STUDENT PERFORMANCE ANALYSIS USING C4.5 ALGORITHM TO OPTIMIZE SELECTION

Hilda Amalia[1*]; Yunita[2]; Ari Puspitasari[3]; Ade Fitria Lestari[4]

[1,2,3] Information Systems; [4] Accounting Information System;
Univeristas Bina Sarana Informatika
bsi.ac.id
[1]hilda.ham@bsi.ac.id; [2]yunita.ynt@bsi.ac.id; [3]ari.arp@bsi.ac.id; [4]ade.afr@bsi.ac.id

(*) Corresponding Author

Abstract— *Education is one of the fields that generate heaps of data. Pile of data that can utilized by higher education institutions to improve tertiary performance. One way to process data piles in the education is to use data mining or called education data mining. The quality assessment of educational institutions conducted by the community and the government is strongly influenced by student performance. Students who have poor performance will have a negative impact on educational institutions. Student data is processed to obtain valuable knowledge regarding the classification of student performance. One method of data mining is the C4.5 algorithm which is known to be able to produce good classifications. In this research and optimization method will be used namely optimize selection on the c4.5 algorithm. Based on the research, it is known that the optimization selection optimization method can improve the performance of algorithm c4.5 from 85% to 87%.*

Keywords: *Student Performance, C4.5 Algorithm, Optimize Selection*

Abstrak— *Dunia pendidikan merupakan salah satu gudang data terbesar yang bisa diolah dan dimanfaatkan untuk kemajuan suatu instansi pendidikan. Salah satu cara mengolah tumpukan data dunia pendidikan adalah dengan menggunakan data mining atau biasa disebut data mining pendidikan. Dalam dunia pendidikan kinerja siswa merupakan hal penting yang wajib dilakukan oleh semua instansi pendidikan. Hal ini disebabkan penilaian kwalitas instansi pendidikan yang dilakukan oleh masyarakat dan pemerintah sangat dipengaruhi oleh kinerja siswa. Siswa yang memiliki kinerja buruk akan memberikan dampak buruk bagi instansi pendidikan. Untuk itu setiap institusi pendidikan berusaha dengan keras untuk meningkatkan kinerja siswanya. Data mining merupakan suatu metode yang dapat melakukan penggalian pengetahuan dari tumpukan data. Data siswa diolah untuk mendapatkan pengetahuan yang berharga mengenai klasifikasi kinerja siswa. Salah metode data mining yaitu algoritma C4.5 yang dikenal mampu menghasilkan klasifikasi yang baik. Dalam penelitian ini akan digunakan metode optimasi yaitu optimize selection pada algoritma c4.5. Berdasarkan penelitian diketahui bahwa metode optimasi optimize selection mampu meningkatkan kinerja algotima c4.5 dari 85% menjadi 87%. Dari hasil penggunaan optimize selection diketahui bahwa dari 22 atribut yang digunakan ada sebanyak 10 atribut yang menghasilkan kinerja baik.*

Kata Kunci*: Kinerja Siswa, Algoritma C4.5, Optimize Selection*

INTRODUCTION

Educational institutions are the main resource to produce good students and provide the best service to the community, therefore each educational institution is guided to be able to understand and improve student competencies (Umadevi & Dhanalakshmi, 2017). The education system requires an innovative effort to improve the quality of education and get the best results and reduce the failure rate (Ashraf et al., 2018). Innovations are made by utilizing information received from the data mining process for planning the educational process (Ashraf et al., 2018). Lack of knowledge about the main processes of education, namely educational planning, evaluation and marketing (Abu-Oda & El-Halees, 2008). For this reason, it is necessary to process student data or educational data so that further information is obtained about student needs in depth so that the goals of educational institutions to produce quality students can be achieved. one of the methods commonly used to extract data into valuable knowledge is data mining.

Data mining is very important in the business world, namely in the decision-making process and obtaining and obtaining knowledge from data stored in an organization (Ashraf et al., 2018). Data mining is used in all aspects of fields

that have data warehouses or data stacks (Patil, 2015). One of them is in the field of education, in the world of education there are piles of student data that are only stored. Data mining in education is currently commonly referred to as educational data mining (Hussain et al., 2018).

Educational data mining is a method of mining knowledge based on data from the world of education, data is collected from the history and operations of the educational institution itself (Kalpana & Venkatalakshmi, 2014). Educational data mining comes with a paradigm for designing models, tasks and methods on educational data to make predictions about student achievement and behavior patterns (Peña-Ayala, 2014). To overcome the problems faced by educational institutions, new knowledge is needed based on educational data mining so that the main tasks of educational institutions can be achieved. educational data mining is currently commonly known as Educational Data Mining (EDM).

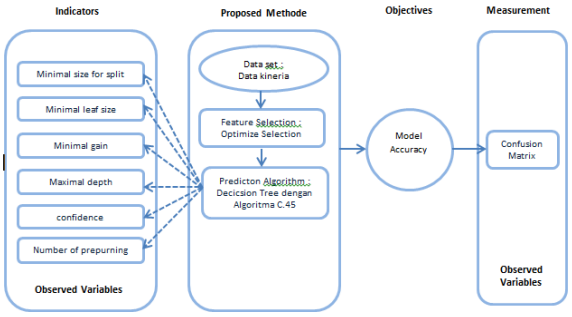
By using educational data mining techniques, the authorities in educational institutions can have ideas and then be able to make the right decisions for students who have academics, through data mining to obtain trends or predictive patterns of student performance (Soni et al., 2018). Predicting student performance also assists educational planning and decision making for educational institution leaders to make adequate change plans (Hamoud et al., 2018) Classification is an effective way to apply or process educational data. One of the tasks of educational data mining is being able to map the results so that they are able to carry out three stages, namely taking precautions, planning and recommendations for students who are likely to have poor performance

Research using educational data has been carried out by Amalia in 2014 using the C4.5 algorithm method (Amalia, 2014). Patil in 2015 conducted research on the use of data mining classification methods for student performance assessment (Patil, 2015). Kavipriya in 2016 predicted student performance using a decision tree, naïve bayes, neural network, SVM and DT-J48 which showed the highest accuracy value obtained by the Decision Tree-J48 (Kavipriya, 2016). Mendood in 2017 conducted research on student performance predictions and risk analysis using decision trees, naïve Bayes ID3 and random forest (Mehboob et al., 2017). Hussain etc In 2018 conducted research using educational data, namely student data from three colleges in India using data mining techniques, namely the classification method, namely the J48, PART, Random Forest and Basiyan classifier using WEKA tools. The results of the study show that the highest accuracy value is obtained from the Random Forest method(Hussain et al., 2018)

In this study, educational data processing will be carried out, namely to obtain predictions of student performance using the data mining classification method, namely the decision tree, namely the C4.5 algorithm, which improves the performance of the data mining method by using the optimize selection method

MATERIALS AND METHODS

In completing this research, a useful frame of mind is made as a guide so that the research is carried out consistently. The following is the framework used:

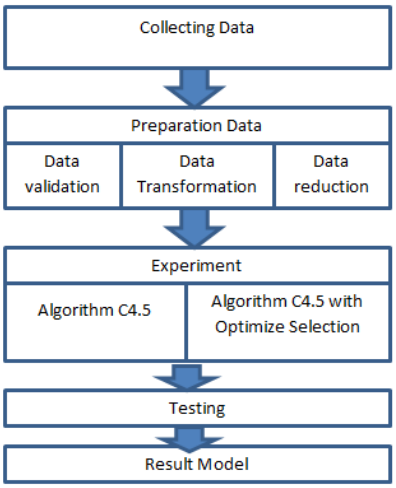


Source: (Amalia et al., 2020)

Figure 1 Research Framework

Figure 1 is the research framework carried out in this study. The dataset used is the student performance dataset on the uci repository web page. The preprocessing stage using an optimized selection algorithm (Proposed Method) by processing input variables and the resulting target dataset with a minimum size for split, minimum leaf size, minimum gain, maximal depth, confidence and number of preparing as indicators. To measure or assess the accuracy of the method used, confusion matrix is used with an assessment of the higher the confusion matrix value, the more accurate the prediction will be(Yunita, 2017). In conducting this research, several stages were taken; the research stages used were presented in Figure 2. Represents the stages or steps carried out in this study. Starting from the initial stages, namely collecting data, processing the initial data which consists of three steps, namely data validation, data transformation and reducing existing data, the next stage is to experiment with processing the dataset with the algorithm used, namely the C4.5 algorithm, and improve its performance by the optimize selection method, the next step is to test the resulting model, and the last

stage is to obtain the most accurate method among the two methods used.



Source: (Amalia et al., 2020)

Figure 2 Stages of Research

RESULTS AND DISCUSSION

Collecting Data

The data is obtained from the data provider site, namely Uci repository. The data used is a dataset of student performance. The dataset used consists of 21 attributes and 1 attribute as a label. The following attribute table is used:

Table 2 Attributes used

Attribute	Decription
Gd	Gender
Cas	Caste
AttdX	Percentage of attained at Class X
IAS	Percentage of Internal Assessments
Ex	Examination of end semester
Tddix	Percentage of attained at Class IX
ARR	failed previous semesters
SM	Status of Mariagge
PoL	Place of Leaving
IoF	Income of Family
FSiz	Family Size
Fqua	Father Qualification
Mqua	Mother Qualification
FOc	Father Occupation
MOc	Mother Occupation
NoF	Number of friend
SH	Study Hours
Attx	Attended at Class X level
ME	Medium
Distance	Vechile
Atd	Class attendance

Source: (Amalia et al., 2020)

Table 2 is an explanation of the attributes used, namely the attribute abbreviations used in the dataset, description or clear names of the attributes and the value or content of each attribute.

Initial data processing

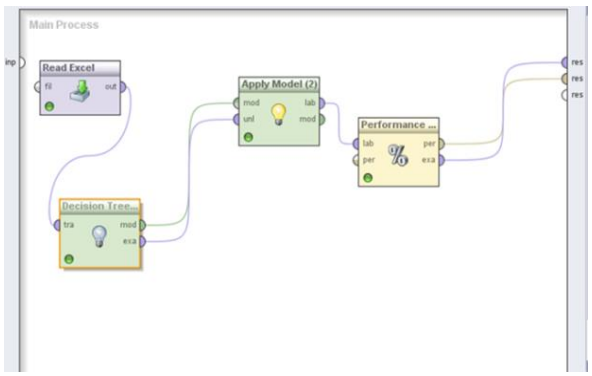
In this study, initial data management was carried out, which aims to obtain representative research results so as to get good results. There are three initial data processing techniques, namely (a) data validation, namely the initial data management technique that aims to eliminate noise or incomplete or repetitive records, (b) data transformation is a technique of changing and unifying the arrangement of records on a dataset but not changing the contents record aims to eliminate (Amalia, 2017) In this study, the data used are secondary and data that have been processed early so that the data used has already taken the three preliminary data processing techniques so that the dataset can be used immediately at the next stage.

Experiment

The experiment was carried out by processing the student performance dataset using the C4.5 algorithm method, then the experiment was carried out again with the student performance dataset using the C4.5 algorithm method which improved its performance by using the optimize selection method. The tool used in processing this student performance dataset is Rapid Miner.

C4.5 algorithm experiment

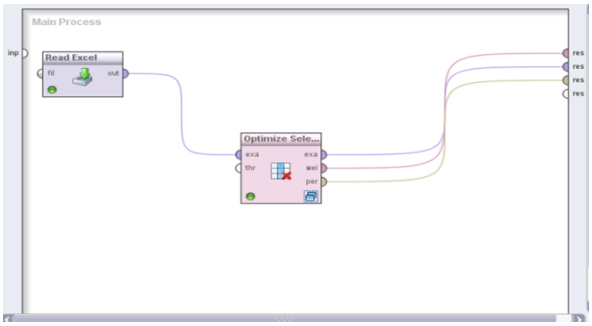
To conduct the experiment using several modules, namely the read excel module, decision tree for data processing using the C4.5 algorithm method, apply the model and performance to produce accuracy values, kappa values and error classification values. Figure 2 presents the steps of the student data processing experiment using the C4.5 decision tree algorithm with the RapidMiner tools:



Source: (Amalia, Yunita, Puspita, & Lestari, 2020)

Figure 3.Experiment using the C4.5 algorithm C4.5

Figure 3 is an experiment in processing the student performance dataset using the C4.5 algorithm method



Source: (Amalia et al., 2020)

Figure 4 Experiments using the C4.5 algorithm and Optimize Selection

In Figure 4, an experiment on the processing of the student performance dataset using the C4.5 algorithm method is presented, which is enhanced by the Optimize Selection method. In this stage of testing, the module used is the read excel module where the dataset is located, the optimize selection module as a method used to improve the performance of the C4.5 algorithm method.

Model Testing

The resulting model from the experimental stage will be tested using the confusion matrix table produced by each method

accuracy: 85.50%

	true Good	true Vg	true Best	true Pass
pred. Good	51	1	1	7
pred. Vg	3	41	4	3
pred. Best	0	0	3	0
pred. Pass	0	0	0	17
class recall	94.44%	97.62%	37.50%	62.96%

Source: (Amalia et al., 2020)

Figure 5 Figure confusion matrix table for the C4.5 method

Figure 5 presents the accuracy value obtained by the C4.5 algorithm method for processing the student performance dataset is 85.50. The labels of this study have four labels and each of them is given a value based on the processing of the c4.5 algorithm method by providing a prediction of attributes that are predicted to be good to produce a value or label of good correctly as many as 51 and produce Vg of 1 and predictable good results in the best value of 1 and produce a pass value of 7. Vg prediction yields good values as much as 3, Vg as much as 41, best as much as 4 and pass as much as 3. The best prediction yields good as much as 0, Vg as much as 0, best as much as 3 and pass as much as 0. good value as much as 0, Vg as much as 0, Best as much as 0 and pass as much as 17. The following table confusion matrix generated by the C4.5 algorithm method is improved by using optimizes selection:

accuracy: 87.79%

	true Good	true Vg	true Best	true Pass
pred. Good	52	4	1	3
pred. Vg	1	36	2	2
pred. Best	0	2	5	0
pred. Pass	1	0	0	22
class recall	96.30%	85.71%	62.50%	81.48%

Source: (Amalia et al., 2020)

Figure 6.The C4.5 confusion matrix algorithm table to optimize selection

Figure 6 presents the accuracy value obtained by the C4.5 algorithm method for processing the student performance dataset is 87.79. The label of this study has four labels and each of them is given a value based on the processing of the c4.5 algorithm method by optimizing the selection to obtain the predicted value for each label attribute used. The prediction for the value of good results in a good value or label of good as much as 52 and produces 4 Vg, 1 best and 3 passes. good as much as 0, Vg as much as 2, best as much as 5 and pass as much as 0. The prediction of the pass that produces good value is 1, Vg is 0, Best is 0 and the pass is 12.

The results of processing the student performance dataset using the optimization algorithm C4.5 method produce attributes that have or do not affect this research. The following table 2 results from the optimize selection:

Table 2 Optimize Selection Results

Attribute	Values
Gd	0
Cas	0
AttdX	1
IAS	1
Ex	1
Tddix	1
ARR	1
SM	0
PoL	0
IoF	0
FSiz	1
Fqua	0
Mqua	1
FOc	1
MOc	1
NoF	1
SH	1
Attx	1
ME	1
Distance	1
Atd	0

Source: (Amalia et al., 2020)

Table 2 presents data regarding the results of the optimize selection method processing. In the table there are two columns, namely attributes and values. Attribute contains the names or abbreviations of the attributes used and the values that contain the numbers 0 and 1. The number 0 is the value given to attributes that have no effect on the student performance dataset research and the number 1 is the value for the attributes that affect the student performance dataset research.

CONCLUSION

Based on research that has been conducted using a student performance dataset totaling 138 records and consisting of 22 attributes using the C4.5 algorithm data mining method, an accuracy value of 85% is obtained, while processing the student performance dataset using the C4.5 algorithm method uses the method for improve data mining performance, namely the optimize selection method, obtained an accuracy value of 87%. So that in this study it can be concluded that the optimize selection optimization method can work well so that it can increase the accuracy value which is better than the accuracy value generated by using only the c4.5 algorithm data mining method. From the results of the research, namely using the optimize selection method, it is obtained that the table for processing student performance data by using optimize selection is known that several influential attributes are attributes that get a value of 1 student performance namely AttdX, IAS, Ex, Tddix, ARR, Fsiz, Mqua, FOc, MOc, NoF, SH, Attx, ME, Distance, as many as thirteen attributes and there are several attributes that have no effect, namely attributes that get a value of 0 in processing student performance datasets, namely Gd, Cas, SM, PoL, IoF, Fqua and Atd or as many as seven attribute

REFERENCE

Abu-Oda, G. S., & El-Halees, A. M. (2008). DATA MINING IN HIGHER EDUCATION : UNIVERSITY STUDENT DROPOUT CASE STUDY. *Annals of DAAAM and Proceedings of the International DAAAM Symposium*, *5*(1), 827–828. https://doi.org/10.2507/daaam.scibook.2009.11

Amalia, H. (2017). Penerapan Naïve Bayes Berbasis Genetic Algorithm Untuk Penentuan Klasifikasi Donor Darah. *Jurnal Teknik Komputer*, *2*(2), 70–76.

Amalia, H. (2014). Prediction of students graduation using algorithm c4.5. *International Seminar on Scientific Issues and Trends (ISSIT) 2014*, 31–35. http://issit.bsi.ac.id/proceedings/index.php/issit2014/article/view/31

Amalia, H., Yunita, Y., Puspitasari, A., & Lestari, A. F. (2020). *Laporan Akhir Penelitian Mandiri: Analisa Kinerja Siswa Dengan Menggunakan Metode Optimize Selection Pada Algoritma C4.5*.

Ashraf, A., Anwer, S., & Khan, M. G. (2018). A Comparative Study of Predicting Student ' s Performance by use of Data A Comparative Study of Predicting Student ' s Performance by use of Data Mining Techniques. *American Scientific Research Journal for Engineering, Technology and Sciences*, *44 No.1*(October), 122–136.

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26. https://doi.org/10.9781/ijimai.2018.02.004

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Kalpana, J. K. J., & Venkatalakshmi, K. (2014). DM: Intellectual Performance Analysis of Students' by using Data Mining Techniques. *India*, *3*(3), 1922–1929.

Kavipriya, P. (2016). A Review on Predicting Students ' Academic Performance Earlier , Using Data Mining Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(12), 101–105.

Mehboob, B., Liaqat, R. M., & Nazar, A. (2017). Student Performance Prediction and Risk Analysis by Using Data Mining Approach. *Journal of Intelligent Computing*, *8*(2), 49–57.

Patil, P. (2015). a Study of Student'S Academic Performance Using Data Mining Techniques. *International Journal of Research in Computer Applications and Robotics*, *3*(9), 59–63.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4 PART 1), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Soni, A., Kumar, V., Kaur, R., & Hemavathi, D. (2018). PREDICTING STUDENT PERFORMANCE USING DATA MINING TECHNIQUES. *International Journal of Computer Sciences and Engineering*, *6*(10), 172–177. https://doi.org/10.26438/ijcse/v6i10.17217

7

Umadevi, B., & Dhanalakshmi, R. (2017). A Comprehensive Survey of Students Performance Using Various Data Mining Techniques. *International Journal of Science and Research (IJSR)*, *6*(4), 2233–2238. https://www.ijsr.net/get_abstract.php?paper_id=ART20172940

Yunita. (2017). Seleksi Fitur Menggunakan Backward Elimination Pada Prediksi Cuaca Dengan Neural Network. *Indonesia Journal on Computer and Information Technology (IJCIT)*, *2*(1), 26–37.