Comparison J48 And Naïve Bayes Methods in Educational Analysis

Agung Triayudi¹, Wahyu Oktri Widyarto²

¹Department of Communication Technology and Informatics, Universitas Nasional
²Industrial Engineering Department, Universitas Serang Raya

*agungtriayudi@civitas.unas.ac.id

Abstract. In the present, data mining can be applied in various fields. One of them is the field of education. By applying data mining in the field of education, education providers can make an analysis of students in their schools. Schools can predict student achievement, make assessments of students more thoroughly, and can also predict students' interests and talents. This study will discuss the prediction of student learning habits and the prediction of student achievement in order to find out the right steps to take afterwards. In this study, two (2) classification algorithms were used, namely J48 and NaïveBayes. This is done to find the best results from each algorithm.

1. Introduction

In this era, all fields are required to carry out analysis and predictions quickly and accurately. One of these fields is the field of education. Improving the quality of education is one of the aspects most concerned by many parties because it has an important role in future progress [1]. To get the best analysis, data mining techniques are used. Data mining is an analytical technique that is used to be able to dig deeper into raw data [2] [3]. Educational Data Mining (EDM) is a scientific discipline that uses data mining techniques in education [4]. EDM in education is very helpful in the process of predicting student achievement, assessing students more thoroughly, predicting student interests and talents, and various other analyzes [5] [6] [7]. Data mining is widely used because it is very helpful in examining data using a variety of different approaches and having their own characteristics. Data mining is also used to simplify data into functional information. Data mining methods that are widely used in EDM are k-nearest neighbors, decision tree, neural network, naïve Bayes, and so on [8] [9]. To be able to perform analysis, many open-source tools can be used to implement data mining. These tools include WEKA. These tools are designed to be able to carry out data investigations and get patterns or structures that can be useful in the future [10].

In previous research, many have conducted studies on data mining with the main objective of predicting student achievement. In research [11], the k-nearest algorithm has the most effective role in classification accuracy. In research [12], there was a study on the prediction of student performance in school and used several parameters such as attendance and assignment value. This study uses the Naïve Bayes algorithm and produces the highest accuracy compared to other classification algorithms. Then, in research [13], the Decision Tree and Neural Network methods were referenced to predict student achievement because they had high accuracy scores. In research [14], a study of student performance was assessed using the Decision Tree method. As a result, this model can only produce an accuracy value of 60%. In research [15], there was a study on measuring student achievement using the Decision Tree and Neural Network methods. As a result, this study shows the effectiveness of applying the method in EDM is higher. In research [16], a study on relative research to analyze several
Decision Tree methods and their effect on educational datasets was carried out. The results show that regression analysis and classification methods are the best combination because they have a high level of compatibility to produce better results.

2. Methodology

2.1 Dataset Description

In this study, this dataset was taken from the learning management system (LMS) of a private university in Jakarta, Indonesia. The dataset consists of 340 data rows and 10 attributes. The 10 attributes used consisted of three categories, namely:

a) Demographics, namely the attributes of gender, citizenship and domicile;
b) Education, namely GPA when entering university, last GPA at university, semester, and major;
c) Habits, namely filling out parental surveys, habit of opening material, and habits of answering forums;

2.2 Classification Techniques Classification

Classification is a technique in data mining that is widely used because it is quite simple. There are two stages in the classification technique, namely developing a model for training and evaluating the model using training data. Classification also has many methods, for example statistical algorithms, correlation analysis, regression analysis, Bayesian models, distance-based algorithms, simple approach, k-nearest neighbors, decision tree, neural network, and rule-based algorithms. In this study, the methods used are Bayesian model, decision tree, and neural network.

a) Bayesian model is a method that uses the technique of applying probability to the data or it is also called the frequentist technique. Calculations using this method yield a direct value to the hypothesized probability.

b) Decision Tree is a method which has a construction condition like a tree. To be able to use this method, two (2) steps are needed, namely building a Decision Tree and then implementing it into the database.

3. Result and Discussion

After processing data on two (2) different algorithms in classification methods (J48 and NaïveBayes), each of which has its own characteristics. In this study, the focused values are CC (Correctly Classified Instances), IC (Incorrectly Classified Instances), Precision, Recall, and F-Measure values.

3.1 J48 or ID3 algorithm

Algorithm J48 is a development of the ID3 algorithm which is a conventional algorithm. This algorithm can classify numeric and discrete data using the Decision Tree method. By using this algorithm, the Correctly Classified Instances value is 71.17% and the accuracy or F-Measure is 73.6%.

Figure 1. Result of Processing Algorithm J48 or ID3
3.2 Naïve Bayes algorithm

The Naïve Bayes Algorithm is an algorithm in the classification method. This algorithm uses probability and statistical calculations. By using this algorithm, we get the Correctly Classified Instances value of 81.17% and accuracy or F-Measure 73.8%.

![Figure 2. Results of the Naive Bayes Algorithm Processing](image)

3.6 Performance Result

The results from processing the data with these five (5) different algorithms are as follows.

| Criteria     | NaiveBayes | ID3 |
|--------------|------------|-----|
| Accuracy / F-Measure (%) | 73.8 | 73.6 |
| CC           | 276        | 242 |
| IC           | 64         | 80  |
| Recall       | 81.2       | 75.2 |
| Precision    | 75.6       | 72.4 |

It can be seen that the five algorithms have a good accuracy rate (more than 70%). The highest accuracy value is found in the NaïveBayes algorithm (73.8%), then ID3 (73.6%). Meanwhile, for CC or Correctly Classified Instances, the highest value is also found in the NaïveBayes algorithm (276 data), then ID3 (242 data). Overall, based on the table above, the Naïve Bayes algorithm has the highest accuracy value based on F-Measure, Recall, and Precision compared to other algorithms. The image below is a graphical representation of the accuracy values (based on the F-Measure value) of the five algorithms.

![Figure 3. Representation Accuracy](image)
4. Conclusion

In the world of education, data mining is needed to perform predictive analysis of student datasets. One of the information or knowledge that can be generated from data mining techniques is the prediction of student performance and achievement, which will then be used to decide the next steps for these students. In this study, the dataset used is data taken from a university's learning management system (LMS). This dataset has 340 data rows and 10 attributes. For processing, two (2) classification algorithms are used, namely J48 and NaïveBayes. These two algorithms have different characteristics. As a result, the two algorithms show good accuracy (> 70%). However, the Naïve Bayes algorithm has the highest accuracy value (based on F-Measure, Recall, and Precision) and the highest Correctly Classified Instances value compared to other algorithms. For further research, so that more data lines are used and also tried using other data mining techniques.

References

[1] Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In 2017 8th International Conference on Information Technology (ICIT) (pp. 909-913). IEEE.

[2] Ramaphosa, K. I. M., Zuva, T., & Kwuimi, R. (2018, August). Educational data mining to improve learner performance in Gauteng primary schools. In 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-6). IEEE.

[3] Kumar, A. D., Selvam, R. P., & Kumar, K. S. (2018). Review on prediction algorithms in educational data mining. International Journal of Pure and Applied Mathematics, 118(8), 531-537.

[4] Triayudi, A., Sumiati, S., Nurhadiyan, T., & Rosalina, V. (2020). Data Mining Implementation to Predict Sales Using Time Series Method. Proceeding of the Electrical Engineering Computer Science and Informatics, 7(2), 1-6.

[5] Jalota, C., & Agrawal, R. (2019, February). Analysis of educational data mining using classification. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 243-247). IEEE.

[6] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining using classification. In 2019 International Conference on Information Technology, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 243-247). IEEE.

[7] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.

[8] Fitri I, Triayudi A, Iksal, Muttaqin Z, Sumiati. Visualization of Data Mining Distribution of COVID-19 in Indonesia Using Self-Organizing Maps Algorithm. Icic Express Letters. 2021, Vol. 15 (3), pp. 241-248.

[9] Rawat, K. S., & Malhan, I. V. (2019). A hybrid classification method based on machine learning classifiers to predict performance in educational data mining. In Proceedings of 2nd International Conference on Communication, Computing and Networking (pp. 677-684).
Springer, Singapore.

[10] Hegde, V., & Prageeth, P. P. (2018, January). Higher education student dropout prediction and analysis through educational data mining. In 2018 2nd International Conference on Inventive Systems and Control (ICISC) (pp. 694-699). IEEE.

[11] Huang, J., Lu, J., & Ling, C. X. (2003, November). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. In Third IEEE International Conference on Data Mining (pp. 553-556). IEEE.

[12] Nakayama, M., Mutsuura, K., & Yamamoto, H. (2018). Using note taking instructions to reform student’s note taking activities and improve learning performance in a blended learning course. International Conference Information Visualisation, 326-331.

[13] Costa, Evandro B., Balduino Fonseca, Marcelo Almeida Santana, Fabrisia Ferreira de Araújo, and Joilson Rego. "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses." Computers in Human Behavior 73 (2017): 247-256.

[14] Silva, Carla, and José Fonseca. "Educational Data Mining: a literature review." In Europe and MENA Cooperation Advances in Information and Communication Technologies, pp. 87-94. Springer, Cham, 2017.

[15] Baker, Ryan S. "Challenges for the future of educational data mining: The Baker learning analytics prizes." JEDM| Journal of Educational Data Mining 11, no. 1 (2019): 1-17.

[16] Fernandes, Eduardo, Maristela Holanda, Marcio Victorino, Vinicius Borges, Rommel Carvalho, and Gustavo Van Erven. "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil." Journal of Business Research 94 (2019): 335-343.