Phi-TERA : Journal Physics in Teaching, Education Research and Application
Vol. 1, No. 1 [Juli 2024], pp. 61-68

The Implementation of Educational Data Mining (EDM) to Classify Determining Factors of Physics Learning Outcomes of Students in Online Learning in Palopo City

Rinto Suppa1*
Program Studi Teknik Informatika, Universitas Andi Djemma, Indonesia
Email: rintosuppa@gmail.com

Surahman Indra 2
Program Studi Teknik Informatika, Universitas Andi Djemma, Indonesia
Email: ammanksinyo147@gmail.com

Solmin Paembonan 3
Program Studi Teknik Informatika, Universitas Andi Djemma, Indonesia
Email: solmin.p@unanda.ac.id

Abstrak. The objective of this research was to apply data mining classification rules in the field of education to identify factors that determined the learning outcomes of Physics students during online learning in the Covid-19 pandemic. The data collection method involved distributing a questionnaire consisting of 21 data entries to be filled out by the students. Rapid Miner 10.2 software was used to analyze the questionnaire results. The classification method employed was Decision Tree C.4.5. The performance results of the algorithm showed an accuracy of 83.37%, falling into the category of good classification. According to the Decision Tree classification results table, the most determining factors or features for students' learning outcomes were the rankings, where students with higher rankings also achieved higher physics scores.

Keywords: Decision Tree, Rapid Miner, accuracy

I. Introduction

The COVID-19 pandemic has left many problems, including in the field of education. This pandemic has necessitated radical policy changes. One concrete example is the implementation of online learning, which has been in place for approximately two years. The implementation of online learning has been facilitated by the significant development of video conference application platforms in the Era of the Fourth Industrial Revolution. In the process of learning physics, there are still challenges that teachers and students may face, such as conducting laboratory experiments. However, this can be addressed with the availability of virtual laboratories that are offered as open source or paid services on the internet.

In its implementation, online learning does not always negatively impact students. In fact, according to research conducted by Aulia Zulfa Shoumi[1], Antika Damayani[2], and Rachel Mia Lorenza Lumban Toruan[3], distance or online learning implemented by Ruang Guru has a positive impact on students. Similarly, Ifa Datul Mustafidah[4] concluded that students' performance improves if online learning is conducted according to the students' ability levels. This proves that certain factors contribute to the success or failure of online learning. Therefore, some students' physics learning outcomes might be better, worse, or sufficient during the COVID-19 pandemic online learning period, depending on the influencing factors.

The online learning method can be implemented in the future, regardless of the presence of another pandemic, as it is considered necessary to adapt to the increasingly digital era. There needs to be an evaluation of the learning that has been conducted to design or implement future methods or curricula. Therefore, knowledge is needed as a basis for policymakers, especially physics teachers, to determine how the learning opportunities currently applied can be effective and efficient in the future if implemented again.

Educational Data Mining (EDM) is the application of data mining methods in the field of education[5]. With EDM methods, teachers, government officials, and researchers/lecturers involved in education can generate information that evolves into new knowledge by utilizing and processing existing data[6]. This new knowledge can serve as a basis for policies aimed at improving the quality of education, one of which is students' learning outcomes[7]. The data mining process is very rational to be applied in this research [8][9][10] because the required data can be directly obtained from students who have experienced online learning by providing questionnaires to students affected by COVID-19 at SMA Negeri 1 and SMA Negeri 6 Palopo.

Based on the above background, the purpose of this research is to apply classification rules in data mining to classify the factors that influence the Physics learning outcomes of students in Palopo City during online learning in the COVID-19 pandemic.

II. Method

This research adapts the data mining research process proposed by Gorunescu, 2011[11], and the dataset to be collected contains features or attributes referring to the research by Paulo Cortez and Alice Silva, 2008, which are related to the profile, demographics, social aspects, school facilities and infrastructure, and methods related to the physics subject, as well as students' [7] extracurricular activities during the COVID-19 pandemic.

Time and Place of Research

The data collection in this research, which includes the profiles of students that will become the factors influencing their physics learning outcomes, was conducted at SMA Negeri 1 and SMA Negeri 6 in Palopo City in February 2023.

Research Procedure

1. Data Collection

In this stage, data is collected directly by visiting the respective schools to request the school's database that includes students' (learners') physics grades and their profiles. If the school does not have a database, questionnaires are distributed to students for them to fill out. Unlike other survey or research questionnaires, the statements (features or attributes) in a data mining questionnaire must be factual and concrete, not opinion-based or abstract.

The data used as features in this research are: 'School Origin,' 'Gender,' 'Number of Siblings,' 'Living With Whom,' 'Mother's Education,' 'Father's Education,' 'Mother's Occupation,' 'Father's Occupation,' 'Favorite Subject,' 'Attending Physics Courses,' 'Parental Guidance,' 'Extracurricular Activity Participation,' 'Internet Facilities at Home,' 'Internet Speed at Home,' 'Physics Learning Applications,' 'Use of Physics Simulations,' 'Virtual Practicum,' 'Learning Tools,' 'Use of Internet for What,' 'Internet Restrictions,' 'Ranking,' and 'Students' Physics Grades.'

2. Prepocessing Data

The collected data will then be compiled into a 'data set' in a single Excel file to facilitate data conversion when processing. In this research, Microsoft Excel is used to create the data set, and RapidMiner is used to process the data. Among the many features or attributes, the 'physics grades' of the students will be used as the label or class because the purpose of this research is to identify which variables or features influence the students' physics grades during online learning in the COVID-19 pandemic. Before the data mining process is conducted, it is essential to ensure that the data set does not contain any missing data or outliers.

3. Modelling

To gain insights from the existing data, it is first necessary to determine which function to use. Referring to the attributes, data types, and attribute values used in this research, the researcher employs Classification rules using a Tree model. The Decision Tree method is chosen for Classification rules because it fits the research needs very well. Additionally, the flexibility of the Decision Tree method makes it attractive, as it provides the advantage of visual recommendations, making the prediction procedure observable or, in other words, the resulting Tree model is easy to understand.

4. Performance Evaluation

The performance evaluation of the classification model is based on testing true and false objects[12]. In this research, the performance evaluation results can be seen from the accuracy value. The better the accuracy value of a classification model, the higher the confidence level in the displayed results[13].

5. Knowledge

The most important result of data mining is finding information. As Eko Prasetyo stated, data mining involves the activities of collecting and using historical data to discover patterns, regularities, and relationships within a dataset[14]. The information in the form of patterns and relationships between variables will become knowledge.

6. Research Flowchart

The research flowchart is as follows

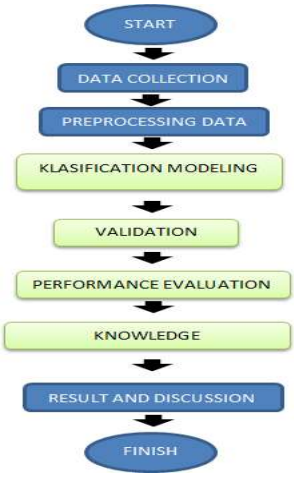

Figure 1. Research Flowchart

III. Result and Discussion

Result

1. Data Colletion

At this stage, the researcher collected data on the physics learning outcomes of students who participated in online learning during the Covid-19 pandemic from two high schools in Palopo City, namely, SMA Negeri 1 Palopo and SMA Negeri 6 Palopo with a total of 125 student data. The variables that are featured in this research are school origin (AS), gender (JK), number of siblings (JB), living together (TB), mother's education (PEN IBU), father's education (PEN AYAH), mother's job (PEK IBU), father's job (PEK AYAH), favorite lessons (Pel Suka), taking physics courses (Kur Fis), parental assistance (OT assistance), extracurricular activity (Active Extracurricular), internet facilities at home (Internet Facilities), internet speed at home (Internet Literacy Devices), physics learning applications (Learning Apps), the use of physics simulations (Simulation), virtual practicum (Virtual Practicum), learning tools (Pembel Tools), using the internet for ? (Using the Internet), internet restrictions (Access Restrictions), rankings (Rankings), and students' physics grades (Grades).

In this study, the feature of students' physics scores is used as a label or class because the *Decesion Tree method* includes *Supervisid Learning* where the goal is to identify and influence factors that affect students' physics learning outcomes in the form of student physics scores. To make it easier to classify them, the researcher grouped the students' physics scores into four groups, namely, 90-100 grades are grouped into A-, 85-89 are grouped into A-, 80-84 B+ and 79 and below are included in group B.

2. *Preprocessing Data*

At this stage, the researcher entered all the data in the form of *an Excel* file to make it easier to convert it in *the Rapid Miner software*. After completion, this Excel-type file is then imported into the *Rapid Miner* application by first setting the data type for each feature.

The data type can be seen in the figure below:

Name	Type	Missing	Statistics		
Label **Nilai**	Nominal	0	Least B (8)	Most A- (55)	Values A- (55), B+ (34), ...[2 more]
AS	Binominal	0	Negative 1	Positive 6	Values 1 (62), 6 (62)
JK	Binominal	0	Negative L	Positive P	Values P (69), L (55)
JB	Integer	0	Min 1	Max 8	Average 3.629
TB	Binominal	0	Negative OT	Positive K	Values OT (92), K (32)
PEN IBU	Nominal	0	Least smp (1)	Most SMA (86)	Values SMA (86), S1 (20), ...[5 more]
PEN AYAH	Nominal	0	Least smp (1)	Most SMA (80)	Values SMA (80), S1 (15), ...[8 more]
PEK IBU	Nominal	0	Least Pns (1)	Most IRT (91)	Values IRT (91), Pegawai (29), ...[2 more]
PEK AYAH	Nominal	0	Least wirausaha (1)	Most Pegawai (54)	Values Pegawai (54), wiraswasta (47), ...[4 more]
PEL SUKA	Nominal	0	Least sejarah (1)	Most Biologi (30)	Values Biologi (30), Penjas (26), ...[13 more]
Kur Fis	Binominal	0	Negative Ya	Positive Tidak	Values Tidak (117), Ya (7)
Pendampingan OT	Binominal	0	Negative Ya	Positive Tidak	Values Tidak (89), Ya (35)
Aktif Ekskul	Nominal	0	Least Kadang (39)	Most Ya (44)	Values Ya (44), Tidak (41), ...[1 more]
Fas Internet	Binominal	0	Negative indihome	Positive paket data	Values indihome (76), paket data (48)
Kec Internet	Nominal	0	Least lambat (4)	Most sedang (82)	Values sedang (82), cepat (38), ...[1 more]
App Pembelajaran	Nominal	0	Least WA (16)	Most GC (68)	Values GC (68), Zoom (40), ...[1 more]
Simulasi	Nominal	0	Least tidak (25)	Most kadang (66)	Values kadang (66), ya (33), ...[1 more]
Praktikum Virtual	Nominal	0	Least ya (19)	Most kadang (62)	Values kadang (62), tidak (43), ...[1 more]
Alat Pembel	Binominal	0	Negative smartphone	Positive laptop	Values smartphone (100), laptop (24)
Menggunakan Internet	Nominal	0	Least main game (9)	Most sosial media (73)	Values sosial media (73), belajar (25), ...[2 more]
Pembatasan Akses	Binominal	0	Negative tidak	Positive ya	Values tidak (68), ya (56)
Rangking	Integer	0	Min 1	Max 35	Average 13.694

Figure 2. Data Sets to be processed

3. Modeling

This study applies *the Decision Tree* model by using the C4.5 *Decision Tree* algorithm . In *Rapid Miner*, the visual programming can be seen in the image below:

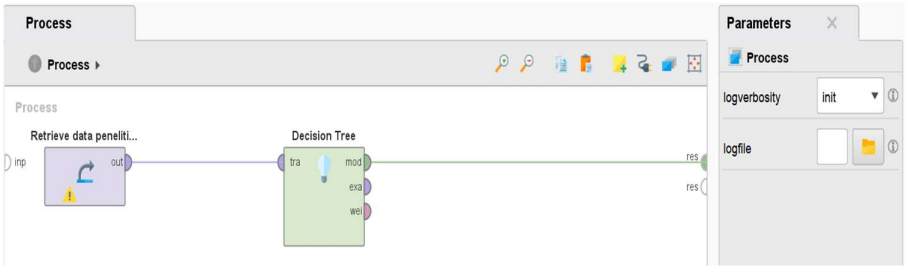

Figure 3. Modeling with Decision Tree

The results in the form of a decision tree from the visual programming above can be seen in figure 4:

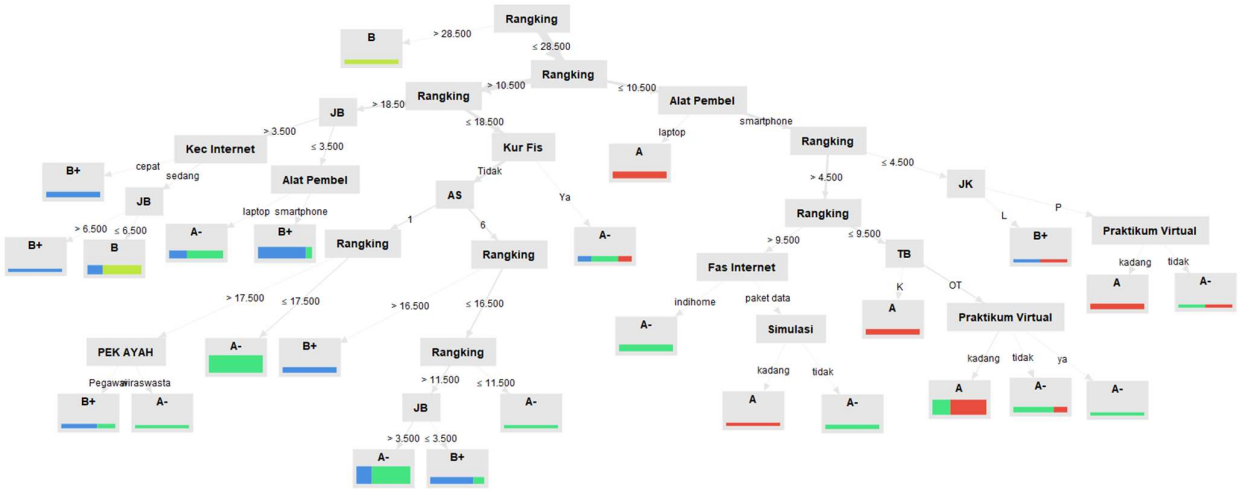

Figure 4. Decision Tree

From the image above, it is clear that the ranking feature is the root node. This proves that the highest ratio gain is in the ranking feature.

4. *Performance/Kinerja*

At this stage, it is very necessary to carry out a performance test or performance test of the results of the determined algorithm to determine the accuracy of the model used.

The visual programming on *Rapid Miner* to determine the *performance* of the *Decesion Tree* C4.5 classification model used can be seen in the image below:

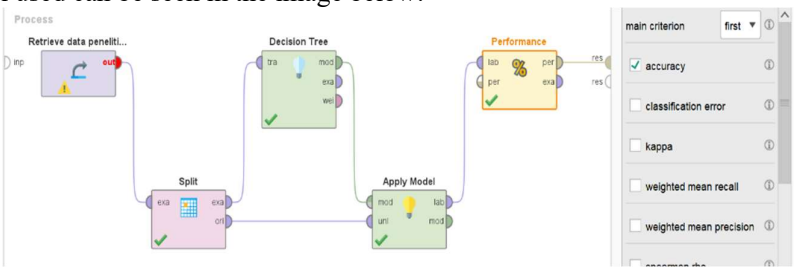

Figure 5. Performance measurement of *Decision Tree* C 4.5 classification model

From the visual programming above, the performance results were obtained in the form of an accuracy value of 83.87%.

accuracy: 83.87%

	true B+	true A-	true B	true A	class precision
pred. B+	25	3	0	1	86.21%
pred. A-	7	48	0	3	82.76%
pred. B	2	0	8	0	80.00%
pred. A	0	4	0	23	85.19%
class recall	73.53%	87.27%	100.00%	85.19%	

Figure 1. Accuracy results with *the Confusion Matrix method*

Discussion

1. Decision Tree Results

Based on the results of the decision tree in figure 4.3 above, it is clear that the ranking feature has a very important factor in determining the physics learning outcomes of high school students in Palopo City during online learning.

In the results of the node division in the figure, it can be seen that students or students who rank 28 and above have the lowest physics score, namely B. while for rank 28 and below, they have diverse physics scores with the first condition that if they are in the top 10 their scores between A and A- only, and there is only one person who gets B+. The difference is that if the student uses a learning tool, namely a laptop, it is likely that the score will be better than that of him or her who uses *a smartphone* during online learning. This is evidenced by the results of the modeling carried out that 100% of students who are in the top ten who use laptops during learning get an A grade.

For rankings above the top 10, they are separated again by ranking nodes, which are between 10 to 18 and between 18 to 28. If a student is ranked between 10 and 18 the defining feature or factor is whether he is taking a physics course or not if he is taking a physics course then there is a great chance that he gets an A- grade and if the student does not take a physics course the grade can be A- or B+ depending on the origin of the school, the student or student whose rank is between 18 to 28 depends on the number of siblings if he has a number of siblings less than 4 his grade between A- or B+ depends on the learning tool he uses, if he uses a laptop he is most likely to get an A- grade and if he uses *a smartphone* he is most likely to get a B+ grade. If the student has more than or equal to four, the grade is in the range of B or B+ depending on the internet speed in his home. If the internet speed at home is relatively fast, then the value is B+ if it is slow, the value is between B or B+.

2. Performance Model Decision Tree C4.5

The performance results of the Decision Tree model can be seen from the accuracy value using the *Confusion Matrix* method. Based on Figure 5, it can be seen that the total accuracy of this model is 83.37%. The accuracy value is included in the category of good classification. This is in line with what Gurunescu said, that the accuracy value of a classification model between 80% and 90% is included in the Good Classification category[11]. With this good accuracy value, it can be ensured that the data set used is suitable for the classification model that is built and can be trusted to be a reference for obtaining knowledge from the collected data[15], in this case the factors that affect the learning outcomes of students in physics during online learning carried out in Palopo City are well validated and can be accounted for.

IV. Conclusion

Based on the results obtained above, it can be concluded that:

1. The data set used is suitable for the *Decision Tree* C.4.5 method to classify the determinants of students' physics learning outcomes in Palopo City because the accuracy obtained is 83.37% which is included in the good classification category.
2. The most influential determinant of students' physics learning outcomes is ranking, where if the ranking is good, then the physics score is also good and vice versa.

3. Based on the results of physics learning, the use of a laptop as a physics learning tool during online learning is better than using *a smartphone*.

References

[1] Shoumi, A. Z. 2019, October. Peran multimedia dalam pendidikan pada aplikasi ruang guru. In *Prosiding seminar nasional cendekiawan* (pp. 2-23).

[2] Damayani, A., & Harsono, S. U. 2021. *Penggunaan Aplikasi Ruangguru Sebagai Media Pembelajaran Online Bagi Siswa Unggul SMAN 1 Baturetno* (Skripsi, Universitas Muhammadiyah Surakarta).

[3] Lumban Toruan, R. M. L. 2021. Efektivitas Aplikasi Ruang Guru sebagai Medium Komunikasi dalam Kegiatan Bimbingan Belajar Daring di Kalangan Siswa SMA di Kota Medan.

[4] Mustafidah, I. D., Mutohar, P. M., & Tanzeh, A. 2022. Prestasi Belajar Siswa di Era Covid 19: Analisis Pelaksanaan Pembelajaran Daring, Motivasi Belajar dan Lingkungan Belajar di Madrasah Ibtidaiyah. *edukatif: jurnal ilmu pendidikan*, *4*(1), 355-363

[5] Romero, C., & Ventura, S. 2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

[6] Romero, C., & Ventura, S. 2010. Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

[7] Cortez, P., & Silva, A. M. G. 2008. Using data mining to predict secondary school student performance.

[8] Hendradi, P. 2015. Improve of use educational data mining (edm) for the implementation of competency based curriculum (cbc).

[9] Galih, G. 2019. Data Mining di Bidang Pendidikan untuk Analisa Prediksi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi pada STMIK Jabar. *J. Teknol. Sist. Inf. dan Apl*, *2*(1), 23.

[10] Ahmed, A. B. E. D., & Elaraby, I. S. 2014. Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, *2*(2), 43-47.

[11] Gorunescu, F. 2011. *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media.

[12] Wu, X., & Kumar, V. (Eds.). 2009. *The top ten algorithms in data mining*. CRC press.

[13] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. 2005. Practical machine learning tools and techniques. In *DATA MINING* (Vol. 2, p. 4).

[14] Prasetyo, Eko. 2014. Data Mining – Mengolah Data Menjadi Informasi menggunakan Matlab. Yogyakarta: Penerbit Andi.

[15] M. Manvar and M. Rao. Predicting students performance in higher education: A Data Mining Approach. Ijser.Org, vol. 5, no. 2, pp. 1024–1027, 2014.