Implementation of Data Mining In Predicting the Study Period of Student Using the Naïve Bayes Algorithm

Sri Wahyuni¹, Murni Marbun²
¹ Departemen Teknik Computer Universitas Pembangunan Panca Budi Medan, Indonesia
² Department of Teknik Informatika STMIK Pelita Nusantara Medan, Indonesia

E-mail: sriwahyuni@dosen.pancabudi.ac.id

Abstract. The period of study of students is one of the parameters in the assessment of accreditation of university. Pembangunan Panca Budi University as one of the best institutions in North Sumatra needs to predict the period of study of students in order to know the strategies implemented so that students graduate on time. This study predicts the period of study using the Naïve Bayes Data Mining Algorithm method to students of the Pembangunan Panca Budi University of Computer System study program bachelor program with a study period of 4 years. Application designed using Naïve Bayes algorithm that works based on the shortest distance between two objects by determining the value of k. The value of k is a parameter to determine the closest distance between a new object and an old object. The data used as variables are student academic data, GPA for the first from 4 semesters, the average value of the National Examination during high school, and majors during high school. The results of the study showed that the data mining application that was designed could predict the period of study of students at the Pembangunan Panca Budi University.

1. Introduction
Data Mining is a semi-automatic process that uses statistical techniques, mathematics, artificial intelligence and machine learning to extract and identify potential and useful knowledge information stored in large databases. Having a very large amount of data, for example lecturers, staff, facilities and infrastructure in universities is currently required to have the ability to compete by utilizing all the resources they have. As mentioned earlier, the data mentioned is student data in the data set, for example data on potential students dropping out. Understanding the profile of potential drop out students can be done by revealing knowledge, also used to find out more about the profile of a student who is going to take the final examination whether it tends to fail or succeed. Knowledge possessed to understand, predict, and prevent failure is very important for the management of higher education, especially University of Pembangunan Panca Budi. Previous research was conducted using the Decision Tree C4.5 method where it is known that the most influential variables are from school origin [1]. Naïve Bayes algorithm works based on the shortest distance between two objects by determining the value of k. The k value is a parameter used in determining the closest distance between a new object and the previous object [2].
This knowledge can be used in helping lecturers to get to know the situation of their students, and can be used as early knowledge in the decision making process. Measures of student success or achievement can be seen from the Grade Point Average (GPA) reflecting all grades obtained by students until the
current semester. GPA is obtained by summing all the quality values of all courses taken and dividing by the total credits (semester credit units).

By using data mining techniques, such as classification algorithms, it is possible to find characteristics of student achievement and use their characteristics to predict future achievements. The research was conducted at University of Pembangunan Panca Budi where student data obtained from the last five years of Pembangunan Panca Budi University are illustrated in table 1 below:

Table 1. Student Data Year 2010 until Year 2015

| Year | Number of Student Entering | Number of Student Graduating |
|------|----------------------------|------------------------------|
| 2010 | 1776 person                | 199 person                   |
| 2011 | 1979 person                | 630 person                   |
| 2012 | 2155 person                | 83 person                    |
| 2013 | 1914 person                | 166 person                   |
| 2014 | 3286 person                | 81 person                    |
| 2015 | 3774 person                | -                            |

From the table above it can be clearly concluded that every year there is an increase in the number of students drop out. Then an application is needed that can predict the study period of students to reduce student numbers drop out. Data mining methods are often applied at the University level to analyze available data about academic data and extract new information and knowledge in support of decision making [3].

2. Research Methods

The stages of the research conducted in the Data Mining Implementation study in predicting the study period of students with the Naïve Bayes algorithm as follows:

![Research Framework](image)

Figure 1. Research Framework
2.1 Data Collection
The data used in this study were obtained from University of Pembangunan Panca Budi. The data used as samples is 20 data. The data are 10 testing data and 10 data. Details of the data training used can be seen in Table 2.

| Name  | GPA Semester 1 | GPA Semester 2 | GPA Semester 3 | GPA Semester 4 | Average National Examination Score | High School Major | Studi Period |
|-------|----------------|----------------|----------------|----------------|------------------------------------|------------------|--------------|
| Andi  | 3.00           | 3.11           | 1.89           | 1.70           | 85                                 | Science          | > 4 Years    |
| Amisa | 2.72           | 2.32           | 1.89           | 1.60           | 78                                 | Social           | > 4 Years    |
| Anggi | 3.28           | 3.37           | 3.32           | 2.00           | 87                                 | Science          | <=4 Years    |
| Farhan| 3.61           | 3.16           | 2.89           | 3.11           | 82                                 | Science          | <=4 Years    |
| Fatin | 3.50           | 3.42           | 3.27           | 2.84           | 89                                 | Social           | <=4 Years    |
| Siaka | 2.83           | 2.63           | 2.58           | 2.25           | 78                                 | Science          | > 4 Years    |
| Fajar | 1.50           | 1.89           | 1.93           | 1.44           | 81                                 | Social           | > 4 Years    |
| Widya | 3.06           | 3.89           | 3.53           | 3.55           | 89                                 | Social           | <=4 Years    |
| Nikita| 3.11           | 3.21           | 2.89           | 3.22           | 88                                 | Science          | <=4 Years    |
| Edi   | 2.17           | 2.16           | 1.29           | 1.65           | 76                                 | Science          | > 4 Years    |

After getting the data to be used as training data, then test data or new data will be made which is seen in the table below and determined whether the new data is at which level.

| No  | Name | GPA Semester 1 | GPA Semester 2 | GPA Semester 3 | GPA Semester 4 | Average National Examination Score | High School Major | Studi Period |
|-----|------|----------------|----------------|----------------|----------------|------------------------------------|------------------|--------------|
| 1.  | Agus| 3.97           | 3.79           | 3.88           | 3.35           | 90                                 | Science          | -            |

Before searching the study period value for testing data, value limits will be determined first for 1-4 semester GPA and national exam scores as shown in table 4 below.
Table 4. Value Limit for 1 - 4 Semester GPA and Average National Examination Score

| No | Attribute                          | Limit          |
|----|------------------------------------|----------------|
| 1. | GPA Semester 1                     | <= 2.8, 2.81-3.2, >=3.21 |
| 2. | GPA Semester 2                     | <= 2.8, 2.81-3.2, >=3.21 |
| 3. | GPA Semester 3                     | <= 2.8, 2.81-3.2, >=3.21 |
| 4. | GPA Semester 4                     | <= 2.8, 2.81-3.2, >=3.21 |
| 5. | Average National Examination Score | <= 80, 81-86, >=87 |

After carrying out the value threshold for criteria that has an integer value type then the classification process will be carried out using the naïve bayes method based on existing training data as will be displayed in the process below.

3. Discussion And Result

3.1 Manual Calculation

P (Ci)

P (Study Period > 4 years) = 5/10 = 0.5
P (Study Period <= 4 years) = 5/10 = 0.5
P (X|Ci)

Testing Data for Semester 1 is: 3.97, so the threshold taken is >= 3.21
1. GPA Semester 1
   P (GPA Semester 1 = “>= 3.21” | Study Period = “> 4 years “) = 0 / 5 = 0
   P (GPA Semester 1 = “>= 3.21” | Study Period = “<= 4 years “) = 3 / 5 = 0.6

2. GPA Semester 2
   Testing Data for Semester 2 adalah: 3.79, so the threshold taken is >= 3.21
   P (GPA Semester 2 = “>= 3.21” | Study Period = “> 4 years “) = 0 / 5 = 0
   P (GPA Semester 2 = “>= 3.21” | Study Period = “<= 4 years “) = 4 / 5 = 0.8

3. GPA Semester 3
   Testing Data for Semester 3 adalah: 3.88, so the threshold taken is >= 3.21
   P (GPA Semester 3 = “>= 3.21” | Study Period = “> 4 years “) = 0 / 5 = 0
   P (GPA Semester 3 = “>= 3.21” | Study Period = “<= 4 years “) = 3 / 5 = 0.6

4. IP Semester 4
   Testing Data for Semester 4 adalah: 3.35, so the threshold taken is >= 3.21
   P (GPA Semester 4 = “>= 3.21” | Study Period = “> 4 years “) = 0 / 5 = 0
   P (GPA Semester 4 = “>= 3.21” | Study Period = “<= 4 years “) = 2 / 5 = 0.4

5. National Examnition
   Testing Data for the average national examination score is: 90, so the threshold taken is >= 87
   P (National Examination Score = “>= 87” | Study Period = “> 4 years “) = 0 / 5 = 0
   P (National Examination Score = “>= 87” | Study Period = “<= 4 years “) = 3 / 5 = 0.6

6. Major
   Testing Data for major is: Science
   P (major = “Science” | Study Period = “> 4 years “) = 3 / 5 = 0.6
   P (major = “Science” | Study Period = “<= 4 years “) = 4 / 5 = 0.8
   P(X | Study Period > 4 years) = 0 * 0 * 0 * 0 * 0.6 = 0
   P(X | Study Period <= 4 years) = 0.6 * 0.8 * 0.6 * 0.4 * 0.6 * 0.8 = 0.055296
\[ P(X|Ci) \times P(Ci) \]
\[ P(\text{Study Period} \geq 4 \text{ years}) \times P(X | \text{Study Period} \geq 4 \text{ years}) = 0 \times 0.5 = 0 \]
\[ P(\text{Study Period} \leq 4 \text{ years}) \times P(X | \text{Study Period} \leq 4 \text{ years}) = 0.055296 \times 0.5 = 0.0276 \]

So for existing data testing, the study period is included in the classification \( \leq 4 \text{ years} \).

### 3.2 Software Implementation

For Input Training data, Training data inputting form, there are data that must be filled by an administrator, namely Student Identity Number, name of student, GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, average national examination score, high school majors, and study periods.

Display testing data processing functions for the admin to display the testing data that has been inputted on the Testing form. The data displayed is Student Identity Number, name of student, GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, average national examination score, high school majors, and study periods. Display output testing processing data can be seen in Figure 2.

![Aplikasi Data Mining Untuk Memperediksi Masa Studi Mahasiswa Menggunakan Algoritma Naive Bayes](image)

**Figure 2. Training Data Processing**

The Input Data Testing page is an image display for Input Data Testing. For the testing data inputting form, there are data that must be filled by an admin, namely Student Identity Number, name of student, GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, average national examination score, high school majors. Input Data Testing page is an image display for input data testing. For the testing data inputting form, there are data that must be filled by an admin, namely Student Identity Number, name of student, GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, average national examination score and high school majors.

Display testing data processing functions for the admin to display the Testing data that has been inputted on the Testing form. The data displayed is Student Identity Number, name of student, GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, average national examination score, high school majors, and study periods. Display testing data processing output. The data testing processing form functions for the admin to display the Testing data that has been inputted on the Testing form. The data displayed is Student Identity Number, name of student, GPA semester 1, GPA semester 2, GPA semester 3, GPA semester 4, average national examination score, high school majors, and study periods.
2. GPA semester 3, GPA semester 4, average national examination score, high school majors, and study periods.

Figure 3. Testing Data Processing

This page displays the classification process using the Naive Bayes method where the table column displayed is student identity number, student name, probability of Naive Bayes probability for study period over 4 years, Naive Bayes probability value for the 4-year study period, and mass classification studies.

Figure 4. Mining Data Classification Process

4. Conclusion
Implementation of Data Mining with the Naïve Bayes Algorithm can predict the period of study the students of Computer Systems study program at the Pembangunan Panca Budi University. Based on a sample of 10 testing data with the variables used, the predicted results of the study period were exactly
4 years if the student came from the IPA department, the national examination scores 9 and IP scores above 3.0. The results of the implementation of a data mining system that was designed can be utilized to create a student graduation strategy on time and emphasize drop-out students.

References
[1] Wahyuni, S. 2018. Implementasi Metode Decision Tree C4.5 Untuk Menganalisa Mahasiswa Drop Out. ETHOS Jurnal Penelitian dan Pengabdian. Vol. 6, No.1:39-48, 2018
[2] Nur Sukma. 2017. KLasifikasi Calon Pendonor Darah Dengan Metode Naive Bayes Classifier. Malang: Universitas Brawijaya Malang
[3] D. Kabakchieva. 2013. Predicting Student Performance by Using Data Mining Methods for Classification. Sofia University
[4] Hariyanto, Eko and Rahim, Robbi. Arnold’s Cat Map Algorithm in Digital Image Encryption. International Journal of Science and Research (IJSR), Vol 5 Issue 10, 2016, pp 1363-65.
[5] Yeni, 2011, “Algoritma Data Mining”. Penerbit Andi.
[6] Wahyuni, S. 2017. Implementasi Rapidminer Dalam Menganalisa Data Mahasiswa Drop Out. Medan: Jurnal Ilmiah Abdi Ilmu. Vol. 10, No.2:1899-1902