testing data at 300 data. The result in Figure 10 represent the result of 300 data after experimentation based on calculation of application.

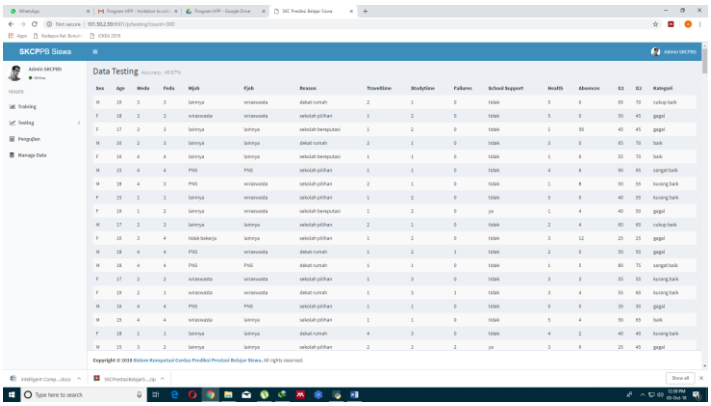

Figure 10. Data Testing with 300 Data

From the 4th experiment, the accuracy value is 78%. A total of 234 data is in accordance with the target data. While 66 data from the experiment is not in accordance with the target data.

Prediction Results of Accuracy Testing

The result of the testing process that had been carried out could determine the accuracy value of each experimentation. The highest accuracy value from experimentation is 83%, while the lowest is 68%. Comparison of accuracy value for each experimentation represent in Table 1.

Table 1. Comparison of Test Accuracy

Test Sequence	Data Testing	Valid Data	Invalid Data	Accuracy
1	50 Data	34 Data	16 Data	68%
2	100 Data	83 Data	17 Data	83%
3	200 Data	152 Data	99 Data	76%
4	300 Data	149 Data	151 Data	78%

In general, the result of testing represents moderate accuracy. This was due to the determinant of student achievement not only in exam score, but external factor can also contributed to the success of learning in this research. After testing the data, the best result of mining will be used to classified target data to class, namely "Very good" or "Good". Student achievement which become the target data will be evaluated based on academic historical data that already been taken by student and students external factors. In addition, the problem that occurs is the selection of the right training data can also affect the results of testing accuracy. In the study only used 60 training data.

5. Conclusion

Testing on student achievement data in Vocational High School with ICT majors on 2014-2017 generation using Naïve Bayes algorithm produces the conclusion that is the highest accuracy values in 83% with lowest accuracy value in 68%. Determining data training first can affect the result of testing process, because the pattern of training data will be used as a rule to determine classes in data testing. The magnitude or small percentage of accuracy level can also be influenced by the determination of training data. In this research the result of mining process using Naïve Bayes algorithm can be used to classify the 3rd year student achievement to five categories. Further research could be

focused at acceptance and success rate of intelligent computing system implementation in Vocational High School. Theoretical frameworks could be used by the researcher are technology and knowledge transfer.

References

1. Prihantoro, C. R.: The Perspective Of Curriculum In Indonesia On Environmental Education. International Journal of Research Studies in Education. 4(1), 77-83 (2015)
2. Hailikari, T., Tuononen, T., & Parpala, A.: Students' Experiences of The Factors Affecting Their Study Progress: Differences in Study Profiles. Journal of Further and Higher Education. 42(1), 1-12 (2016)
3. Yuan, L.: An Improved Naive Bayes Text Classification Algorithm in Chinese Information Processing. In: Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCSCT '10), pp. 267-269. Jiaozuo, P. R. China (2010)
4. Cortez, P., Silva, A.M.G.: Using Data Mining to Predict Secondary School Student Performance. In: Brito, A., Teixeira, J. (eds.) Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008). Porto, Portugal (2008)
5. Zhang, Y., Oussena, S., Clark, T., & Hyensook, K. Using Data Mining to Improve Student Retention in Higher Education: A Case Study. In: 12th International Conference on Enterprise Information Systems (ICEIS), pp. 1-8. Madeira. Portugal (2010)
6. Shieh, C.J. & Yu, L.: A Study on Information Technology Integrated Guided Discovery Instruction towards Students' Learning Achievement and Learning Retention. Eurasia Journal of Mathematics, Science & Technology Education. 12(4), 833-842 (2016)
7. Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., Willingham, D.T.: Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology. Psychological Science in the Public Interest. 14(1), 4-58 (2013)
8. Ramesh, V., Parkavi, P., & Ramar, K.: Predicting Student Performance: A Statistical and Data Mining Approach. International Journal of Computer Applications. 63(8), 35-39 (2013)
9. Al-Obeidat, F., Tubaishat, A., Dillon, A., & Shah, B.: Analyzing Students' Performance Using Multi-Criteria Classification. Cluster Computing. 20(78), 1-10 (2017)