Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU)

M T Sembiring*, R H Tambunan

Industrial Engineering Department, Faculty of Engineering, Universitas Sumatera Utara, Medan, Indonesia

*E-mail: meilita_tryana@yahoo.co.id

Abstract. One indicator of higher education performance is student graduation presentation on time with a passing percentage of more than or equal to 50%. The abundance of data in the Higher Education Academic Information System (SIA) is not used in measuring the performance of a College, especially the Study Program. The Industrial Engineering Department summarizes student data in the SIA as a database and processes it according to what is needed. During this time, the data were abundant in SIA has not been utilized to assess student performance. When the Study Program conducts a student academic evaluation, the Study Program evaluates it based on the variable number of credits, GPA, and constraints. In the Study Program, an appropriate classification is needed to make it easier to evaluate student performance. The method that is used is the Naïve Bayes algorithm in classifying students’ academic performance USU Department of Industrial Engineering. Attributes are used, among other NIM, Name, Gender, Region of Origin, Origin School, Entrance, and GPA. From the results of testing 173 sample data using the Naïve Bayes algorithm, the pattern formed has a match accuracy of 70.83%, which means that it is useful in predicting student graduation.

1. Introduction

Advances in information technology have overgrown in all areas of life. Many data is generated by information technology, ranging from industry, economics, science and technology, and various other life fields. These data are abundant but still hidden so that they are not appropriately utilized [1].

In industry 4.0, technology has a positive impact on organizations such as increasing productivity and increasing communication and coordination vertically (within the organization) and horizontally (between organizations, suppliers, and customers). Therefore, to obtain this, the adoption of industry 4.0 technology is required. Similar to other technology adoptions, industry 4.0 adoption requires a strategy for adoption. Formulation of this strategy, an evaluation of the organization's existing conditions is needed [2].

Big data processing is also still hidden, done through the data mining process, or is commonly called data mining. As part of Knowledge Discovery in Database, data mining is an activity of extracting high-value information or knowledge from large databases. The working concept specifies the patterns found in data sets to be more useful in real life [3].

The development of the business scope shows a significant increase in the growth of communication. It makes data access easier for organizations or companies, thus giving effect to database management changes to an advanced level supported by data warehousing facilities and web-based databases as an organizational strategy to improve organizational performance and profit.
The Industrial Engineering Department summarizes student data in the SIA as a database and processes it according to what is needed. So far, the abundant data in SIA has not been used to assess student performance. When the Study Program conducts a student academic evaluation, the Study Program evaluates it based on the variable number of credits, GPA, and constraints. Meanwhile, other AIS data variables have not been used to assess student performance, such as gender, school origin, regional Origin, entry route, and even class groupings based on entry pathways. In the Study Program, an appropriate classification is needed to make it easier to evaluate student performance. The method used is the algorithm Naïve Bayes classifying students' academic performance USU Department of Industrial Engineering.

Some of the preliminary studies that have been carried out are related to research topics, including research utilizing evaluating student academic performance by applying the Naïve Bayes classifier as a method by using the achievement index indicator as a determinant of whether or not student graduation is appropriate at University of Brawijaya [5]. Research with the application of data mining for mapping students needs guidance and counseling using the naïve Bayes algorithm at the State University of Padang State University [6]. This research is expected to help the department's performance increase the number of student graduations on time by considering research analysis and evaluation.

2. Research Methodology

2.1. Place and Time of Research
This research was conducted at the Department of Industrial Engineering, Faculty of Engineering, the University of North Sumatra, which processed industrial engineering student data. The location is located on Jalan Almamater USU Padang Bulan, Medan Baru District, Medan City, North Sumatra 20222, from February to present.

2.2. Type of Research
Quantitative research is research based on the philosophy of positivism, which is used to examine specific populations or samples using quantitative data collection and analysis instruments [7]. This study seeks to predict graduation by paying attention to the accuracy level to get recommendations for improvement.

2.3. Research Object
Academic performance of USU Industrial Engineering Department students.

2.4. Research Variables
There are two variables, namely the dependent and independent variables, with their respective definitions where the dependent variable is a variable that is influenced by other variables. Meanwhile, the independent variable is an independent variable that is not influenced by other variables, and its nature affects the dependent variable.

2.4.1. Dependent Variable
Students graduating on time (on time) and late (not on time). Determination of whether a student is right or late is determined by the length of the student's study.

2.4.2. Independent Variable
1. Student identification number (student ID)
2. Attribute the student's full name (name)
3. Gender as the M / W categorical independent variable
4. Regional Origin as the independent variable for the Medan and outside Medan categories
5. Origin of school as the independent variable for MAN, SMAN, SMAS, SMTA Others
6. Entry path as SNMPTN, SPMB, and SPMPRM categorical independent variables.
7. GPA (Grade Point Average) as a numerical independent variable.

The study results are expressed by a specific index accumulated from the credits and materials taken. CPI shows the results of investigations by the exams were conducted while taking courses calculated by the following formula:

\[
\frac{\text{Number of Credits} \times \text{Weighted Value}}{\text{The Number of Credits Collage}}
\]

2.5. Thinking Framework
The frame of mind is an explanatory frame of mind and is used to solve problem-solving [8]. This framework shows the researcher's line of thought in carrying out the research process to achieve the expected output. The conceptual framework image in this study can be seen in Figure 1.

![Figure 1. Framework for thinking research](image)

2.6. Data Collection Method
The data used in this study is secondary data obtained from the database of students of the Industrial Engineering Department, University of North Sumatra. Datasets are used as follows: First, the student data with 80 attributes and 483 records comprising no, ID, name, gender, city of birth, date of birth, religion, home school, graduation year, parents' names, and so on. Second, the graduation data consists of 21 attributes and 178 records comprised of a number, ID, name, weight, height, study program, length of study, GPA, a predicate of graduation, etc.

2.7. Data Processing Method
The data processing carried out in predicting student graduation on time can be broadly explained by the block diagram as follows:

![Figure 2. Block diagram of research](image)
3. Result and Discussion

The Naïve Bayes algorithm is one of the leading learning machine learning methods with the maximum effective and efficient data mining results. This algorithm's competitive performance is recognized in the classification even though it assumes that attribute associations are not calculated [9]. This assumption is rare. If there is a violation of this condition, naïve Bayes classification performance is still high, supported by empirical evidence.

In the classification process, each row's class label is owned as the value of the class variable in question. The calculation of the probability value can be carried out. Naïve Bayes assumes each free attribute, and then the following equation is obtained:

\[
\text{Chance } p(C = c_i \mid X = x_j), \text{ wherein Naïve Bayes, class } C \text{ is qualitative, while attribute } X_i \text{ can be qualitative or quantitative.}
\]

- Attribute \( X_i \) with quantitative type makes the probability \( p(X = x_i \mid C = c_j) \) very small. The probability equation cannot be relied on for the case of a quantitative type attribute [10].

Data processing using the Naïve Bayes algorithm requires training data [11].

### Table 1. Data training

| No | ID       | Name          | Gender | Origin     | School Origin | Entrance | GPA     | Length of Study |
|----|----------|---------------|--------|------------|---------------|----------|---------|-----------------|
| 1  | 130403001| Esa Delviana  | F      | OUTSIDE MEDAN | SMAN          | SNMPTN     | 3.30137 | LATE            |
| 2  | 130403002| Ulfa Audina   | F      | MEDAN       | SMAN          | SNMPTN     | 3.47000 | LATE            |
| 3  | 130403004| Ummu Habibah | F      | OUTSIDE MEDAN | MAN           | SNMPTN     | 3.32877 | LATE            |
| 4  | 130403005| Wildani Kurniasari | F | OUTSIDE MEDAN | MAN           | SNMPTN     | 3.40753 | LATE            |
| 5  | 130403006| Siti Khairennisa | F | OUTSIDE MEDAN | SMAN          | SNMPTN     | 3.60274 | LATE            |
| 6  | 130403007| Agni Fitaloka | F      | OUTSIDE MEDAN | SMAN          | SNMPTN     | 3.46000 | ON TIME         |
| 7  | 130403008| Wawan Andrian | M      | MEDAN       | SMAN          | SNMPTN     | 3.23630 | ON TIME         |
| 8  | 130403010| Aji Prasetyo  | M      | MEDAN       | SMAN          | SNMPTN     | 3.35274 | ON TIME         |
| 9  | 130403011| Akbar Al Ayubi | M       | OUTSIDE MEDAN | SMAN          | SNMPTN     | 3.24000 | ON TIME         |
| 10 | 130403013| Intan Hartanti| F      | MEDAN       | SMAN          | SNMPTN     | 3.53425 | LATE            |
| 11 | 130403014| Muhammad Burha | M       | OUTSIDE MEDAN | MAN           | SNMPTN     | 3.29110 | LATE            |
| 12 | 130403015| Abdi Santoso  | M      | OUTSIDE MEDAN | MAN           | SNMPTN     | 3.17466 | LATE            |

125 training data have been applied in the data cleaning process. Then from the table above, it can be calculated using the Naïve Bayes Classification formula, as for how it works as follows:

For classification problems, what is calculated is \( P(H \mid X) \), which is the probability that the hypothesis is correct (valid) for the observed sample X data.

\[
P(X \mid H) = \frac{P(X \mid H)P(H)}{P(X)} \tag{1}
\]

Where:
- X: sample data with a new class (label).
- H: a hypothesis that X is a class of data (label) C. \( P(H) \) is the probability of the hypothesis H.
- \( P(X) \): the probability of the observed sample data.
- \( P(X \mid H) \): the probability of sample data X, if it is assumed that the hypothesis is correct (valid).

The summary of calculation results can be seen in the following table.
| Table 2. Summary of probability calculation results for each categorial attribute |
|------------------------------------------|----------|----------|
| **P(PS)**  | **ON TIME** | **LATE** |
| P(PS)       | 0.304     | 0.696    |
| **Total**   | 1.000     |          |
| **P(JK)**   | **ON TIME** | **LATE** |
| W           | 0.500     | 0.471    |
| M           | 0.500     | 0.529    |
| **Total**   | 1.000     | 1.000    |
| **P(AD)**   | **ON TIME** | **LATE** |
| MEDAN       | 0.368     | 0.368    |
| Out side MEDAN | 0.632 | 0.632    |
| **Total**   | 1.000     | 1.000    |
| **P(AS)**   | **ON TIME** | **LATE** |
| SMAN        | 0.474     | 0.644    |
| MAN         | 0.000     | 0.080    |
| SMAS        | 0.316     | 0.184    |
| SMTA etc    | 0.211     | 0.080    |
| **Total**   | 1.000     | 0.989    |
| **P(JM)**   | **ON TIME** | **LATE** |
| SNMPTN      | 0.316     | 0.598    |
| SPMB        | 0.500     | 0.161    |
| SPMPRM      | 0.184     | 0.241    |
| **Total**   | 1.000     | 1.000    |

Calculation of numerical attributes such as student GPA uses the Gauss Density formula, namely:

\[
\hat{f} = N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.
\]  

**Information:**
- \( \hat{f} \): Opportunity
- \( X_i \): Attribute it to \( i \)
- \( x_i \): The value of the \( i \)th attribute
- \( \mu \): Mean
- \( \sigma \): Standard deviation

\[ (2) \]
The mean and standard deviation are then used in calculating the GPA probability value in influencing student graduation. The calculation of the P GPA applied to the testing data is shown in table 3 below.

There are a total of 48 testing data in this study and data cleaning has been applied first for noise removal. Recapitulation of the calculation of the probability of the CPI can be seen in the table at the bottom of this.

**Table 3. Recapitulation of P calculation results (GPA) of right and late graduation**

| NO. | ON TIME | LATE  | NO. | ON TIME | LATE  |
|-----|---------|-------|-----|---------|-------|
| 1   | 0.75308 | 0.50897 | 25  | 0.46670 | 0.00552 |
| 2   | 0.41806 | 0.87636 | 26  | 0.81369 | 0.43176 |
| 3   | 0.14209 | 0.97689 | 27  | 0.48564 | 0.00621 |
| 4   | 0.00973 | 0.38133 | 28  | 0.69098 | 0.58412 |
| 5   | 0.32088 | 0.95437 | 29  | 0.75954 | 0.02918 |
| 6   | 0.15805 | 0.98779 | 30  | 0.46161 | 0.83511 |
| 7   | 0.03127 | 0.64440 | 31  | 0.95677 | 0.15384 |
| 8   | 0.67153 | 0.60702 | 32  | 0.94775 | 0.21025 |
| 9   | 0.95280 | 0.19220 | 33  | 0.95294 | 0.19158 |
| 10  | 0.95266 | 0.19279 | 34  | 0.72528 | 0.02411 |
| 11  | 0.07190 | 0.85180 | 35  | 0.73978 | 0.02613 |
| 12  | 0.22652 | 0.99743 | 36  | 0.32869 | 0.94906 |
| 13  | 0.59468 | 0.69484 | 37  | 0.00017 | 0.03241 |
| 14  | 0.14661 | 0.98042 | 38  | 0.02406 | 0.57938 |
| 15  | 0.30143 | 0.96667 | 39  | 0.85718 | 0.05233 |
| 16  | 0.00881 | 0.36272 | 40  | 0.15778 | 0.98764 |
| 17  | 0.01075 | 0.40062 | 41  | 0.95572 | 0.17627 |
| 18  | 0.48734 | 0.80946 | 42  | 0.46688 | 0.82993 |
| 19  | 0.51515 | 0.78458 | 43  | 0.95479 | 0.13820 |
| 20  | 0.47175 | 0.82510 | 44  | 0.90946 | 0.29187 |
| 21  | 0.00055 | 0.07126 | 45  | 0.93715 | 0.23798 |
| 22  | 0.64882 | 0.63342 | 46  | 0.00178 | 0.14877 |
| 23  | 0.09133 | 0.90405 | 47  | 0.00013 | 0.02651 |
| 24  | 0.08612 | 0.89185 | 48  | 0.00001 | 0.00318 |

The probability value of each attribute is then multiplied according to the data to be predicted. Data that will be predicted next will be called data testing. The testing process will later involve testing data to predict whether graduation is correct or late. The prediction results table for testing data is presented in table 4 below.

**Table 4. Student graduation prediction results on data testing**

| No. | Nim  | Name                  | Length of Study |
|-----|------|-----------------------|-----------------|
| 1   | 160403001 | Dwinitha Aura        | LATE           |
| 2   | 160403003 | Jernita Nadeak       | ON TIME        |
| 3   | 160403005 | Abdul Rahiem Nasution | LATE         |
| 4   | 160403006 | Muhammad Imam Prakasa | LATE         |
| 5   | 160403008 | Bayu Suwandira       | LATE           |
| 6   | 160403010 | Tamadhar Al Khansa   | LATE           |
The results of the graduation prediction were then tested using the rapid miner 7.3 software which can be seen in the image below.

![Figure 3. Accuracy test results using rapiminer 7.3](image)

The results of the accuracy of the naïve Bayes model show the level of accuracy is 70.83%, meaning that the passing classification model is seen from the level of precision which reaches 70.83%, which is considered significant (moderate), but this needs to be reviewed from the perspective of complexity and the number of datasets.

4. Conclusion

Based on the research that has been done,

a. The training data used in this study were 125 records from students from 2013 to 2015. While the testing data used were 48 records came from 2016 student data.

b. The attributes used include the student identification number, student name, gender, area of Origin, school of Origin, a path of entry, and GPA.

c. From the results of testing 173 sample data using the Naïve Bayes algorithm, the pattern formed has a match accuracy of 70.83%, which means that the way is useful in predicting student
The model to predict student graduation is successfully created using the Naïve Bayes algorithm.

References
[1] Faisal M 2008 Malang Network Management Information System: UIN-Malang Press
[2] Sitepu M H et al. 2020 A Socio-Technical Approach to Assess Readiness of Organizations for Industry 4.0 J. Phys.: Conf. Ser. 1542 012031
[3] Kapiudin M 2007 Data Mining for Customer Classification with Ant Colony Optimization, Vol 8 No 1
[4] Regulation of the National Higher Education Accreditation Board Number 3 of 2019 concerning Higher Education Accreditation Instruments
[5] Ridwan M 2013 Application of Data Mining to Evaluate Student Academic Performance Using Algoritma Naïve Bayes Classifier
[6] Kurnia Saputra H 2018 Data Mining Analysis For Mapping Students Who Need Guidance And Counseling Using Algoritma Naïve Bayes Classifier
[7] Sinulingga S 2015 Introduction to Industrial Engineering 2nd Edition Medan : USU Press
[8] Ningrum 2017 The Influence of Using Problem Solving Based Methods on The Learning Outcomes of Class X Student Economics Class X Even Semester Metro Man 1 Academic Year
[9] Ashari Muin A 2016 The Naive Bayes Method for Predicting Graduation (Case Study: College New Student Data)
[10] Pudjiantoro et al 2011 Data Mining Application To Analyze Possible Resignation of Prospective New Students. http://yudiagusta.files.wordpress.com/2008/09/051-060-knsi2011-009-implement- data- mining- for- analyzing- possible resignation-prospective-new-student.pdf (downloaded on March 21, 2020)