JurnalMantik
Volume 4, Number 3, Nov 2020, pp. 2100-2105 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Data Mining Predicts The Graduation of Students of Stmik Atma Luhur Information System Using Neive Bayes Algorithm

Agus Dendi Rachmatsyah[1], Benny Wijaya[2], Kiswanto[3]

[1,3]Information System Study Program, STMIK Atma Luhur, Pangkalpinang, Kep.Babel, Indonesia,
[2]Informatics Engineering Study Program, STMIK Atma Luhur, Pangkalpinang, Kep.Babel, Indonesia,

E-mail: dendi@atmaluhur.co.id

ARTICLE INFO

Article history:
Received: 01/11/2020
Revised: 09/11/2020
Accepted: 26/11/2020

Keywords:
data mining, timeliness of graduation, neive bayes

ABSTRACT

In studying in higher education, the accuracy of graduation at the end of the semester is important, as is the case for the final semester students at STMIK Atma Luhur Pangkalpinang. STMIK Atma Luhur Pangkalpinang has three Study Programs, one of which is the Information Systems Study Program which receives less than 150 to 200 students each year and is expected to graduate on time. With the number of students who enter, it might not match those students who have completed their studies on time. In this study, the classification of the graduation punctuality was performed using the Naif Bayes algorithm. Neive Bayes predicts future opportunities based on past experience known as Bayes Theorem. The implementation in this study uses the RapidMiner 7.0.001 software. To help find accurate values, the attributes used in this study are NIM, Name, Level, Study Program, Province of Origin, Gender, SKS, GPA, and Year of Graduation. STMIK Atma Luhur.

1. Introduction

STMIK Atma Luhur Pangkalpinang is the first computer-based private college in Bangka Belitung Island Province. Atma Luhur High School of Computer Science Management has three Courses including: Information System Study Program, Informatics Engineering Study Program and Informatics Management Study Program.

In this study, the authors conducted research on one study program only, namely the Information System study program. The Information System Study Program is one of the most in-demand courses for students. The Information System study program has graduated approximately 958 students and only about 58% of the time of graduation is on time, from that data there are still many students who graduate not on time.

It is also very influential with the accreditation standards of institutions and study programs between the ratio of incoming students and those who graduate on time. Many factors that cause students to graduate are not on time, among others: Lack of attendance, Students lacking understanding of the courses they have taken and other factors.

The technique used in this study uses the Neive Bayes Data Mining Algorithm. The Neive Bayes algorithm is a classification method using probability and statistical methods put forward by British scientist Thomas Bayes. The Naive Bayes algorithm predicts future opportunities based on previous experience so that it is known as the Bayes Theorem.

Bayes theorem is a theorem with two different interpretations. In Bayes's interpretation, this theorem states how far the degree of subjective belief should change rationally when there are new clues.

With Naive Bayes Algorithm can also determine the probability of each parameter, so it can be known which parameters are the most influential cause of the student's graduation time.

2. Theory

2.1 Data Mining

Data Mining is a series of processes for extracting value-added information that has not been known manually from a database by extracting patterns from data with the aim of manipulating data into more valuable information obtained by extracting and recognizing important or interesting patterns from data contained in the database

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Data mining (Segall et.all,2008) commonly referred to as "Data or Knowledge discovery" or finding hidden patterns in data mining is the process of analyzing data from different perspectives and concluding it into useful information.

Data mining (Han and Kember, 2006:5) is defined as the process of extracting or saving the necessary knowledge from a large amount of data.

Ordinary data mining is also known other names such as: Knowledge discovery (mining) in databases (KDD), knowledge extraction data analysis /pattern and business intelligence and is an important tool to manipulate data for the presentation of information according to the needs of users with the aim to assist in the analysis of behavioral observation collection, in general the definition of data mining can be interpreted as follows:

a) Interesting pattern discovery process of large amounts of stored data.
b) Extraction of useful or interesting information (non-trivial, implicit, previously
c) it is not yet known the potential usefulness) of patterns or knowledge of large amounts of stored data.
d) Exploration of automated or semiautomatic analysis of large amounts of data to look for meaningful patterns and rules.

2.2 Rapid Miner

Rapid Miner is one of the software for data mining processing. The work done by RapidMiner text mining is to range from text analysis, extract patterns from large data sets and combine them with statistical methods, artificial intelligence, and databases. The purpose of this text analysis is to obtain the highest quality information from the processed text.

RapidMiner provides data mining and machine learning procedures, including: ETL (extraction, transformation, loading), data preprocessing, visualization, modelling and evaluation. The data mining process is composed of nestable operators, described with XML, and created with a GUI. The presentation is written in the Java programming language.

The process of analyzing documents using RapidMiner is done using the First Installed Text Processing Operator by selecting the Help menu and then Update RapidMiner. After that, the required operator is selected from several operators presented in the list, including the Text Processing operator. Once installed, the Text Processing Operator will appear in the Operators list, as will the Operator Feature Selection Extension which is also installed in the same way.

2.3 Neive Bayes Algorithm

Naive Bayes algorithm is one of the algorithms found in classification techniques. Naive Bayes is a classification of probability and statistical methods by British scientist Thomas Bayes, which predicts future opportunities based on previous experiences known as the Bayes Theorem. The theorem is combined with Naive where it is assumed that the conditions between attributes are mutually free. The Classification of Naive Bayes assumed that it existed or not. Certain characteristics of a class have nothing to do with the characteristics of the other class.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Fig 1. Bayes Formula.

Description:

x : Data with unknown class
H : The data hypothesis is a specific class
P(H|x) : Probability hypothesis based on condition(posteriori probability)
P(H) : Hypothetical probability (prior probability)
P(x|H): Probability based on conditions in hypothesis
P(x) : Probability c

3. Method

3.1 Research Development Model

For design purposes, a method is needed to reference the software development process. Therefore the design method used is waterfall model.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

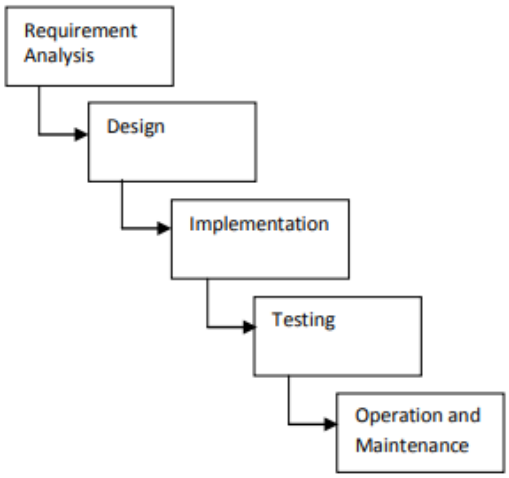

Fig 2. Waterfall Method Stage

a) Requirement Definition collects needs in full and then analyzes and defines the needs that must be met by the software to be built.
b) System and Software Design The search process needs to be intensified and focused on software. It aims to facilitate the understanding of the processes that occur, explaining the flow of the system in the software. In this process there are four attributes including data structure, software architecture, interface representation, and procedural algorithms.
c) Implementation and Unit Testing Program design is translated into codes using the specified programming language. Programs built directly are tested per unit.
d) Integration and System Testing This stage is an implementation of the design stage that will be technically done by the programmer. The unit unification of the program is then tested in its entirety (system testing).
e) Operation and Maintenance This stage is an error-free software shutdown stage, and the result should be completely in accordance with the needs that have been defined before, lastly done software maintenance. In this study using descriptive methods, where this research aims to solve phenomena or problems that exist at this time, in this study the case taken is the data of students of the Information System study program at STMIK Atma Luhur Pangkalpinang. Descriptive methods have the following features:
    1) It centers on solving problems in the present, and on actual problems.
    2) The collected data is first compiled, explained and analyzed because this method is often called an analytical method.

3.2 Research Development Methods

Research methodology is a step and procedure carried out in the collection of data or information to solve problems and test research hypotheses.

The method used in carrying out research is the basis of the preparation of research design and is the description of scientific methods in general. In this study, the system to be built is to classify the timeliness of graduation for information system students at STMIK Atma Luhur Pangkalpinang using Naive Bayes algorithm.

Related to the process of quantification of ordinary data is classified into 4 factors namely: Nominal Variable Factor, i.e. variable stipulated based on classification process; these variables are mutually exclusive between one category and the other; example: gender, marital status, occupational type

a) Ordinal Variable Factor, i.e. variables arranged based on the level in a particular attribute. The highest level is used to be numbered 1, the level below is given the number 2, then below it is given the number 3 and so on. (ranking)
b) Interval Variable Factor, which is the variable resulting from the measurement, which in that measurement is realized there is the same unit of measurement. Examples: variable intervals e.g. learning achievement, attitude to a program expressed in scores, earnings and so on.
c) Variable ratio factor, is a variable that in quantification has absolute zero.

3.3 Research Steps

a) The initial stage of research begins by determining the needs of research data such as naïve bayes algorithm theory, active student data, after which data is collected and prepared research tools and

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

materials.

b) Literature study is done by studying and understanding the theories used, namely looking for factors that become naïve bayes algorithm theory, calculation of probability scores in each class, and calculation of student scores. The data is searched by collecting literature, journals, internet browsing and readings related to topics in the form of textbooks or papers.

c) Observation is done by conducting a live interview to the relevant part of the problem taken to obtain accurate data, as well as studying naïve bayes algorithm.

d) Results from literature studies and observations found naïve bayes theory data.

e) Software Engineering: Naïve Bayes After that from the data collected first will be created naïve bayes algorithm design. In the Naïve Bayes algorithm it takes data to be calculated, in order to get a probability result.

f) Software Engineering: Dempster Shafer

The probability result in naïve bayes algorithm becomes a measure of the level of trust in the graduation result so that it will produce a certainty value whether the mahasiwa passed according to the specified time.

3.4 Research Sites

In the study titled "Data Mining Predicts The Graduation of Information System Students Using Neive Bayes Algorithm" was conducted at STMIK Atma Luhur Pangkalpinang.

3.5 Data Collection Techniques

The data collection techniques used are in the following ways:

a) Library Study Instruments

Instruments for data collection by library study method. This library study instrument is a researcher studying literature on the concept and workings of systems that will be created according to case studies in books and scientific articles with themes corresponding to case studies.

b) Observation Instruments

Data collection techniques by making direct observations in case studies to look for existing problems as well as solutions to solve them.

3.6 Data Analysis Techniques

The technique of specification of system needs means doing details about what is needed in the development of the system and making planning related to the system project. Analysis of functional needs and non-functional needs of the system is also required. Functional needs relate to software features to be built, system integration, ordering, and transactions. Non-functional needs are not directly related to a feature in the software, such as system performance.

In this study the system analysis technique used is to look for information about the background in case studies. This analysis process is useful to provide an alternative form proposed as a single problem solving technique.

4. Results And Discussions

4.1 Application Main Module

From the results of the study using acgoritma classification of data mining namely Naive Bayes based forward selection for the prediction of the graduation of students Precisely or Not On Time by using 144 student data of STMIK Atma Luhur Pangkalpinang Information System Department, then it can be analyzed from the results that the addition of Forward Selection feature in prediction can improve accuracy and enter into the excellent clasification section with an accuracy rate of 97.92% with a precision level of 98.92% compared to predictions using only Naive Bayes algorithm which is 97.68% For research advice it can then add another data mining classification algorithm in addition to naive bayes main algorithm, and it is expected to perform attribute selection and because data set validation is so important that it produces accuracy that is balanced with the method used.

4.2 Analysis and design

After carrying out the initial stage and the initial preparation stage, then carry out the analysis and design process using Repeat Miner using the Neive Bayes Algorithm method.

Based on the research conducted, a new pattern, information, and knowledge has been generated in the data mining process for the classification of student graduation based on student data of STMIK Atma Luhur Information System Department. From the research produced a new pattern, information, and knowledge in accordance with the purpose of mining data, namely the calculation pattern of mining data containing training data and testing data and looking for probability of each attribute based on training data and testing data to produce a new information, whether in the data of students of stmik atma sublime information system more students graduate on time or students graduate not on time. Then to test the level of accuracy,

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Rapidminer is used as a tool in the process of testing the accuracy of the classification. From the calculation process of data mining using naïve bayes algorithm and the level of accounting, a new information is generated that is the calculation of mining data based on students of STMIK Atma Luhur Information System Department, showing students graduate "yes"/on time with total multiplication prior probability worth 0, while students pass "no" / not on time with total multiplication prior probability worth 0.00055. For accuracy based on the classification process using the naïve bayes algorithm, through all stages it is ensured that no important parts are missed, resulting in an accuracy rate of 97.92%. Based on the results of data mining calculation and accuracy level testing process using Rapidminer, it can be drawn the conclusion that students pass "no" / not on time greater than the year class of graduating "yes" / on time. While the analysis conducted on the accuracy level using naïve bayes algorithm shows that the value produced by the naïve bayes algorithm has a fairly high level of strength. This is evidenced by the calculation result reaching a value of 97.92 %, a value of 97. 92% prove that the built model can be used to classify the graduation of mahsiswa. A value of 97.92 % can also be caused by a lack of complex data that results in the model being able to predict accurately.

4.3 Prototype Design and Manufacturing Stage

The method used in carrying out research is the basis of the preparation of research design and is the description of scientific methods in general. In this study, the system to be built is to classify the timeliness of graduation for information system students at STMIK Atma Luhur Pangkalpinang using Neive Bayes Algorithm.

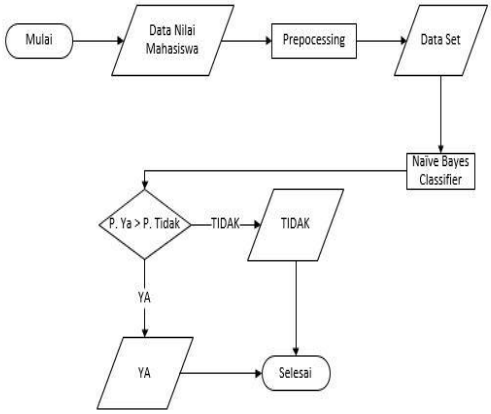

Fig 3. Flowchart of the system flow method

4.4 Algorithm Analysis Stage

For this stage of the process will be described in figure 4.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

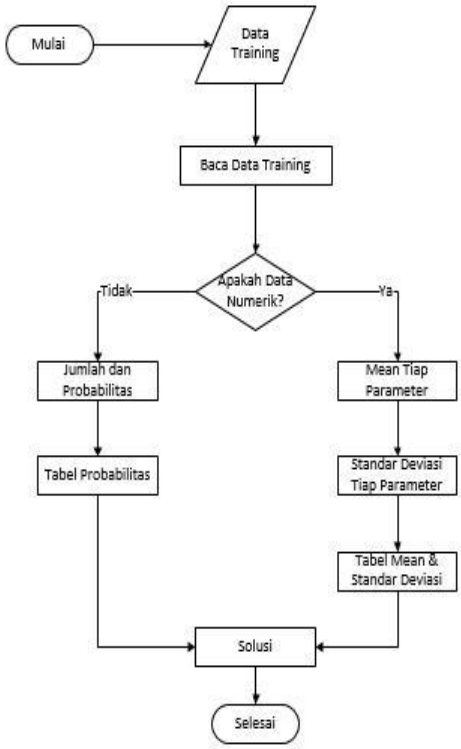

Fig 4. Flowchart Process Algorithm Naïve BayesClassifier

Figure 4 shows the flow on the Naive Bayes Classifier algorithm process. Starting by reading the training data, then the system checks the attributes on the data, numeric-type attributes or not. If the attribute is not numeric, it is calculated in number and probability. If the attribute is numeric, it is calculated mean and the standard deviation, then calculated the probability.

5. Conclusion

Based on the results of the discussion, this study has the following conclusions:

a) Based on data mining calculations using naïve bayes algorithm, it can be concluded that the year class pass "no" / do not pass on time is greater than the year class pass "yes" / pass on time.

b) The Naive Bayes Classifier algorithm can be implemented for the classification of student pass timeliness and produces a pretty good classification percentage, more than 80% of the test scenarios performed.

6. References

[1] A. Adetokunbo and A. Basirat, "Software Engineering Methodologies: A Review of The Waterfall Model and Object-Oriented Approach," *Int. J. Sci. Eng. Res.*, vol. 4, no. 7, pp. 427–434, 2014.

[2] N. Amalia, Shaufiah, and S. Siti Sa'adah, "Penerapan Teknik Data Mining Untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma Naive Bayes Classifier," *J. Telkom Univ.*, vol. 1, no. 2, pp. 1–11, 2014.

[3] Bustami, "Penerapan Algoritma Naive Bayes," *J. Inform.*, vol. 8, no. 1, pp. 884–898, 2014.

[4] A. P. Fadillah and B. Hardiyana, "Penerapan Naïve Bayes Classifier Untuk Pemilihan Konsentrasi Mata Kuliah," *J. Teknol. dan Inf.*, vol. 8, no. 2, 2018.

[5] S. Kom, E. Dewi, S. Mulyani, S. Kom, and I. R. Nurhasanah, "Penerapan Data Mining Classification Untuk Prediksi Perilaku Pola Pembelian Terhadap Waktu Transaksi Menggunakan Metode Naïve Bayes," *Konf. Nas. Sist. dan Inform.*, pp. 9–10, 2015.

[6] F. H. Messerli and R. B. Devereux, "Introduction: Left ventricular hypertrophy-Good or evil?," *Am. J. Med.*, vol. 75, no. 3 PART A, pp. 1–3, 1983.