Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm

Ida Bagus Adisimakrisna Peling[1], I Nyoman Arnawan[2], I Putu Arich Arthawan[3], and IGN Janardana[4]

[1][2][3] Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[4] Department of Electrical and Computer Engineering, Faculty of Engineering, Udayana University
Email: adi.peling@gmail.com

***Abstract*—The quality of universities, especially study programs in Indonesia is measured based on accreditation conducted by BAN PT. According to BAN PT the quality is measured based on 7 main standards, one of them is Student and Graduate. One of the problems that still be the subject of discussion related to student failure is about the students who graduated not on time. Students graduating not on time are students who can not complete their studies in accordance with the provisions of time given. The existence of a graduate student is not timely of course cause problems and potentially drop out that affect the quality of education and accreditation. A system that predicts students' graduation is required by evaluating their learning outcomes. The timeliness of graduating students can be done with data mining techniques to find graduation patterns of students who have graduated which then used as a basis to predict students' graduation in the next year. This study showed that Naïve Bayes was able to classify the correct data testing on average by 86.16% and 13.84% error. In addition, other information obtained from the data testing used that the students who entered from the PMDK Pass graduated on time as much as 40%, other paths graduated on time by 26.7%, and pass filter exam on time 13.3%.**

Index Term—Data Mining, Naive Bayes, Prediction.

I. Introduction

The quality of the college, especially study programs in Indonesia are measured based on the accreditation conducted by the Badan Akreditasi Nasional Perguruan Tinggi or BAN PT. According BAN PT [1] the quality is measured based on the 7 main standards, one of them is a students and graduates.

To achieve the highest quality level quality of college system is to explore the knowledge from the education data as the main learning attributes that affect the achievement of the students.

The student is one of the important aspects in the success of the implementation program study at a university. The reflection of the quality of a college can be seen from the high low levels of students success or failure.

The success or failure of a student to complete the study in time can be made of evaluation to continuously improve and enhance the quality of the college in terms of both management for the quality of education and accreditation. Each college seeks to continue to continuously improve management to improve the quality of education and improve the accreditation. One element of college accreditation is graduating on time [2].

One of the problems that still a matter of discussion related to the failure of the graduate students is about not timely. Students do not pass on time is not able to complete his studies in accordance with the provisions of the given time. The existence of the student graduate not the right time is certainly cause problems and potentially drop out affecting the quality of education and accreditation.

Based on the above, it needs a system that can predict the graduation of students through learning how to evaluate results. Timeliness graduate students can do with data mining techniques to discover patterns that have passed the graduation which is then used as the basis for predicting graduation in the next year.

Data mining is the process of adding additional value of a set of data in the form of knowledge that had been unknown to manually [3]. Data mining technique is a process of extracting information to gain knowledge (knowledge discovery) and found the pattern (pattern recognition) on a pile of data in the database are usually large-scale [4]. Data Mining itself has several techniques exist one classification. Classification is the process of finding a model (or functions) that describe and distinguish classes of data or concept that aims to make the bias is used to predict the class of the object class label is unknown [5].

Data mining techniques to be used in this study is a classification algorithm Naive Bayes classifier which is a simple probability that applying Bayes' Theorem. The basic idea of Bayes' Theorem is dealing with hypothetical namely designing a classification function to separate objects [6].

This research will provide predictions by applying data mining using Naive Bayes classification method in order to analyze the possibility of graduating students over eight semesters with the classification of the data set of students who have graduated.

II. Related Research

A. *Implementation of Data Mining To Predict Period of Students Study Using C4.5 Algorithm (Case Study: University Dehasen Bengkulu)*

The purpose of this study is to use the C4.5 decision tree algorithm-based and implemented into an application that RapidMiner is expected to improve the accuracy of the analysis of the study period the student. This research was conducted at the Dehasen University of Bengkulu. In the study discussed by monitoring the results of studies at the university in the form of the GPA and the number of credits that have not been accurate for determining a student to graduate on time or not. In this study was to classify the grading of student used data mining techniques with C4.5 algorithms and implemented into Rapid Miner, it aims to see the results of the development can graduate on time or not. From the research results prove that the algorithm C4.5 is more accurate than analysis conducted by analysts students. This is evidenced by the results of the evaluation study found C4.5 algorithms capable of analyzing the punctuality of students completing their study [7].

B. *Naive Bayes Method To Predict Graduation (Case Study: New Student Data Universities)*

Timetables system, genetic algorithm, timetable at universityClassifier created from a set of data. Bayesian classifier is a statistical classifier for predicting the probability of a particular class membership. This research will try to perform data classification for prediction of new student graduation, Naive Bayes algorithm method used for nave Bayes classification performance high enough ability to predict future opportunities based on experience or data in the past. Imlementation WEKA algorithms in applications that will explore the characteristics of the dataset with superficial attributes Pass options. Evaluation results show classified data correctly (correct classified instances) in accordance with the grouping of choice graduated first choice, second choice and do not pass by the algorithm as much as 93.6288% or as much as 338 data and classified data, but does not match the class predicted (incorrect classified instances) which should be a group of two or Pass options but are included in the group First choice as many as 6.3712% or as much as 23 data. Value Percentage accuracy demonstrated the effectiveness dataset Admissions applied to the methods Nave Bayes Classification, which reached 94% [8].

C. *Application of Data Mining Techniques for Classification Timeliness of Information Engineering University Graduate Student Telkom Using Nave Bayes algorithm Classifer*

In this study, researchers used Nave Bayes methods for determining the timing of graduate student. For the measurement accuracy of the classification confusion matrix

used by the two classes. This research scheme uses 90% of the data for traning and 10% for data testing. After testing found the number of students who graduate on time is less than the graduate students who are not on time. Unknown probability of timeliness pass on the number of credits is not too large. However attribute GPA greatly affect the probability of timely completion of studies. Probability obtained using GPA attribute is 85%. Nave Bayes algorithm performasi Classifer showed pretty good results of the testing that has been done. From the testing that was done generate Correctly classified by 86% [6].

D. *Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing*

Many companies like credit card, insurance, bank, retail industry require direct marketing. Data mining can help those institutes to set marketing goal. Data mining techniques have good prospects in their target audiences and improve the likelihood of response. In this work we have investigated two data mining techniques: the Naïve Bayes and the C4.5 decision tree algorithms. The goal of this work is to predict whether a client will subscribe a term deposit. We also made comparative study of performance of those two algorithms. Publicly available UCI data is used to train and test the performance of the algorithms. Besides, we extract actionable knowledge from decision tree that focuses to take interesting and important decision in business area [9].

III. Literature Review

A. *Data Mining*

Data Mining is the process of extracting information from a very large set of data through the use of algorithms and withdrawal techniques in the field of statistics, machine learning and database management systems [10]

Data Mining (DM) is the process of analyzing data from various perspectives and summarize it into useful information, where the information can be used to increase revenue, cut expenses or both [11]. Evolution of Data Mining begins when the first data collecting times especially for business and bioinformatics applications that are stored in the computer and proceed with the increase in data access technology.

Stages of Data Mining [12]:

1. Cleaning the data (to remove the data are inconsistent and noise)
2. Data integration (merging data from multiple sources)
3. Transformation of data (the data is converted into a form suitable for the mining)
4. Application of Data Mining techniques
5. Evaluate patterns found (to find interesting / valuable)
6. Presentation of knowledge (with a visualization technique)

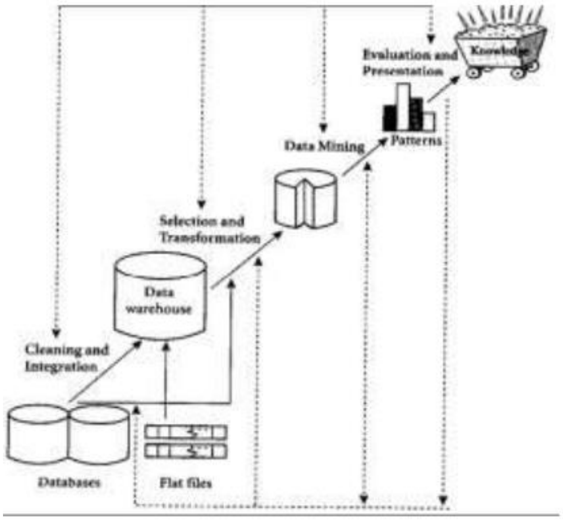

Fig. 1. Stages of Data Mining

B. *Naive Bayes*

Naive Bayes is a simple probabilistic classifier that calculate a set of probabilities by summing the frequency and value combination of given dataset. The algorithm uses Bayes's theorem and assumes all attributes independent or non-interdependent given by the value of the class variable[13]. Another definition says Naive Bayes is a classifier with a probability method and statistics brought by British scientist Thomas Bayes, predicting opportunities in the future based on previous experiences[14].

Naïve Bayes algorithm is a probabilistic method used to classify the class of a data. Outline Naïve bayes method is a statistical analysis where the initial probability (prior probability) being estimated from data traning[15].

For each parameter probability is determined based on initial probability. Mathematically this algorithm can be written as in equation 1. The choice of this method is relatively easy to use because there is no matrix multiplication or numerical optimization, more efficient when used to predict in very large quantities, and has a relatively high degree of accuracy in the prediction results.

$$P(H|E) = \frac{P(E|H) \; X \; P(H)}{P(E)} \qquad (1)$$

That is :
P(H|E) = Positional probability of probability (conditional probability)
P(E|H) = Probability of parameter E on hypothesis H
P(H) = The prior probability (prior) hypothesis H
P(E) = Initial probability (prior) parameter E
Advantages of Bayesian Naive:

1. Handling quantitative and discrete data
2. It only requires a small amount of training data to estimate the parameters (average and variance of variables) required for classification
3. Handle the lost value by ignoring the agency during the estimated opportunity calculation
4. Fast and space efficiency
5. Strong against irrelevant attributes

Disadvantages of Naive Bayesian :

1. Not applicable if the conditional probability is zero, if zero then the predicted probability will be zero as well
2. Assume independent variables

IV. Discussion

In this research designed a system by applying data mining techniques to find an information that has a value that is able to predict graduation students right or not on time from the student data using the naive bayes classification method. Flow chart of this system can be seen in picture 2.

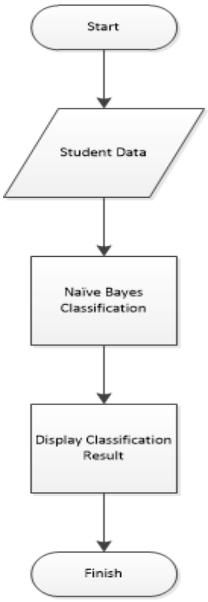

Fig. 2. Flowcart System

The steps to complete this research are:

1. Data collection
   All the data obtained from the Academic Section Veterinary Unud, the data sampled is data from the 2012 force that has passed. Attributes used are NIM, Name, Entrance Path, Gender, SKS Semester 1 to 4, number of credits Semester 1 to 4, duration of study. Where for the length of study is divided into two categories, namely timely and not on time, generally undergraduate students take an average of 8 semesters.
2. Data Processing

- Creating tables from sample data and performing data cleansing discards inconsistent data and noise.

- Data integration combines data from multiple sources, created tables according to the attributes used as described above and specify the reference attribute of the table ie pass on time or pass not timely.
- Transforming data by converting data into a form suitable for mining. The data transformation done here is to change the form of data from Semester IP attribute and number of credits.
- Analyzing data tables (obtained from Steps by Naive Bayes method)
- Evaluate the pattern found in the analysis of the data to find the information worth.

Qualification results are then analyzed to measure the performance of this system by calculating accuracy and error. To calculate the accuracy can be seen in equation 2.

$$Accuracy = \frac{totaldatavalid}{totaldatatesting} X\ 100\% \qquad (2)$$

And to measure the percentage of error from nave bayes algorithm can be seen in equation 3.

$$Error = \frac{totaldatainvalid}{totaldatatesting} X\ 100\% \qquad (3)$$

For test data taken from the students with the number of 80 people, with data for knowledge nave bayes 65 people (81,25%) and test data as much as 15 people (18,75%). To test results from nave bayes algorithm can be seen in the table 1.

TABLE I.
NAIVE BAYES TESTING

NIM	Original Value	Naive Bayes Value	Conclusion
1209005071	Not on time	Not on time	Valid
1209005072	On time	On time	Valid
1209005073	On time	On time	Valid
1209005074	Not on time	On time	Invalid
1209005075	On time	On time	Valid
1209005076	On time	On time	Valid
1209005076	On time	On time	Valid
1209005078	On time	On time	Valid
1209005079	On time	On time	Valid
1209005080	On time	On time	Valid
1209005081	On time	Not on time	Invalid
1209005082	On time	On time	Valid
1209005083	Tepat	Tidak Tepat	Invalid
1209005084	Tepat	Tepat	Valid
1209005085	Tepat	Tepat	Valid

From the results of the calculation of accuracy and error that has been done is obtained from the algorithm naive bayes accuracy of 80% and error of 20%. There were also tests of 80 student data with data testing of 10%, 20% and 30% of the total data to determine the accuracy of Naive Bayes. Where the accuracy of naive bayes algorithm can be seen in figure 3.

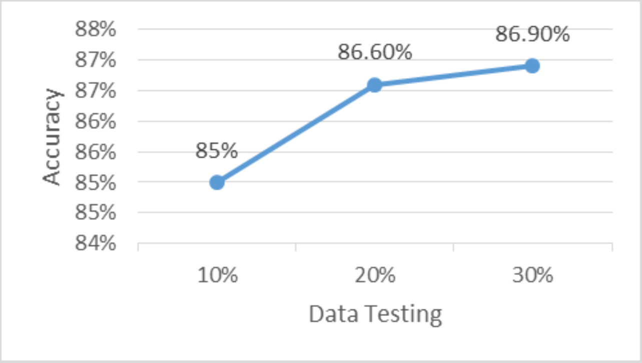

Fig. 3. Accuracy of Naive Bayes

From the prediction results obtained also other information that students who come from the Path PMDK graduated in time as much as 40%, other band pass time of 26,7%, and the test filter graduated on time of 13.3%. Can be analyzed the results that students who go through the path PMDK has a tendency to pass more quickly than students through the other path and the path of the filter test.

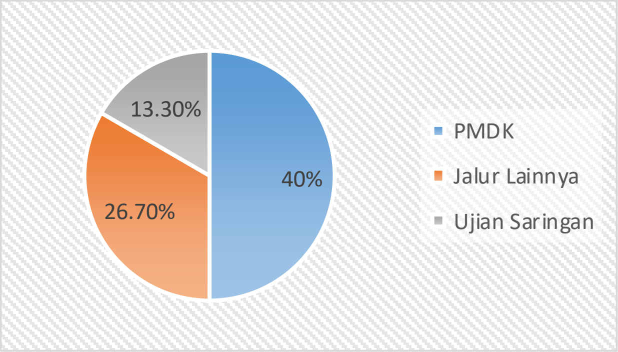

Fig. 4. Graph of Graduate On Time Based On Entrance Path

V. CONCLUSION

From the research that has been done can be concluded that the Nave Bayes classification algorithm has a good performance in determining the prediction period of study. This study shows that Nave Bayes is able to correctly classify data testing. Determination of training data and data testing can affect the test results, because the training data pattern will be used as a rule to determine the class on the data testing. And can be concluded from the results of testing algorithm Naive Bayes has an average accuracy of 86.16% and error of 13,84%. With the number of samples of 15 students used the test data using naive bayes method obtained the result that students who will graduate on time amounted to 12 students or about 80% of the sample while the student who will graduate is not on time amounted to 3 students or about 20%.

From the analysis based on the data can be used by university institutions to further increase the quota of new admissions through the PMDK Line to increase the average time of graduation of students, so that the quality of accreditation assessment for student and graduate point becomes better.

REFERENCES

[1] BAN - PT, B. (2011). Akreditasi Intitusi Perguruan Tinggi - Buku III Pedoman Penyusunan Borang, pp 4.

[2] Untari, D. (2014). Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5.

[3] Banjarsari, M. A., Budiman, H. I., & Farmadi, A. (2015). Penerapan K-Optimal Pada Algoritma KNN Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer FMIPA UNLAM Berdasarkan IP Sampai Dengan Semester 4. Kumpulan Jurnal Ilmu Komputer(KLIK). Vol 02 No 02 September 2015.

[4] Larose, D. T., Discovering Knowledge In Data. New Jersey. John Wiley & Sons, Inc .(2005).

[5] Han, J., Kamber, M., & Pei, J. Data Mining Concepts and Techniques. Waltham. Morgan Kaufmann Publishers. (2012).

[6] Amalia, N., Shaufiah, & Saadah, S. (2016). Penerapan Teknik Data Mining Untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma Nave Bayes Classifier.

[7] Siska H., Aji S, Eko S. (2015). Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu). Jurnal Media Infotama Vol. 11 No. 2, September 2015

[8] Syarli, Asrul A. (2016) Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi). Jurnal Ilmiah Ilmu Komputer, Vol. 2, No. 1, April 2016

[9] Masud K., Rashedur M.R. (2013). Decision Tree and Naive Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. Journal of Software Engineering and Applications, April 2013

[10] Taruna R., S., Hiranwal, S., 2013, Enhanced Naive Bayes Algorithm for Intrusion Detection in Data Mining, *International Journal of Computer Science and Information Technologies*, Vol.6, No. 4, Hal 960-962.

[11] Goele Sangeeta, Chanana Nisha, Data Mining Trend in Past, Current and Future, 2012, International Journal of Computing & Business Research.

[12] Lindawati, Data Mining dengan Teknik Clustering Dalam Pengklasifikasian Data Mahasiswa Studi Kasus Prediksi Lama Studi Mahasiswa Universitas Bina Nusantara, 2008, Seminar Nasional Informatika 2008 (semnasIF 2008),

[13] Patil, T. R., Sherekar, M. S., 2013, Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *International Journal of Computer Science and Applications*, Vol. 6, No. 2, Hal 256-261.

[14] Bustami., 2013, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, *TECHSI : Jurnal Penelitian Teknik Informatika*, Vol. 3, No.2, Hal. 127-146.

[15] Jantawan, B., Tsai, C. (2014). A Classification Model on Graduate Employability Using Bayesian Approaches: A Comparasion. International Journal of Innovative Research in Computer and Communication Engineering. Vol 2 Issue 6 Juni 2014.