Prediction of student learning outcomes using the Naive Bayesian Algorithm (Case Study of Tama Jagakarsa University)

Arini Aha Pekuwali

Department of Computer Science, Christian University of Wira Wacana Sumba, R. Soeprapo 35 Street, Prailiu, Waingapu – Sumba Timur

E-mail: arini_pekuwali@apps.ipb.ac.id

Abstract. Assessment of student learning outcomes is the most important part in the learning process. Student achievement can be determined based on the achievement of final grades in certain subjects. Student final grades can be used to evaluate and predict student achievement in the future. This research was conducted to analyze the Naïve Bayesian Classifier (NBC) algorithm in predicting the final grades of students in the future based on student final grade data in the previous semester. This study is useful for students to improve their grades, according to their predicted weaknesses through this research (wake-up calling). The results of this study indicate that NBC successfully classifies data with an accuracy of 94.2446%.

1. Introduction

The main purpose of higher education institutions is to provide quality education for students. One way to achieve the highest level of quality in the higher education system is to find knowledge from educational data to learn the main attributes that can affect student achievement.

Research using Data Mining in the field of education is increasing. It is a new field, called Educational Data Mining (EDM), which processes data from the educational context into information. The information obtained is useful for supporting decisions in the form of constructive recommendations to academic planners in higher education institutions, cutting failure rates, to better understand student behavior, and many other benefits.

The quality of higher education can also be seen from the value of student learning outcomes. In general, higher education already has an information system that is used to manage student learning outcomes data assessment. The information system manages values based on established assessment points. Assessment points used by each educational institution are governed based on academic regulations owned by the institution. Based on Government Regulation No. 49 of 2014 explains that the assessment of learning is a minimum criterion regarding the assessment of student learning processes and outcomes in the context of fulfilling graduate learning outcomes. Assessment is done based on the value of assignments, quizzes, homework assignments, midterms, final examinations. In its implementation, the assignment value is the combined value of several other tasks, such as quizzes, homework assignments, practical reports, and presentation values.
This study predicts the value of student learning outcomes in the current semester based on data on student learning outcomes in the previous semester. The previous semester's assessment data is called training data, and the data that will be predicted to its final value is called test data. The assessment data consists of a number of values namely attendance value, assignment value, Midterm Exam scores and Final Semester Exam scores of the students Tama Jagakarsa University.

The three algorithms commonly used in EDM are the Decision Tree, Neural Network, and Naïve Bayes Classifier (NBC) [1]. NBC provides the highest accuracy compared to the other two algorithms. The variables used in this classification are GPA, student demographics, high school background, scholarships, and social interaction [1]. The researcher proposes the use of the NBC Algorithm in this study. The NBC algorithm extracts patterns from the training data and the test data testing process.

The problems are:
• How to determine the assessment components of the student's final grade that can be used in predicting the value of student learning outcomes?
• How to extract and analyze the value of student learning outcomes?
• How is the accuracy of the algorithm Naïve Bayesian Classifier to predict the value of student learning outcomes?

The purposes of this study are:
• Knowing the assessment components that can be used in predicting the value of student learning outcomes.
• Extracting and analyzing the value of student learning outcomes.
• Knowing the accuracy of the algorithm Naïve Bayesian Classifier to predict the value of student learning outcomes.

2. Literature Review

2.1 Data Mining
The concept of data mining (DM) is extracting hidden patterns and trace relations between variables in a large data set [2]. Patterns are found used as the basis of decision making for the future. An important characteristic of DM is very large data volume [3].

DM divided into several stages:
a. Data cleansing
b. Data integration
c. Transformation of data
d. Application of DM techniques
e. Evaluation of the pattern that was found
f. Presentation of knowledge

DM includes several methods for the process of data analysis, namely classification, clustering and association rules [2]. Classification is predicting data testing based on patterns from training data. Clustering is a process to classify data that has characteristics that equal to one group, and the other to another group.

2.2 Education Data Mining (EDM)
Research Data Mining in the Education field now increases day by day. Education Data Mining is processing data from the educational field to more useful information. The information is useful for supporting to make a decision in the form of recommendations for the academic planner in university. EDM functions are cut the rate of failure, to better understand students behavior, and many benefits other [4].
Decision Tree, Neural Network, and Naive Bayes Classifiers are ordinarily used in EDM. Naive Bayes Classifier provides the highest accuracy compared to the other two algorithms. Variables of classification are student’s GPA, demographics of the student, scholarships and social interaction [1].

2.3 Naive Bayesian Classifier Algorithm
The classification method used in this study is Naïve Bayesian Classifier (NBC). The NBC method is one of the classification methods that assume all the attributes of the examples are independent of each other in the class context. Although in general, the assumptions are bad, in practice, the NBC method shows very good performance. Probability Bayes can be used to calculate the conditional probability, the chance of occurrence if an event is known [5]. This method can predict the likelihood of members of a class based on samples from members of the class [6]. Because it is assumed to be an independent variable, only the variance of a variable in a class is needed to determine the classification, not the whole of the covariance matrix.

The first step in this classification is to calculate the average value and standard deviation of the training data features in each class [7,8]. The average value and standard deviation are used to calculate the probability $P(x_i | C_i)$

$$P(x_i | C_i) = \frac{1}{\sqrt{2\pi} \sigma_{C_i}} e^{-\frac{(x_i - \mu_{C_i})^2}{2\sigma_{C_i}^2}}$$

(1)

where $g(x_i, \mu_{C_i}, \sigma_{C_i})$ is the Gaussian density function for the attribute $A_k$, during $\mu_{C_i}$ and $\sigma_{C_i}$ is the average and standard deviation, each of which gives the attribute value $A_k$ for the training data from the class.

After obtaining the probability $P(x_i | C_i)$ for each feature in each class, the probability value is $P(x_i | C_i)$ multiplied $P(X | C_i) = \prod_{i=1}^{n} P(x_i | C_i)$ so we get the probability $P(X | C_i)$ for each class. The probability of $P(X | C_i)$ multiplied by the prior probability of each class will produce a posterior probability $P(C_i | X)$.

$$P(C_i | X) = \frac{P(X | C_i) P(C_i)}{P(X)}$$

(2)

The class determination of the test data follows the rules of $P(C_j | X) > P(C_i | X)$, for $1 \leq j \leq m, j \neq i$.

Accuracy for classification results can be searched using the formula:

$$\text{Akurasi} = \frac{\sum_{\text{data uji benar}}}{\sum_{\text{data uji}}} \times 100\%$$

(3)

2.4 An important variable in predicting student performance
Shahiri stated that the variables most often used in predicting the performance of the student are GPA and values of college [1]. Ten of twenty-three journals reviewed by Shahiri been using GPA as a primary variable in predicting the performance of students [1].

Internal assessment tools are assignments, quizzes, practice, Middle Test, and Final Test. Furthermore, student demographic variables and external assessments. Demographics students include gender, age, and senior high school background. Men and women have a different style in the process of learning [9, 10].
3. **Research Method**

Figure 1 shows the research framework.

![Research Framework Diagram]

**Figure 1.** Research framework

3.1. **Distributing questionnaires**

Retrieval of data using a questionnaire is intended to know the point of view of Tama Jagakarsa University lecturers in assessing student performance. Student performance can be measured through GPA variables, internal assessment, student demographics, and external assessment. The questionnaire distributed was aimed to find out whether there were subjective factors that influenced the lecturers’ assessment of students.

3.2. **Collecting student data**

Student data collection according to predictive variables. Student data were obtained from the academic section of Tama Jagakarsa University.

3.3. **Data preprocessing and classification process**

Prediction variables are normalized, so they have the same scale; this stage is called data preprocessing. *K-fold cross-validation* is used to divide the data into a data train and test data. The method is doing the loop as much as *k* times to divide a set of examples is randomly into *k* - subset are mutually freely. Each retest is left with a subset for testing, and the rest is used for training. The results of the *training* data training stage are the characteristic patterns of data that become the reference stages of testing the test data.

3.4. **Research location**

This research was conducted at Tama Jagakarsa University.

3.5. **Software**

The software which is used in the research:
Microsoft Excel 2013
Tabulation of data is done using the device software Microsoft Office Excel 2013.

Matlab R2014b
Processing of the data train and test data and the application of Naïve Bayesian classifier using Matlab R2014b.

3.6. Variable selection
The data value of student learning outcomes at the University of Tama Jagakarsa, Faculty of Engineering, Department of Engineering Informatics. Variables that are used as follows:
1. Value of attendance
2. Quiz Value
3. Practical Value
4. Task Value
5. Mid Test Value
6. Final Test Value

4. Results and Discussion

4.1 Results

| Question                                                                 | Value | TP | J  | S  |
|-------------------------------------------------------------------------|-------|----|----|----|
| 1  I will give good grades to students who always arrive on time at lecture hours |       | 0  | 3  | 7  |
| 2  I will give fewer marks to students who often leave class without permission even though the lecture has not been completed |       | 0  | 3  | 7  |
| 3  I will give good grades to students who are always on time in completing the assignments |       | 0  | 5  | 5  |
| 4  I will give good grades to students who are active during lectures by either asking or answering questions |       | 2  | 2  | 6  |
| 5  I will give good grades to students who help friends who have difficulty in lecturing material |       | 1  | 4  | 5  |
| 6  I will give fewer marks to students who don't pay attention to lecture material |       | 3  | 5  | 2  |
| 7  I will give good grades to students who always pay attention to lecture material |       | 0  | 3  | 7  |
| 8  I will give good grades to students who are always polite in attending lectures |       | 0  | 7  | 3  |
| 9  I will give good grades to students who always record responses from lecturers and classmates |       | 0  | 4  | 6  |
| 10 I will give good grades to students who always permit when they want to leave the classroom and enter the classroom |       | 0  | 8  | 2  |

**Total**

6  44  50

**Description:**
TP: Never
J: Rarely
S: Often

4.2 Data processing stages
Data values are divided into the data train and test data using K-fold cross-validation. After that, the training data undergo a training process using the Naïve Bayesian Classifier algorithm. This process generates a training pattern. Then the process conducts a testing
The process of testing this using a pattern that is generated by the training process. The testing process results in classification accuracy.

**Table 2. Student grades**

| Courses                     | 2015/2016 | 2016/2017 | Sum |
|-----------------------------|-----------|-----------|-----|
| 1 Data Structure            | 23        | 20        | 43  |
| 2 Algorithm and Programming 1| 43        | 42        | 85  |
| 3 Compilation Engineering   | 28        | 19        | 47  |
| 4 Web Programming 3         | 29        | 15        | 44  |
| 5 Database 1                | 47        | 42        | 89  |
| 6 Information Systems Modeling| 26        | 16        | 42  |
| 7 Concept of Information Systems| 29        | 42        | 71  |
| 8 Analysis and Design Information Systems| 22        | 19        | 41  |
| 9 Algorithm and Programming 4| 19        | 20        | 39  |
| 10 Computer Network         | 24        | 28        | 52  |
| 11 Algorithm and Programming 2| 18        | 26        | 44  |
| 12 Assembly Language        | 23        | 41        | 64  |
| 13 Neural Network           | 21        | 15        | 36  |

**Total** 697

Table 2 displays the amount of data students grades. There were 13 courses with 697 total data. The total data collected is separated according to letter values (Table 3).

**Table 3. Student grade data according to letter grades**

| Letter Grades | Sum |
|---------------|-----|
| 1 A           | 300 |
| 2 B           | 211 |
| 3 C           | 125 |
| 4 D           | 11  |
| 5 E           | 50  |

**Total** 697

The number of data values the letter A is 300 the data or the amount of data most lots. Value letter D is the amount of data the bit, i.e., 11 data. Data grades students are divided into the data train and data test. Details of the amount of training data and data tests are presented in Table 4. Data sharing uses *K-fold cross-validation*.

**Table 4. Data training and testing**

| Data types     | Sum |
|----------------|-----|
| 1 Data Training| 558 |
| 2 Data Testing | 139 |

**Total** 697
Data that has been divided, then processed using the Naïve Bayesian Classifier Algorithm. The results of the test data classification are shown in the Confusion Matrix (Table 7).

### Table 7. Confusion matrix

| Actual | Prediction | A | B | C | D | E |
|--------|------------|---|---|---|---|---|
| A      | 58         | 2 | 0 | 0 | 0 | 0 |
| B      | 3          | 39| 0 | 0 | 0 | 0 |
| C      | 0          | 24| 24| 0 | 1 | 0 |
| D      | 0          | 0 | 0 | 0 | 2 | 0 |
| E      | 0          | 0 | 0 | 0 | 10| 0 |

The correct classification of grade A was 58 data and incorrect classification was 2 data. 39 data was classified into grade B, but 3 data was classified incorrectly to grade A. The correct classification of grade C was 24 data and incorrect classification was 1 data into grade E. There is no correct classification in grade D, and all of the data classified to grade E. All of data grade E had correct classification.

\[
\text{accuracy} = \frac{\text{Sum of true classified}}{\text{Sum of data}} \times 100\%
\]

\[
\text{accuracy} = \frac{131}{139} \times 100\% = 94.2446\%
\]

So, the accuracy of the classification results is 94.2446%.
5. Conclusions and Suggestions
DHDNS component values consist of Presence, Tasks, Mid Test, and Final Test. The level of accuracy of the classification of 94.2446% by using a measurement of the effectiveness Confusion Matrix. For the development, can add the amount of data, so that the pattern of values is more varied.

Acknowledgments
Thank you to DPRM Higher Education through a grant DIKTI PDP has financed this research.

References
[1] Shahiria A M, Husaina W and Rashida N A 2015 The third information systems international conference a review on predicting student’s performance using data mining techniques Procedia Computer Science 72 414-22
[2] Ahmad F, Ismail N H and Aziz A A 2015 The prediction of students’ academic performance using classification data mining techniques Journal Applied Mathematical Sciences HIKARI 9(129) 6415–26
[3] Hermawati F A 2013 Data Mining (Yogyakarta: CV. ANDI OFFSET)
[4] Abuteir M M and 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) 140-6
[5] Manning C D, Raghavan P and Schutze H 2009 An Introduction to Information Retrieval (Cambridge: Cambridge University Pr.)
[6] Rosen G, Garbarine E, Caseiro D, Polikar E and Sokhansanj 2008 Metagenome fragment classification using n-mer frequency profiles Advances in Bioinformatics
[7] Han J and Kamber M 2001 Data Mining: Concepts and Techniques (San Francisco: Morgan Kaufmann Publishers)
[8] Pekuwali A, Kusuma W A and Buono A 2018 Optimization of spaced k-mer frequency feature extraction using genetic algorithms for metagenome fragment classification Journal of ICT Research and Applications 12(2) 123-37
[9] U. bin Mat, Buniyamin N, Arsad P M and Kassim R 2013 An overview of using academic analytics to predict and improve students’ achievement: A proposed proactive intelligent intervention IEEE 5th Conference on Engineering Education (ICEED) 126–30
[10] Meit S S, Borges N J, Cubic B A and Seibel H R 2007 Personality Differences in Incoming Male and Female Medical Students Available from: https://files.eric.ed.gov/fulltext/ED490525.pdf