Implementation of Decision Tree and Naïve Bayes Classification Method for Predicting Study Period

N Pandiangan, M L C Buono, and S H D Loppies

Information System, Faculty of Engineering, Musamus University, Merauke, Papua, Indonesia.

Lela24pandiangan@gmail.com

Abstract. student learning performance influences the quality of a university. One indicator of assessment of student learning performance is the student’s study period. By predicting the student’s study period, universities can evaluate performance to strive to improve the quality of universities. Data mining is one of the choice technique in terms of predictions. But, each algorithm in data mining has the advantages of each, including Decision Tree and Naïve Bayes Algorithm which were tested in this research. This research shows Decision Tree is better than Naïve Bayes Algorithm to be the best choice to predicting study period with accuracy level tested for several ratios of training data and test data are 60:40,65:35,70:30,75:25 and 80:20 produce the highest accuracy values obtains by decision tree with accuracy values are 90%, 89,14%, 89,3%, 88,8% and 88%. But, the amount of training data in this research does not affect the value of accuracy. This is shown from the value of accuracy in a certain ratio is very small but sometimes also the value of accuracy is high.

1. Introduction
The period of student study is a important indicator for accreditation assessment to supports the quality of university. The University need to regard student performance through the study period as a form of evaluation to improve the student learning performance. Data mining is one technique that can be used to process student data for get useful information [1]. Data mining is a technique that can be to find information and knowledge in extensive data scope [2]. Data mining is a powerful, flexible and efficient technique. Classification is one of several techniques in data mining [3]. Naïve Bayes algorithm and Decision tree are included in the classification technique.

Naïve bayes algorithm is simple and able to classify data correctly [4]. While, Decision Tree Algorithm is also able to classify data objectively and show results easily so that becomes the best choice by decision makers in making decisions [5]. Algorithms with their respective advantages make the writer want to compare the ability of the Naïve Bayes Algorithm and Decision Tree Algorithm for classifying data, with case studies in this research is to predicting study period with some parameters. These parameters are obtained from multiple combinations previous research are college entrance
2. Related Work

Implementation of Data Mining has given great benefits in various problems. Data mining techniques are useful in finding information more accurately [8]. Naïve bayes is a data mining method can be used in making decisions that are multi-criteria, and able to manage data quickly, and produce final probability with a combination of various variables [9]. Naïve Bayes is simple in classifying final probabilistic by adding up the frequency and combination value from the dataset, the final result of accuracy is able to reach 75% for case studies taking car purchase decisions [4]. In another research, Naïve Bayes was able to classify data with 93% accuracy in classifying tweet speech speeches containing positive and negative elements [10]. Some research also has made a comparison between Naïve Bayes Algorithm and the other Classification Algorithms, such as the comparison of the Naïve Bayes and KNN Algorithms in predicting student performance in the Ministry of Education, the results of the Naïve Bayes Algorithm have the highest accuracy 93.17% [11]. Different from Naïve Bayes, Decision Tree Algorithm is the most popular algorithm because easy and high accuracy. This algorithm predicts with model decision tree structures for decision making [12].

Decision tree is simple and easy to use. The decision tree in the case certain can be become the basis of supporting Quality Control Decisions. Decision Tree is able to determine the root causes of product defects and give instructions for handling product defects happen [13]. In the field of education, Decision Tree be an alternative to predict student graduation. C4.5 Decision Tree Algorithm get an accuracy value of 77.01% compared to Random Tree and REPTree. However, the research shows fluctuating results so that the research data and the sample data does not affect results [6]. In other cases, Decision Tree produces an accuracy of 66.65% and Naïve Bayes got 79.95% to classify the feasibility of blood donation with the accuracy for decision tree is 78.5% and for Naïve Bayes is 81.5%. Other research, decision tree combined with Naïve Bayes predicting study period, and showed NBtree adapted accuracy of 77.31%, Decision Tree 68.04% and Naïve Bayes produced 65.97%. A little different from the research, in this research the authors compared decision tree algorithms and Naïve Bayes algorithms in the field of education to predict the study period by combining the parameters Predicate index (IP) for semester 1, Predicate index for semester 2, university entrance, type of secondary school, Gender, and Predicate student graduation [7], and additional parameter is college entrance value [6].

3. Design Research

The data used in this research is artificial data, with parameters for predicting study period of students are the results of the College entrance value, Predicate index (IP) for semester 1, Predicate index for semester 2, University Entrance, Type of Secondary School, Gender, and Predicate Student Graduation. Dataset use 500 data by testing the accuracy of each algorithm based on the ratio 60:40, 65:35, 70:30, 75:25 and 80:20 [6]. The student will be classified into 2 categories namely more than 4 years and less than or equal to 4 years. Dataset in this research can be seen in Table 1.

3.1. Data Processing

In this stage, the data will be converted to the .CSV format to be processed in the Weka Workbench for analysis.

3.2. Attribute Selection

In this stage, the system selects the classification attributes and method used.
3.3. Classification
In this stage, the system will classify using the Decision Tree Algorithm and Naïve Bayes Algorithm using Weka with the ratio of training data and test data are 60:40, 65:35, 70:30, 75:25, and 80:20.

3.4. Evaluation
At the stage an analysis of the classification results is carried out by calculating the accuracy value of the algorithms used.

Table 1. Dataset

| No | College entrance value | IP1   | IP2   | University Entrance | Type of Secondary School | Program Study | Gender | Predicate Student Graduation | Study Period |
|----|------------------------|-------|-------|----------------------|--------------------------|---------------|--------|------------------------------|--------------|
| 1  | 181                    | 3.715277 | 2.187607 | SNMPTN              | SMA                      | SI            | Woman | Cumlaude                      | > 4 Years    |
| 2  | 377                    | 2.025047 | 2.9183  | SBMPTN              | SMA                      | TI            | Man    | Cumlaude                      | > 4 Years    |
| 3  | 105                    | 3.498597 | 2.044178 | Mandiri             | MA                       | TI            | Man    | Very Satisfactory             | > 4 years    |
| 4  | 241                    | 3.162167 | 2.208021 | Mandiri             | MA                       | TI            | Man    | Satisfactory                  | ≤ 4 years    |
| 5  | 325                    | 3.235097 | 2.383476 | SBMPTN              | SMA                      | SI            | Woman | Satisfactory                  | > 4 years    |
|    |                        |        |        |                      |                          |               |        |                              |              |
|    |                        |        |        |                      |                          |               |        |                              |              |
| 498| 474                    | 4.793034 | 4.317007 | Mandiri             | MA                       | TI            | Woman | Cumlaude                      | > 4 Years    |
| 499| 160                    | 4.797025 | 4.320083 | Mandiri             | MA                       | SI            | Woman | Cumlaude                      | ≤ 4 years    |
| 500| 159                    | 4.801015 | 4.323159 | SBMPTN              | SMA                      | SI            | Woman | Cumlaude                      | > 4 years    |

4. Result and Discussion
From 500 dataset testing of the accuracy for Decision Tree and Naïve Bayes Algorithm, distributing training data and testing data with by ratio 60:40, 65:35, 70:30, 75:25 and 80:20 can be seen in Table 2 until Table 6. In the table, we can see the value of precision and recall of each algorithm. The value of precision represents many data positive and correctly identified from many existing data. While, the recall value represents the success rate of the system in correctly identifying data that is positively from many data positive in case and accuracy values that can be seen in graphical from in the Figure 1 will show many data of cases correctly identified by the system from many data in case [1]
Table 2. Results of Classification Analysis by Ratio 60:40

| Algorithm                        | TP Rate | FP Rate | Precision | Recall | F-Measure | Roc Area |
|----------------------------------|---------|---------|-----------|--------|-----------|----------|
| Decision Tree (More than 4 years)| 0.937   | 0.124   | 0.831     | 0.937  | 0.881     | 0.975    |
| Decision Tree (less than or equal to 4 years) | 0.876   | 0.063   | 0.955     | 0.876  | 0.914     | 0.975    |
| Naïve Bayes (more than 4 years)  | 0.468   | 0.388   | 0.440     | 0.468  | 0.454     | 0.555    |
| Naïve Bayes (less than or equal to 4 years) | 0.612   | 0.532   | 0.638     | 0.612  | 0.624     | 0.555    |

Table 3. Results of Classification Analysis by Ratio 65:35

| Algorithm                        | TP Rate | FP Rate | Precision | Recall | F-Measure | Roc Area |
|----------------------------------|---------|---------|-----------|--------|-----------|----------|
| Decision Tree (More than 4 years)| 0.881   | 0.102   | 0.843     | 0.881  | 0.861     | 0.966    |
| Decision Tree (less than or equal to 4 years) | 0.898   | 0.119   | 0.924     | 0.898  | 0.911     | 0.966    |
| Naïve Bayes (more than 4 years)  | 0.418   | 0.417   | 0.384     | 0.418  | 0.400     | 0.531    |
| Naïve Bayes (less than or equal to 4 years) | 0.583   | 0.582   | 0.618     | 0.583  | 0.600     | 0.531    |

Table 4. Results of Classification Analysis by Ratio 70:30

| Algorithm                        | TP Rate | FP Rate | Precision | Recall | F-Measure | Roc Area |
|----------------------------------|---------|---------|-----------|--------|-----------|----------|
| Decision Tree (More than 4 years)| 0.864   | 0.088   | 0.864     | 0.864  | 0.864     | 0.965    |
| Decision Tree (less than or equal to 4 years) | 0.912   | 0.136   | 0.912     | 0.912  | 0.912     | 0.965    |
| Naïve Bayes (more than 4 years)  | 0.373   | 0.407   | 0.373     | 0.373  | 0.373     | 0.535    |
| Naïve Bayes (less than or equal to 4 years) | 0.593   | 0.627   | 0.593     | 0.593  | 0.393     | 0.535    |

Table 5. Results of Classification Analysis by Ratio 75:25

| Algorithm                        | TP Rate | FP Rate | Precision | Recall | F-Measure | Roc Area |
|----------------------------------|---------|---------|-----------|--------|-----------|----------|
| Decision Tree (More than 4 years)| 0.860   | 0.093   | 0.860     | 0.860  | 0.860     | 0.964    |
| Decision Tree (less than or equal to 4 years) | 0.907   | 0.140   | 0.907     | 0.907  | 0.907     | 0.964    |
| Naïve Bayes (more than 4 years)  | 0.440   | 0.320   | 0.478     | 0.440  | 0.458     | 0.581    |
| Naïve Bayes (less than or equal to 4 years) | 0.680   | 0.560   | 0.646     | 0.680  | 0.662     | 0.581    |
Table 6. Results of Classification Analysis by Ratio 80:20

| Algorithm                           | TP Rate | FP Rate | Precision | Recall | F-Measure | Roc Area |
|-------------------------------------|---------|---------|-----------|--------|-----------|----------|
| Decision Tree (More than 4 years)   | 0.842   | 0.097   | 0.842     | 0.842  | 0.842     | 0.967    |
| Decision Tree (less than or equal to 4 years) | 0.903   | 0.158   | 0.903     | 0.903  | 0.903     | 0.967    |
| Naïve Bayes (more than 4 years)     | 0.421   | 0.194   | 0.571     | 0.421  | 0.485     | 0.572    |
| Naïve Bayes (less than or equal to 4 years) | 0.806   | 0.579   | 0.694     | 0.806  | 0.746     | 0.572    |

The results of measuring the accuracy for Decision Tree and Naïve Bayes at various ratios of training data and test data, which can be seen Figure 1.

Figure 1. Comparison of the value of accuracy between the Decision Tree Algorithm and the Naïve Bayes Algorithm

Based on the results of the comparison of the accuracy values between the Decision Tree Algorithm and the Naïve Bayes Algorithm, the highest accuracy value is obtained by the decision tree algorithm for trials with different ratios. And the accuracy value for the highest Decision tree algorithm get at the smallest ratio of 60:40 with an accuracy value of 90%. While the Naïve Bayes algorithm obtain the highest accuracy value found at ratio 80:20 with an accuracy value of 66%. Based on the results, the amount of training data does not affect the accuracy better. In both algorithms, the results of the accuracy did not increase every the ratio of training data increases.

The accuracy value of the Naïve Bayes Algorithm is erratic according to the ratio of training data and test data. This is shown in the calculation of the accuracy value of the naive bayes algorithm at the smallest ratio 60:40 gets an accuracy value of 55.5%. But when the training data ratio rises to 65%, Naïve Bayes produces an accuracy value down to 52% and drops again to 50.57% at a ratio of 70:30.
5. Conclusion
The results of the research showed Decision Tree Algorithm is the best choice in this research to predict the study period, because it has the highest accuracy value compared to the Naïve Bayes Algorithm in several ratios of training data and test data set. The amount of training data used does not affect the accuracy values of the algorithm, proven by testing against several different ratios are 60:40, 65:35, 70:30, 75:25 and 80:20. In the test, Decision Tree and Naïve Bayes show the value of accuracy that does not increase. The accuracy value at a certain ratio is sometimes small but also sometimes the accuracy value is high at a certain ratio.

References
[1] P. Prasetyawan. I. Ahmad. R. I. Borman. Y. A. Pahlevi. and D. E. Kurniawan. “Classification of the Period Undergraduate Study Using Back-propagation Neural Network.” 2018 Int. Conf. Appl. Eng., pp. 1–5. 2018.
[2] N. Mougnutou. T. Tedondjio. and T. Djamegni. “Informatics in Medicine Unlocked The prediction of good physicians for prospective diagnosis using data mining.” Informatics Med. Unlocked. vol. 12. no. March. pp. 120–127. 2018.
[3] B. Predic. M. Ilic. P. Spalevic. S. Trajkovic. S. Jovic. and A. Stanci. “Data mining based tool for early prediction of possible fruit pathogen infection.” Comput. Electron. Agric.. vol. 154. no. August. pp. 314–319. 2018.
[4] F. Harahap. A. Yugo. N. Harahap. E. Ekadiansyah. and U. P. Utama. “Implementation of Naïve Bayes Classification Method for Predicting Purchase.” 2018 6th Int. Conf. Cyber IT Serv. Manag.. no. Citsm. pp. 1–5. 2018.
[5] H. Ming and L. Xu. “An improved decision tree classification algorithm based on ID3 and the application in score analysis.” pp. 1876–1879. 2009.
[6] A. Supianto. “Decision Tree Usage for Student Graduation Classification : A Comparative Case Study in Faculty of Computer Science Brawijaya University.” 2018 Int. Conf. Sustain. Inf. Eng. Technol.. pp. 308–311. 2018.
[7] N. Renaningtias. “Hybrid Decision Tree and Naïve Bayes Classifier for Predicting Study Period and Predicate of Student ’ s Graduation.” vol. 180. no. 49. pp. 28–34. 2018.
[8] Z. Lei and G. Junfeng. “Customer Classification of Discrete Data Concerning Customer Assets Based on Data Mining.” 2019 Int. Conf. Intell. Transp. Big Data Smart City. pp. 352–355. 2019.
[9] F. Burdi. “Application of the Naive Bayes Method to a Decision Support System to provide Discounts ( Case Study : PT. Bina Usaha Teknik ).” 2016.
[10] N. R. Fatahillah. “Implementation Of Naive Bayes Classifier Algorithm On Social Media ( Twitter ) To The Teaching Of Indonesian Hate Speech.” pp. 128–131. 2017.
[11] I. A. A. Amra. “Naïve Bayesian.” pp. 909–913. 2017.
[12] “2018 International Conference on Enhanced Research and Industrial Application 17-18.” p. 25101. 2018.
[13] N. Kusumawardani. S. Putri. S. Ibrahim. K. S. Puika. and L. Darmawan. “Defect Classification using Decision Tree.” 2018 Int. Conf. Inf. Manag. Technol.. no. September. pp. 1–9. 2018.