APPLICATION OF DECISION TREE AND NAIVE BAYES ON STUDENT PERFORMANCE DATASETS

Hilda Amalia[1*)] ; Ari Puspita[2]; Ade Fitria Lestari[3]; Frieyadie[4]

[1,2] Sistem Informasi; [3] Sistem Informasi Akuntansi;
Univeristas Bina Sarana Informatika
www.bsi.ac.id
hilda.ham@bsi.ac.id [1*)] ; ari.arp@bsi.ac.id[2]; ade.afr@bsi.ac.id[3]

[4]Sistem Informasi
Universitas Nusa Mandiri
www.nusamandiri.ac.id
[4]frieyadie@nusamandiri.ac.id

(*) Corresponding Author

Abstract— Student performance is the ability of students to deal with the entire academic series taken during school. Student performance produces two labels, namely successful and unsuccessful students. Successful students can graduate with excellent, excellent, and suitable performance labels. At the same time, students who have a label on average are students who get poor performance. Measurement of student performance is needed for every educational institution to take strategic steps to improve student performance. This study aimed to obtain a data mining method that worked well on student performance datasets. In this study, student performance datasets were processed, which had 11 indicators with one result label. Student performance datasets are processed using data mining methods, namely decision tree and nave Bayes, while the tool used for dataset processing is WEKA. The research results from processing student performance datasets obtained that the accuracy value for the decision tree method was 94.3132%, and the accuracy produced by the naive Bayes method was 84.8052%.

Keywords: student performance, decision tree, naïve Bayes

Abstrak— *Kinerja siswa adalah kemampuan siswa dalam menghadapi seluruh rangkaian akademik yang ditempuh selama sekolah. Kinjera siswa meghasilkan dua label yaitu siswa yang berhasil dan tidak berhasil. Siswa yang berhasil merupakan siswa yang mampu lulus dengan label kinerja luar biasa, sangat baik, dan baik. Sedangkan siswa yang memiliki label yaitu rata-rata merupakan siswa yang mendaparkan kinerja kurang baik. Pengukuran kinerja siswa diperlukan untuk setiap lembaga pendidikan untuk dapat mengambil langkah-langkah strategis untuk meningkatkan kinerja siswa. Tujuan penelitian ini adalah mendapatkan metode data mining yang bekerja baik pada dataset kinerj siswa. Dalam penelitian ini dilakukan pengolahan dataset kinerja siswa yang memiliki 11 indikator dengan satu label hasil. Pengolahan dataset kinerja siswa dilakukan dengan menggunakan metode data mining yaitu decision tree dan naïve bayes sementara alat yang digunakan untuk pengolohan dataset adalah WEKA. Hasil penelitian dari pengolahan dataset kinerja siswa diperoleh hasil nilai akurasi untuk metode decision tree yaitu 94.3132% dan akurasi yang dihasilkan oleh metode naïve bayes yaitu 84.8052%.*

Kata Kunci*: kinerja siswa, decision tree dan naïve bayes*

INTRODUCING

Student performance is essential for quality education to produce quality graduates (Fajar, Hussain, Sarwar, Afzal, & Gilani, 2019). Quality human resources can improve the life of a nation. Student performance is the result obtained by students during their education. Good student performance is needed by the students themselves and by educational institutions. The success of an educational institution can be measured by the output it produces, namely its graduates. The success of student performance affects the accreditation of educational institutions. Accreditation of educational institutions is a value given by the government in providing education. The accreditation value obtained is one of the critical assessments considered by the community when choosing a place to continue their education. Educational institutions are a significant factor in producing quality graduates. Student success needs to be predicted by educational institutions to identify difficulties experienced by students and provide proactive action on student difficulties so

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2714

that they can help students get good grades (Masangu, Jadhav, & Ajoodha, 2021). For this reason, an educational institution can analyze and multiply information based on its educational dataset to find patterns of student success.

Student performance needs to be measured, and this aims to determine the number of students who have less success and preventive measures can be taken (Khasanah & Harwati, 2017). One way to analyze student performance is to use data mining techniques. Data mining is a method that can be used to gain valuable knowledge from existing data (Amalia, Yunita, Puspita, & Lestari, 2020), 2020). Data mining in education or educational data mining has developed rapidly to analyze student performance, and the success of EDM is strongly influenced by the attributes and datasets used (Asraf, Anwer, & Khan, 2018). Educational institutions can use the results of EDM processing to make new policies (Hussain, Dahan, Ba-Alwi, & Ribata, 2018). Finding the best data mining model can improve student performance (Sumitha & Vinothkumar, Prediction of Students Outcome Using Data Mining Techniques, 2016). Factors that affect student performance are parental education (Olufemi, Adederin, & Oyediran, 2018), parental income (Brew, Nketiah, & Korentang, 2021). The use of different attributes can affect the value resulting from student performance. From some literature, it is known that different attributes in the use of student performance datasets will produce different accuracy values for each data mining method.

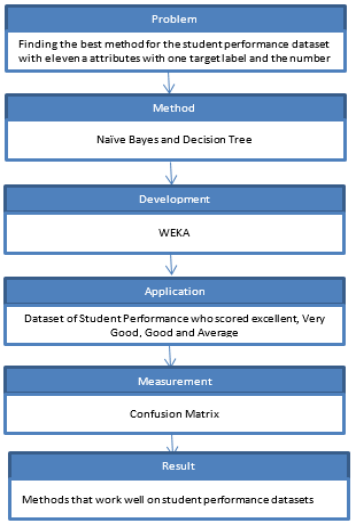
Previous research on the management of educational data mining has been carried out with various datasets and different attributes. Among them was carried out by Amalia in 2014 using a student dataset at a university with nine attributes and one label, namely graduating on time and late using the C4.5 algorithm (Amalia, 2014). In 2015, research on educational data mining for assessment assessments uses the data mining classification method (Patil, 2015). In 2016 a study was conducted by comparing several data mining methods, and the method with the highest value in the dataset was the decision tree (Kavipriya, 2016). In 2017 educational data mining research was conducted for processing student performance data at the tertiary level by producing the best working method, namely the decision tree (Khasanah & Harwati, 2017). In 2019, educational data mining was processed with four data mining methods, and the method that produced the highest accuracy value was nave Bayes (Yaacob, Nasir, Yaacob, & Sobri, 2019). In 2020, the dataset for student performance is processed with 16 attributes and one label, namely the target class with the excellent, good, Above average, average and Fail categories. From the research results, the highest accuracy value was the decision tree method (Walia N. , Kumar, Nayar, & Metha, 2020). Research in 2020 was also carried out by Amalia using the 4.5 algorithms, which improved its performance with the optimized selection method with a student performance dataset of 21 attributes and one label producing which attributes affect the dataset owned (Amalia, Yunita, Puspita, & Lestari, 2020) . In 2021, research on student performance data processing used five data mining methods, and the best method was the decision tree (Lynn & Emanuel, 2020).

In this study, a student performance dataset with nine attributes and one target label will be used which has four categories, namely Excellent, Very good, Good, and average, where students who are included in the excellent performing students are students who are labeled excellent, very good and good, meanwhile for students who are included in students who perform poorly, namely students in the average category. In this study, we will use two data mining methods that have been proven in previous studies to produce the best accuracy values for student performance datasets, namely nave Bayes and decision trees. The tools used are WEKA.

WEKA is one of the tools used to manage data mining which is known to be able to produce and visualize data mining results well and can be implemented in almost all machine learning methods (Hussain, Dahan, Ba-Alwi, & Ribata, 2018)

MATERIALS AND METHODS

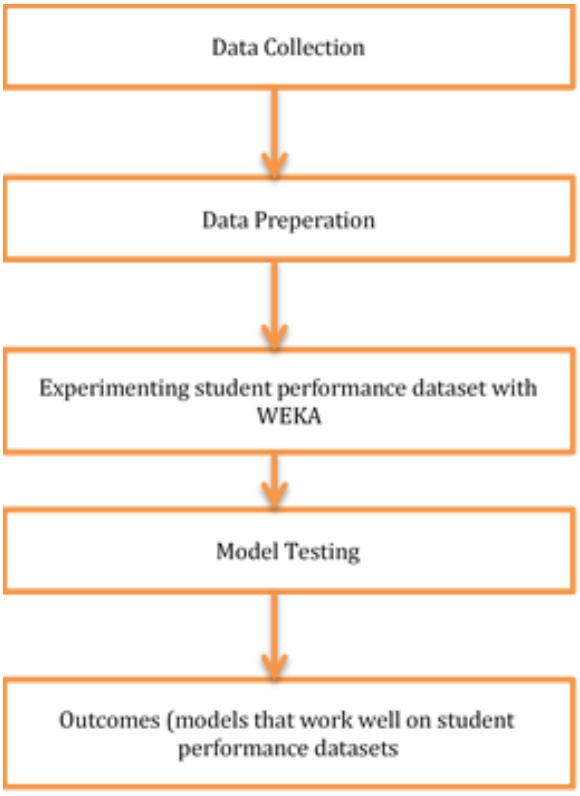
In this research, a framework of thought is made. The following is a framework of thought carried out:



Source: (Amalia, Puspita, & Lestari, 2021)

Figure 1 Thinking Framework

Figure 1 is the framework used in this research. Find solutions to problems in the dataset used, namely the student performance dataset of 666 records, consisting of 9 attributes and one label. The method used in this research is the naive Bayes method and the decision tree. The two data mining methods are data mining methods that have excellent performance for processing educational data mining. The development tool used is WEKA to apply the student performance dataset with the Excellent classification, Very Good, Good, and Average. The measurement of the data mining method used is nave Bayes, and the decision tree uses a confusion matrix. The results obtained are getting the best data mining method on student performance. In addition to creating a framework for this research. The research method carried out is presented in Figure 2 below:



Source: (Amalia, Puspita, & Lestari, 2021)

Figure 2 Research Stages

Figure 2 describes the steps taken in carrying out this research. The first step is data collection, and then the collected data is processed so that the dataset is more valid to use. The next step is to experiment with the existing dataset with the method used. Further testing of the model results from the experiment. The final step is to obtain the results of the model that works well on the student performance dataset

RESULTS AND DISCUSSION

Data collection

The data used in this study is student performance data with nine attributes and one label, which has four categories, namely Excellent, Very good, Good, and Average. Secondary data collection from dataset provider sites for research is the UCI repository web page. The following are the attributes used in processing student performance datasets:

Table 1 Attributes used

Attribute	Value
Gender	Male Female
Caste	General SC OBC ST
Coaching	NO WA OA
time	ONE TWO THREE FOUR FIVE SEVEN
Class_10	CBSE OTHER SEBA
Class_12	AHSEC CBSE OTHER
Medium	ASSAMESE ENGLISH OTHER
Class_10_Persentage	Excellent Very Good Good Average
Class_12_Persentage	Excellent Very Good Good Average
Father Occupation	Bank Official Business Collage Teacher Schoolteacher Cultivator Doctor Engineer Other
Mother Occupation	Bank Official Business Collage Teacher Schoolteacher Cultivator Doctor Engineer Other House Wife

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2714

Attribute	Value
Performance	Excellent
	Very Good
	Good
	Average

Source: (Hussain, Dahan, Ba-Alwi, & Ribata, 2018)

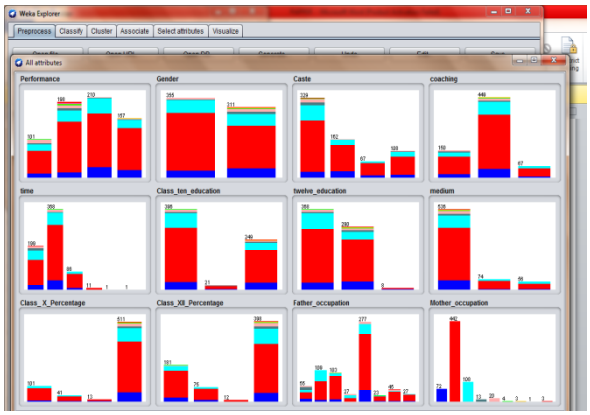
Table 2 presents the attributes used in the student performance dataset. The student performance dataset in this study has 11 attributes and one target label.

Initial data preparation

Preparation of initial data to get a dataset representing a dataset of student performance well. In preparing the initial data, three kinds of stages: Data validation to remove incomplete and duplicate records. Data transformation is the technique of changing and unifying data to represent the data better. The datasets obtained from the UC repository data provider site have gone through the above stages for immediate use in this study.

Experiment

In the experimental stage of the student performance dataset, data processing was carried out using 90 percent training data and 10 percent testing data using two data mining methods, namely the decision tree method and the Naive Bayes method.



Source: (Amalia, Puspita, & Lestari, 2021)

Figure 3 Number of attributes used

In Figure 3, the attributes used along with the number of values for these attributes are presented. There are 11 attributes with one target label. The performance attribute, which is the target label in the student performance dataset, has 101 records that have an Excellent score, 198 Very Good, 210 Good, and 157 Average. For the gender attribute, the number of records for the male value is 355, and the female is 311. For the Caste, the attribute has 329 records for General, 162 for OBC, 67 for SC, and 108 for ST. For the Coaching attribute, 150 records are worth NO, 449 are worth WA, and 67 are worth OA. In the time attribute, there are 199 records for the first value, 368 for the second value, 86 for the third value, 11 for the fourth value, 1 for the fifth value, and 1 for the seventh value. In the class_ten_education attribute, there are 396 records with SEBA values, 21 with OTHERS values , and 249 with CBSE values. In the twelve_education attribute, there are 368 records with AHSEC values, CBSE 290, and OTHERS 8. In the medium attribute, there are 536 records with ENGLISH values, 74 for OTHERS values , and 56 for ASSASEME. In the Class_X_percentage attribute, there are 101 records with a value of Vg, 41 for Good, 13 for Average, and 511 for Excellent values. In the Class _XII_percentage attribute, there are 181 records for Vg, 75 records for the value of Good, 12 records for the Average, and 398 records for the value of Excellent. In the father occupation attribute, there are 55 records for Doctor scores, 109 records for school teachers, 103 records for business, 27 records for College teacher scores, 277 records for Others scores, 23 records for Bank Official scores, and 45 records for Engineer scores and 27 records. For cultivators. In the mother occupation attribute, there are 72 records for the value of Others, 442 records for the value of housewife, 108 records for the value of school teacher, 13 records for the value of doctor, four records for the value of bank official, one record for the value of cultivator and three records for the value of engineer. And for the attribute that is the target label, namely performance, there are 101 records for the Excellent value, 198 records for the Vg value, 210 records for the excellent value, and 157 records for the Average value.

Table 2 Experimental Results with the Decision Tree method

Tp Rate	FP Rate	ROC	Class
0.465	0.051	0.861	Excellent
0.646	0.186	0.826	Vg
0.581	0.208	0.780	Good
0.854	0.047	0.969	Average

Source: (Amalia, Puspita, & Lestari, 2021)

Table 3 Experimental Results with the Naïve Bayes Method

Tp Rate	FP Rate	ROC	Class
0.376	0.057	0.816	Excellent
0.500	0.184	0.763	Vg
0.605	0.296	0.728	Good
0.771	0.055	0.959	Average

Source: (Amalia, Puspita, & Lestari, 2021)

In table 2, the experimental results on the student performance dataset using the decision tree P method produce TP, namely True Positive. TP is the number of positive cases correctly classified as

positive case value. Meanwhile, FP is False Positive, which is the number of positive values classified incorrectly as positive cases. In table 2, using the decision tree method, there are four Tp rates and Fp rates for the four values in the target label, namely performance. Meanwhile, in table 3, the resulting values for Tp and Fp are presented in the trials of the Naive Bayes method. From the experimental results obtained, the ROC value. ROC is a manifestation of the accuracy value obtained so that the ROC value and the accuracy value obtained will be directly proportional if a method works well for a dataset, but if the ROC value is inversely proportional to the accuracy value, then the method cannot work well for a data set. The ROC value of each target label value, namely performance, is presented in table 2 for the decision tree method and table 3 for the Naive Bayes method.

Model Testing

After the experimental results model, test the model using the confusion matrix to assess the accuracy value obtained from the two data mining models tested on the student performance dataset presented in table 4.

Table 4 Confusion Matrix Results

Methods	Value Accuracy
Decision Tree	94.3132 %
Naïve Bayes	84.8052 %

Source: (Amalia, Puspita, & Lestari, 2021)

CONCLUSION

Educational data mining analysis is an important thing that every educational institution must do. The study results concluded that the student performance dataset has 666 records with 11 attributes with one target label, namely performance with four values, namely Excellent, Vg, Good and Average. This study carried out experiments with the decision tree and nave Bayes methods. Decision tree and naive Bayes are two data mining methods that produce high accuracy scores on various student performances dataset; this has been proven in several previous studies. For the student performance dataset used in the study, the accuracy value for the decision tree method was 94.3132%, and the accuracy value obtained from the naive Bayes method was 84 8052%. The conclusion is that the data mining method, namely the decision tree, works better than the Naive Bayes method.

REFERENCE

Amalia, H. (2014). Prediction of Student Graduation Using Algoritma C4.5. *International Seminar on Scientific Issues and Trens (ISSIT)*, A31-A36.

Amalia, H., Puspita, A., & Lestari, A. F. (2021). *Research Report.*

Amalia, H., Yunita, Puspita, A., & Lestari, A. F. (2020). Student Performance Analysis Using Algoritma C4.5 Optimize Selection. *Pilar Nusa Mandiri*, 10-18.

Asraf, A., Anwer, S., & Khan, M. (2018). A Comparative study of Predicting Student's Performance of Data Mining techniques. *American Scientist Research Journal for Engineering, Technology, and Sciences (ASRJETS), 44*(1), 122-136.

Brew, E. A., Nketiah, B., & Korentang, R. (2021). A Literature Review of Academic Performance, an Insight into Factors and Their Influences on Academic Outcomes of Students at Senior High School. *OALIB*, 1-14.

Fajar, S., Hussain, M., Sarwar, H., Afzal, M., & Gilani, S. A. (2019). Factors Affecting Academic Performance of Undergraduate Nursing Students. *International Journal of Social of Undergraduate Nursing Students*, 7-16.

Hussain, S., Dahan, N. A., Ba-Alwi, F. M., & Ribata, N. (2018). Educational Data Mining and Analysis of Student's Academic Performance Using WEKA. *Indonesia Journal of Electrical Engineering and Computer Science*, 447-459.

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*, 101-105.

Khasanah, A. U., & Harwati. (2017). A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. *Material Science and Engineering (MOIME)*, 1-7.

Lynn, N. D., & Emanuel, A. W. (2021). Using Data Mining Techniques to Predict Students' Performance. a Review. *Materials Science and Engineering* , 1-9.

Lynn, N., & Emanuel, A. W. (2020). Using Data Mining Techniques to Predict Students' Performance. a Review. *Confeerrence Series: Materials Science and Engineering*, 1-9.

Masangu, L., Jadhav, A., & Ajoodha, R. (2021). Predicting Student Academic Performance Using Data Mining Techniques. *Advances in Science, Technology and Engineering System Journal (ASTES)* , 153-163.

Olufemi, O. T., Adederin, A. A., & Oyediran, W. O. (2018). Factors Affecting Students' Academic Performance in Collages of

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 18, No. 1 March 2022 | DOI: 10.33480/pilar.v18i1.2714

Education in Southwest Nigeria. *British Journal of Education*, 43-56.

Patil, P. (2015). A STUDY OF STUDENT'S ACADEMIC PERFORMANCE USING DATA MINING TECHNIQUES. *INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS*, 59-63.

Sumitha, R., & Vinothkumar, E. S. (2016). Prediction of Students Outcome Using Data Mining Techniques. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, 132-139.

Walia, N., & Kumar, M. (2020). Students' Academic Performance Prediction in Academic Using Data Mining. *International Conference on Intelligent Communication and Computational Research (ICICCR-2020)*, 1-5.

Walia, N., Kumar, M., Nayar, N., & Metha, G. (2020). Student's Academic Performance Prediction in Academic using Data Mining Techniques. *Inernational Conference in Intelligent Prediction in Academic Using Data Mining Techniques*, 1-5.

Yaacob, W. F., Nasir, S. A., Yaacob, W. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indonesian Journal of Electrical Engineering and Computer Science*, 1584-1592.