Prediction Graduate Student Use Naive Bayes Classifier

Allsela Meiriza\(^1\), Endang Lestari\(^2\), Pacu Putra\(^3\), Ayu Monaputri\(^4\), and Dini Ayu Lestari\(^5\)

\(^1\)allsela@ilkom.unsri.ac.id, Faculty of Computer Science Sriwijaya University, Indonesia
\(^2\)endanglestari@unsri.ac.id, Faculty of Computer Science Sriwijaya University, Indonesia
\(^3\)pacuputra@ilkom.unsri.ac.id, Faculty of Computer Science Sriwijaya University, Indonesia
\(^4\)09031181520031@students.ilkom.unsri.ac.id, Faculty of Computer Science Sriwijaya University, Indonesia
\(^5\)09031181520005@students.ilkom.unsri.ac.id, Faculty of Computer Science Sriwijaya University, Indonesia

Abstract

Student graduation is important in the accreditation assessment process. Because student graduation there are standards that must be achieved by the Study Program items, namely a four-year study period and a 3.0 GPA. Therefore we need a prediction that can anticipate from the beginning of the graduation standard level that has been set. This study aims to predict student graduation using Naïve Bayes Classifier with a data mining approach. Naïve Bayes provides accurate prediction results with a minimum error rated compared to all other data mining components. With the prediction of the student, graduation can be used as input, especially the Information System Study Program in making policies for improvement in the future. The software used in data processing is WEKA. The test results showed that from the Information Systems Study Program Faculty of Computer Science University of Sriwijaya in 2015 as many as 141 students as training data and in 2016 as many as 127 students as testing data, the prediction accuracy was 97.6378%.

Keywords: predicted, data mining, Naïve Bayes classifier

Introduction

Graduation is one of the supporters of the items in the accreditation assessment process. Because student graduation there are standards that must be achieved by the Study Program items, namely a four-year study period and a 3.0 GPA. Therefore we need a prediction that can anticipate from the beginning of the graduation standard level that has been set. This study aims to predict student graduation using Naïve Bayes Classifier with a data mining approach. Naïve Bayes provides accurate prediction results with a minimum error rated compared to all other data mining components. With the prediction of the student, graduation can be used as input, especially the Information System Study Program in making policies for improvement in the future. The software used in data processing is WEKA. The test results showed that from the Information Systems Study Program Faculty of Computer Science University of Sriwijaya in 2015 as many as 141 students as training data and in 2016 as many as 127 students as testing data, the prediction accuracy was 97.6378%.

While the method used to predict graduation uses the classification method. One classification method is the Naïve Bayes classification method. The reason for choosing the Naïve Bayes classifier method is because Naïve Bayes is reported as the best text classifier [3], and have high accuracy and speed [4]. Naïve Bayes is one method that can classify probabilities easily [5]. Also, according to previous research [3], Naïve Bayes accuracy results are the best among others classifiers with values correctly classified instances 97%.

Therefore, it is very necessary to predict student graduation to support policymaking and improve the quality of study program accreditation.

Literature References

State of the art

In this study, researchers explore information about theories related to the title used to get a foundation of scientific theory. As for some of the basis of scientific theory, namely:

Research conducted by Shadab Adam Pattekari and Asma Parveen with the title "Prediction System For Heart
Disease Using Naive Bayes”. The main objective of this research is to develop an Intelligent System using the data mining modelling technique, namely, Naive Bayes. It is implemented as a web-based application in this user answers the predefined questions. The equation of this research with research by researchers is equally aiming to predict something using the Naive Bayes method. While the difference is this research is to predict heart disease, while research by researchers is to predict student graduation.

Research conducted by Trilok Chand Sharma and Manoj under the heading "WEKA Approach for Comparative Study of Classification Algorithm". This paper explores the techniques of data mining to process a dataset and determine the importance of information from the classification study. Our work shows the method of WEKA analysis of file converts and selection of mining attributes and comparison with information Extraction of Evolutionary Learning not only, the classifications of data mining, but also the best efficient tool in learning is the biological, evolutionary algorithms. In this paper, the researcher obtains more data on WEKA framework as a method used by researchers to help predict student graduation.

Data mining

Data mining is the process of finding a pattern or interesting information from large amounts of data stored in the database, data warehouse or can be stored in a storage area other information using pattern recognition techniques such as statistical and mathematical techniques [6]. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. Data mining sources include database, data warehouse, the web, other information repositories[7].

PREDICTION

Prediction is the process of estimating systematically about something that is most likely to occur in the future based on past and present information. The predictive model is methods that produce predictions, regardless of its underlying approach: Bayesian or frequentist, parametric or nonparametric, a data mining algorithm or statistical model, etc.[8].

Naïve Bayes Classifier

A Naïve Bayes Classifier is a term dealing with a simple probabilistic classification based on applying Bayes’ theorem. In simple terms, a Naive Bayes Classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [9]. Naïve Bayes is a statistical classification technique that can be used to design a possibility in the future of a class. The Naïve Bayes theorem comes from the classification technique of the same decision tree and neural network. Naïve Bayes is proven to have high accuracy and speed when used in large databases with data.

A relatively simple algorithm to understand and design. A relatively fast algorithm produce class predictions, compared to all other classification algorithms. It can be easily trained using a small dataset.

WEKA

WEKA has been developed at Waikato University in New Zealand; the name is Waikato Environment for Knowledge Analysis The program is written in Java and distributed under the terms of the GNU General Public License [11]. Many classification methods have been developed with the help learning algorithms such as Bayesian, DecisionTree, K-NN (K-Nearest Neighbour), Support Vector Machine(SVM) and boosting, many classification methods have been developed. All these classifiers are are forms of training and set of rules are implemented. Bayesian classifier originates from the Theory of Bayesian Decision[12]. In this study, the software used in data processing that is using tools WEKA. WEKA is a software that implements various machine learning algorithms to perform several processes related to information retrieval or data mining. Features found on WEKA like Classification, Regression, Clustering, Association Rules, Visualization, and Data Preprocessing.

Cross-Industry Standard Process (CRISP-DM)

In this study, researchers used the method Cross Industry Standard Process - Data Mining (CRISP-DM). For standard data mining process as well as research methods. CRISP-DM is a standard methodology for data mining. CRISP-DM methodology is the most referenced and used in the practice of the data mining methodology [13]. The steps in this method include:

CRISP-DM has 6 steps that are first business understanding, then data understanding, data preparation, modelling, evaluation and finally deployment.

RESULTS AND DISCUSSION

In this study using CRISP-DM methodology to identify in the application of data mining using the Naive Bayes Classifier method, as follow:
Business Understanding Phase

This phase is the first phase that will be carried out for research. The business understanding phase can also be referred to as the Business Understanding phase. This study aims to classify the data of Sriwijaya University Information Systems students based on the Naïve Bayes algorithm using WEKA software. So that information will be obtained about the prediction of student graduation.

Data Understanding phase

In this phase, there is a process of the understanding of the data that has been obtained. The data is the data of students Department of Information Systems Faculty of Computer Science, University of Sriwijaya force in 2015 as a data sample and force in 2016 as the data to be predicted. The data has several attributes as follows:

Table 1. Table Attributes of Data Row

| Attribute       | Information                                       |
|-----------------|---------------------------------------------------|
| Gender          | Student sex                                      |
| Origin High School | Originally derived from the high school students featured / no |
| Entrance        | Driveway when the students admitted to the University of Sriwijaya |
| Organization    | Students once-being or does not belong to / several organizations during the lecture |
| Scholarship     | Students never-was or was not receiving a scholarship during the lecture |
| GPA             | GPA student                                      |

In table 1, it can be seen that in this study, 6 attributes will be used to predict student graduation, namely gender, origin high school, entrance, organization, scholarship and GPA.

Data Preparation Phase

At this stage, there are usually three steps are performed on the data, so that the data obtained can be processed to the research process. The steps are as follows:

1. Data Cleansing

Data have been obtained can not be directly used for research. Such data must pass through a preprocessing stage first, to eliminate the inconsistent data, duplicate the data, or data that is not associated with the research. Usually on data obtained still contains a lot of data are missing or empty data. Steps that can be taken on an empty data is to fill the empty value (imputations) with a mean, median, min/max, and others. Handlers were performed on data containing noise by identifying the values that should not be in the data with simple statistical techniques (such as mean and standard deviation) and find and eliminate the wrong data.

2. Select Cases or Variables

In this stage, the selection of cases to be the focus of research.

3. Transformation of Data

In conducting data mining requires an appropriate data format before being able to process data processing. At this stage of transformation, the names of the variables used and the aggregation of variables are carried out.

Phase Modelling

After the data preparation phase then obtained attributes used to predict a student's graduation. The attribute is gender, origin high school, entrance, organizations, scholarship and GPA. Then the data is calculated using the Naïve Bayes method. To be counted, the attribute used will be transformed first.

1. Gender:
   - Man = 0
   - Women = 1
2. Origin High School:
   - Not Featured = 0
   - Seed = 1
3. GPA:
   - 1.00 – 1.99 = 0
   - 2.00 – 2.99 = 1
   - 3.00 – 4.00 = 2
4. Entrance:
   - SNMPTN = 0
   - SBMPTN = 1
   - USM = 2
5. Scholarships:
   - Never = 0
   - Ever = 1
6. Organizations:
   - Never = 0
   - Ever = 1

Naïve Bayes method of calculation by using WEKA software, training data used were 114 and testing the data used was 127 data. Here are the results of data processing Training and Testing Data on WEKA:

![Figure 2. Figure Data Processing Training on WEKA](image)

The figure 2, is the result or output from WEKA which is the result of training data (2015 students) which can be seen that incorrectly classified instances the percentage is 64.9123% while for incorrectly classified instances the percentage is 35.0877%.

![Figure 3. Figure Testing Data Processing on WEKA](image)

Based on Figure 3, the process of calculating the average of agreed presentations using equation (1) and the error rate in the confusion matrix test will then be performed using equation (2) as follows:

**Accuracy** = \( \frac{\text{the number of correct predictions}}{\text{total number of predictions}} \times 100\% \)

\[
\text{Accuracy} = \frac{27 + 97}{27 + 3 + 0 + 97} \times 100\% = 97.6378\% 
\]

\[\text{(1)}\]

**Error Rate** = \( \frac{\text{the number of incorrect predictions}}{\text{total number of predictions}} \times 100\% \)

\[
\text{Error Rate} = \frac{3 + 0}{27 + 3 + 0 + 97} \times 100\% = 2.3622\% 
\]

\[\text{(2)}\]

The accuracy (correctly classified instances) and error rate (incorrectly classified instances) of data testing using naïve Bayes has an accuracy value of 97.6378% and an error rate of 2.3622%.

![Figure 4. Figure Predicted Result Graduation on WEKA](image)

Figure 4, is a display or the results of the prediction of testing data (2016 students) that have been produced by WEKA. From the results of predictions using WEKA, it is known that as many as 30 information system students will graduate on time while for 97 other students will be late to graduate.

Then to find out the gap between the 2015 Unsri SI class students and the 2016 Unsri class students who graduate on time, the following calculations will be carried out:

\[
\text{Graduated On-Time} = \frac{\text{SI 2015 graduated on time} - \text{SI 2016 graduated on time}}{2} = \frac{60 - 30}{2} = 15
\]

\[\text{(3)}\]
Late to Graduate = \[ \frac{SI \ 2015 \text{ late graduated} - SI \ 2016 \text{ late graduated}}{2} \]

\[= \frac{54 - 97}{2} = -21.5 \quad (4)\]

The 2015 Information System student data is data training and totals 114 data. Based on these data it is known that as many as 60 students graduated on time and 54 others were late graduates. For 2016 student information system data is data testing and totaling 127 data. From the results of WEKA processing, it is known that as many as 30 students are predicted to graduate on time while 97 others are predicted to be late graduating.

Evaluation Phase

Results of prediction graduation Force 2016 Department of Information Systems Faculty of the Computer Science University of Sriwijaya using Naïve Bayes method performed by calculation, using precision and recall. This evaluation compared the results with the prediction of actual results.

The data used were 114 training and testing the data used was 127 data. After testing a prediction of the data, the results are as follows:

Table 2. Table Evaluation

| Predictions | True | False |
|-------------|------|-------|
| Actual Value | TP (True Positive) | FP (False Positive) | FN (False Negative) | TN (True Negative) |
| True | 27 | 3 | 0 | 97 |
| False | - | - | - | - |

\[
\text{Precision} = \frac{TP}{FP+TP} \times 100\% = \frac{27}{3+27} \times 100\% = 90\% \quad (5)\]

\[
\text{Recall} = \frac{TP}{FN+TP} \times 100\% = \frac{27}{0+27} \times 100\% = 100\% \quad (6)\]

The process of predicting student graduation is influenced by several variables or fields, namely gender, origin high school, organizations, entrance, scholarships, GPA. From the results of data processing using WEKA tools with the naïve Bayes classifier method, the accuracy value is 97.6378\%, the error rate is 2.3622\%, precision is 90\% and recall is 100\%. This shows that the naïve Bayes classifier method is well used in predicting graduation of Information Systems Study Programs at the Faculty of Computer Science, Sriwijaya University.

REFERENCES

[1] L. T. Kunjumon, “An Intelligent System to predict Students academic performance using Data Mining,” Int. J. Inf. Syst. Comput. Sci., vol. 8, no. 2, pp. 128–131, 2019.

[2] M. M. Abu Tair and A. M. El-Hales, “International Journal of Information and Communication Technology Research Mining Educational Data to Improve Students’ Performance: A Case Study,” Int. J. Inf. Commun. Technol. Res., vol. 2, no. 2, pp. 140–146, 2012.

[3] S. L. Ting, W. H. Ip, and A. H. C. Tsang, “Is Naïve Bayes a good classifier for document classification?,” Int. J. Softw. Eng. its Appl., vol. 5, no. 3, pp. 37–46, 2011.
[4] A. Kesumawati and D. Waikabu, “Implementation Naïve Bayes Algorithm for Student Classification Based on Graduation Status,” *Int. J. Appl. Bus. Inf. Syst.*, vol. 1, no. 2, pp. 6–12, 2018.

[5] C. C. Le, P. W. C. Prasad, A. Alsadoon, L. Pham, and A. Elchouemi, “Text classification: Naïve Bayes classifier with sentiment Lexicon,” *IAENG Int. J. Comput. Sci.*, vol. 46, no. 2, pp. 141–148, 2019.

[6] A. A. N. Mostafa, “Review of Data Mining Concept and its Techniques,” no. April 2016.

[7] A. R. L. Francisco, *Data Mining Concepts Techniques*, vol. 53, no. 9. 2013.

[8] G. Shmueli, “To explain or to predict?,” *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, 2010.

[9] U. N. Dulhare, “Prediction system for heart disease using Naive Bayes and particle swarm optimization,” *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018.

[10] P. Kaviani and S. Dhotre, “International Journal of Advanced Engineering and Research Short Survey on Naive Bayes Algorithm,” no. March, pp. 607–611, 2017.

[11] S. Kalmegh, “Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News,” *Int. J. Innov. Sci. Eng. Technol.*, vol. 2, no. 2, pp. 438–446, 2015.

[12] T. C. Sharma and M. Jain, “WEKA Approach for Comparative Study of Classification Algorithm,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 4, pp. 1925–1931, 2013.

[13] O. Niakšu, “CRISP Data Mining Methodology Extension for Medical Domain,” *Balt. J. Mod. Comput.*, vol. 3, no. 2, pp. 92–109, 2015.