Educational Data Mining Analysis Using Classification Techniques

Agung Triayudi\textsuperscript{1*}, Wahyu Oktri Widyarto\textsuperscript{2}
\textsuperscript{1}Department of Communication Technology and Informatics, Universitas Nasional
\textsuperscript{2}Industrial Engineering Department, Universitas Serang Raya

*agungtriayudi@civitas.unas.ac.id

Abstract. Currently, data mining is being needed in various fields to obtain analysis results from certain aspects that are needed. One of the fields that make use of data mining is education. The education sector uses data mining to determine the level of performance or achievement of its students. With data mining, the education sector can also evaluate student achievement and take the next step in increasing that achievement. Data mining in education is also known as Educational Data Mining (EDM). In this study, the theme discussed was the prediction of student learning habits and steps that could be taken to improve student achievement at the university. In this study, three (3) classification algorithms were used, namely Multilayer Perceptron, Random Forest, and Support Vector Machine. This is done to find the best results from each algorithm.

1. Introduction

In this era of industrial revolution 4.0, almost all fields in the world need data mining analysis. For example, the petroleum sector requires data mining analysis to be able to see the strategic potential of oil fields \cite{1}. The health sector requires data mining analysis to be able to predict hypertension in pregnancy. In this research, the field to be discussed is data mining in education. Data mining in education is called Educational Data Mining. Educational Data Mining utilizes data related to education; the data mining process can be done to find useful information for advancement in education \cite{2,3}. One of the purposes of using educational data mining is to determine the predictions of student interest and talents, predict student achievement, predict the length of study at university, and various other analyzes \cite{4,5,6}. In processing data, educational data mining has quite a variety of algorithms \cite{7,8}. For example, the decision tree, k-nearest neighbors, naïve bayes, neural network, and so on. For processing implementation, educational data mining can be processed using applications such as WEKA, Rapid Miner, GNU Octave, Orange, KNIME, MATLAB, and so on \cite{9}.

In previous research, many have conducted studies on data mining with the main objective of predicting student achievement. In research \cite{10}, the authors analyzed previous research conducted on the prediction of student performance analysis and assessment by applying different data mining techniques using the Algorithm Support Vector Machine. In research \cite{11}, the k-nearest algorithm has the most effective role in classification accuracy. As a result, this model can only produce an accuracy value of 60\%. In research \cite{12}, the WEKA application was used to prognosis students' scores in the final year and was based on two different dataset parameters. There is one general information in each dataset, namely that various students can be drawn from one course in the last two years. In research \cite{13}, 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 [14], 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

The dataset used in this study amounted to 340 rows of data with 10 attributes. Data is taken from a learning management system (LMS) from a university. The 10 attributes taken are divided into 3 categories, namely Demography (attributes in the form of gender, citizenship, and domicile), Education (attributes in the form of high school final grades, GPA at universities, semesters, and study programs), and Habits (attributes in the form of questionnaire results, parents of students, the number of clicks on the subject matter, and the number of student discussions in the learning forum).

| No | Attribute          | Description                                                                 |
|----|--------------------|----------------------------------------------------------------------------|
| 1  | GENDER             | Gender (Male, Female)                                                      |
| 2  | CITIZATIONSHIP     | Citizenship (Indonesia, Thailand, America, Australia)                       |
| 3  | GPA_HS            | School Examination Score when Entering University (Fair, Average, Good, Very Good) |
| 4  | DOMICILE           | Domicile (Jakarta, Bogor, Depok, Tangerang, Bekasi)                        |
| 5  | SEMESTER           | Study Current Year (1st Year, 2nd Year, 3rd Year, 4th Year)                |
| 6  | PARENT_SURVEY      | Filling in Parents Questionnaire (Filled, Not Filled)                      |
| 7  | RESOURCE_OPEN      | Opening Material on LMS (Sometimes, Always)                               |
| 8  | RAISEDHAND         | Answering Questions during Forums (Sometimes, Often, Always)               |
| 9  | GPA                | Last GPA Score in Universities (Fair, Average, Good, Very Good)            |
| 10 | MAJORS             | Departments at Universities (Informatics, Information Systems, Management) |

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 decision tree and neural network.

a) Neural Network is a method that provides a format for data representation. This method is similar to a network of biological learning systems that occur from a network of nerve cells (neurons) that are connected to each other.

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 three (3) different algorithms in classification methods (SVM, Random Forest, and Multilayer Perceptron), 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 Support Vector Machine algorithm

Algorithm Support Vector Machine (SVM) or SMO algorithm is an algorithm used for regression analysis and classification methods. At WEKA, SVM is referred to as SMO. By using this SVM or SMO algorithm, we get the Correctly Classified Instances value of 80.58% and accuracy or F-Measure of 72.5%.

3.2 Random Forest algorithm

The Random Forest algorithm is one of the algorithms in the Decision Tree which is a combination of each tree and then combined into a particular model. By using this algorithm, we get the Correctly Classified Instances value of 78.82% and accuracy or F-Measure of 72.9%.
3.3 Multilayer Perceptron Algorithm

Multilayer Perceptron Algorithm is one of the artificial neural network (ANN) methods which can produce a more accurate classification because it has better weight characteristics than other modeling. By using this algorithm, we get the Correctly Classified Instances value of 74.70% and accuracy or F-Measure of 73%.

3.6 Performance Result

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

Table 2. Performance Results

| Criteria          | Classifier       |
|-------------------|------------------|
|                   | RandomForest     | MLP             | SVM             |
| Accuracy / F-Measure (%) | 72.9             | 73              | 72.5            |
| CC                | 268              | 254             | 274             |
| IC                | 72               | 86              | 66              |
| Recall            | 78.8             | 74.7            | 80.6            |
| Precision         | 70.0             | 71.7            | 65.8            |

It can be seen that the five algorithms have a good accuracy rate (more than 70%). The highest accuracy value is found in the MLP algorithm (73%), then RandomForest (72.9%), and SVM (72.5%).
Meanwhile, for CC or Correctly Classified Instances, the highest value is found in the SVM algorithm (274 data), then RandomForest (268 data), and MLP (254 data). Overall, based on the table above, the SVM 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 4. Representation Accuracy](image1)

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, three (3) classification algorithms are used, namely SVM, RandomForest, and MultilayerPerceptron. These three algorithms have different characteristics. As a result, the three algorithms show good accuracy (>70%). However, it does not rule out that there are algorithms other than SVM, RandomForest, and MultilayerPerceptron that have a higher level of accuracy. For further research, so that more data lines are used and also tried using other data mining techniques.

References

[1] G. Zhai, Y. Yang, H. Wang and S. Du, "Multi-attention fusion modeling for sentiment analysis of educational big data," in Big Data Mining and Analytics, vol. 3, no. 4, pp. 311-319, Dec. 2020, doi: 10.26599/BDMA.2020.9020024.

[2] S. M. Merchan Rubiano and J. A. Duarte Garcia, "Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance," in IEEE Latin America Transactions, vol. 14, no. 6, pp. 2783-2788, June 2016, doi: 10.1109/TLA.2016.7555255.
[3] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," in IEEE Access, vol. 8, pp. 55462-55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[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] M. Munk, M. Drlík, L. Benko and J. Reichel, "Quantitative and Qualitative Evaluation of Sequence Patterns Found by Application of Different Educational Data Preprocessing Techniques," in IEEE Access, vol. 5, pp. 8989-9004, 2017, doi: 10.1109/ACCESS.2017.2706302.

[6] 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.

[7] J. Luis Cavalcanti Ramos, R. Euiller Dantas e Silva, J. Carlos Sedraz Silva, R. Lins Rodrigues and A. Sandro Gomes, "A Comparative Study between Clustering Methods in Educational Data Mining," in IEEE Latin America Transactions, vol. 14, no. 8, pp. 3755-3761, Aug. 2016, doi: 10.1109/TLA.2016.7786360.

[8] M. Á. Prada et al., "Educational Data Mining for Tutoring Support in Higher Education: A Web-Based Tool Case Study in Engineering Degrees," in IEEE Access, vol. 8, pp. 212818-212836, 2020, doi: 10.1109/ACCESS.2020.3040858.

[9] 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.

[10] O. Moscoso-Zea, J. Castro, J. Paredes-Gualtor and S. Luján-Mora, "A Hybrid Infrastructure of Enterprise Architecture and Business Intelligence & Analytics for Knowledge Management in Education," in IEEE Access, vol. 7, pp. 38778-38788, 2019, doi: 10.1109/ACCESS.2019.2906343.

[11] Hu, Q., Huang, Y. and Deng, L., A Multivariate Learning Evaluation Model for Programming Course in Online Learning Environment. In 2019 14th International Conference on Computer Science & Education (ICCSE) (pp. 737-740). IEEE.

[12] Ifenthaler, D., Gibson, D.C. and Zheng, L., The Dynamics of Learning Engagement in Challenge-Based Online Learning. In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT) (pp. 178-182). IEEE.

[13] Anggrawan, A. and Jihadil, Q.S., Comparative Analysis of Online E-Learning and Face To Face Learning: An Experimental Study. In 2018 Third International Conference on Informatics an Computing (ICIC) (pp. 1-4). IEEE.

[14] Sakulwichitsintu, S., Colbeck, D., Ellis, L. and Turner, P., A Peer Learning Framework for Enhancing Students' Learning Experiences in Online Environments. In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT) (pp. 168-169). IEEE.