Application of Decision Tree C4.5 in Predicting the Best Predicate in Madrasah Ta'hiliyah Ibrahimy

Ahmad Huday[a1], Zaehol Fatah[a2]
Information Systems Study Program, Faculty of Science & Technology, Ibrahimy University, Indonesia
E-mail: [1]ahmadhuday94887@gmail.com, [2]zaeholfatah@gmail.com

Abstrak

Untuk meningkatkan proses evaluasi dalam menilai perkembangan siswa, prediksi terhadap predikat terbaik menjadi langkah penting dalam upaya meningkatkan kualitas pendidikan. Dengan mengetahui siswa yang berpotensi meraih predikat terbaik, institusi pendidikan dapat lebih fokus dalam merancang metode pengajaran yang tepat dan strategi pembelajaran yang terarah. Langkah ini menjadi kunci untuk memastikan bahwa proses pembelajaran berjalan sesuai dengan tujuan institusi, yaitu menghasilkan siswa yang unggul dalam pengetahuan dan keterampilan. Dalam penelitian ini, kami memanfaatkan algoritma C4.5, yang merupakan salah satu metode pohon keputusan paling dikenal dalam data mining, untuk memprediksi predikat siswa. Algoritma C4.5 terkenal akan kemampuannya dalam mengklasifikasi data serta menemukan pola-pola tersembunyi di dalam dataset. Dengan pendekatan ini, kami bertujuan untuk menganalisis faktor-faktor yang memengaruhi keberhasilan siswa sekaligus memberikan wawasan yang dapat dimanfaatkan oleh pendidik dan pengelola sekolah. Penelitian ini dilakukan pada siswa di Madrasah Ta'hiliyah Ibrahimy, di mana algoritma pohon keputusan diterapkan untuk memprediksi predikat terbaik berdasarkan data akademik historis. Dari eksperimen ini, diperoleh tiga aturan atau pola yang dapat digunakan untuk memprediksi predikat siswa, dengan tingkat akurasi sebesar 74,17%. Hasil ini menunjukkan potensi besar dari pendekatan berbasis data dalam mendukung pengambilan keputusan akademik serta memberikan arah yang lebih jelas dalam merancang intervensi guna meningkatkan kinerja siswa di masa depan.

Kata kunci: *C4.5, Pohon Keputusan, Prediksi, Predikat Rapid Maner*

Abstract

To improve the evaluation process in assessing student progress, predicting the best grades plays a crucial role in enhancing the quality of education. By identifying the top-performing students, educational institutions can refine their teaching methods and create targeted strategies to foster better learning outcomes. This step is vital for ensuring that the learning process aligns with the institution's goals to produce highly skilled and knowledgeable students. In this research, we focused on utilizing the C4.5 algorithm, a widely recognized decision tree method in data mining, to predict student achievements. The C4.5 algorithm is known for its ability to classify and uncover hidden patterns within datasets, making it a powerful tool for educational data analysis. Through this approach, we aim to analyze the factors influencing student success and provide actionable insights for educators and administrators. The study was conducted on students from Madrasah Ta'hiliyah Ibrahimy, where we applied the decision tree algorithm to predict the best grades based on historical academic data. The experiment resulted in three distinct rules or patterns derived from the data, with an overall accuracy of 74.17%. These findings demonstrate the potential of data-driven approaches in supporting academic decision-making and guiding future interventions to further enhance student performance.

Keywords: *, C4.5, Decision Tree, Grade, Prediction, Rapid Maner*

1. Introduction

School is the most important educational medium in Indonesia in developing society to be able to have a better life. In Indonesia, there are many schools spread throughout Indonesia

to support an even better level of community education.[1] Education is indispensable in the development of science. Science is useful for logical values, ethical values, and aesthetic values found in humans themselves.[2]

Madrasah Ta'hiliyah Ibrahimy is an Islamic educational institution located at the Salafiyah Syafi'iyah Islamic Boarding School, Sukorejo, Situbondo, East Java. In general, Madrasah Ta'hiliyah Ibrahimy aims to produce a generation that is insightful both in religious science, with a strong moral and ethical foundation according to Islamic teachings. To realize this, Madrasah Ta'hiliyah Ibrahimy intensely conducts exams for students to know the development of students' knowledge or skills, besides that it can also be used as evaluation material for schools or teachers. Evaluation is the activity of collecting the widest and deepest data related to student capabilities in order to find out the causes and consequences and learning outcomes of students that can encourage and develop student learning abilities.[3]

Therefore, at Madrasah Ta'hiliyah Ibrahimy Sukorejo by conducting an exam is a very important thing in knowing the development of students. From the exam, the students get a predicate based on the size of the exam score. The goal is to find out the extent of mastery of knowledge by students. The predicate of success is the actual ability of a person in the form of mastery of knowledge, attitudes, and skills to achieve the final goal of the learning process.[4]

To find out the prediction of student rankings, it is by predicting using the C4.5 algorithm. The C4.5 data mining algorithm is one of the algorithms used to classify or segment or group and is predictive. Classification is one of the processes in data mining that aims to find valuable patterns from relatively large to very large data. Data Mining is a process of automatically searching for useful information in a large data storage area.[5]

The C4.5 data mining algorithm is one of the algorithms of the decision tree.[6] Decisions tree is a decision-making method that uses a tree structure to describe and analyze the consequences of various decisions, this method can be used in data classification or regression, where data is divided into groups based on certain conditions to estimate the value or target category.[7]

This study focuses on the problem of determining students with the best academic achievements using an appropriate and measurable method. The objective of this research is to identify students who achieve the highest grades based on exam score data collected from Madrasah Ta'hiliyah Ibrahimy. Through this approach, the findings are expected to serve as valuable evaluation material for assessing individual student progress and providing relevant insights to improve the quality of the teaching and learning process within the institution.

Furthermore, this study aims to offer a more comprehensive understanding of students' positions in the context of academic competition or overall evaluation. Thus, the results not only contribute to understanding individual student success but also serve as a strategic reference for the school in designing educational policies that are more effective and oriented toward optimal academic achievement.

2. Research Methods / Proposed Methods

Methodology is a theoretical framework used by the author to analyze, work on/overcome the problems faced. Theoretical frameworks or scientific frameworks are scientific methods that will be applied in the implementation of tasks.[8]

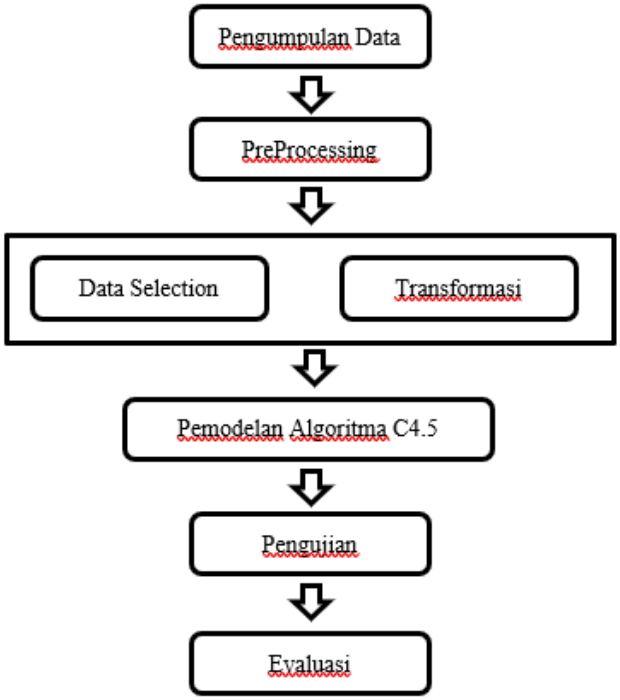


Figure 1. Research Methodology.

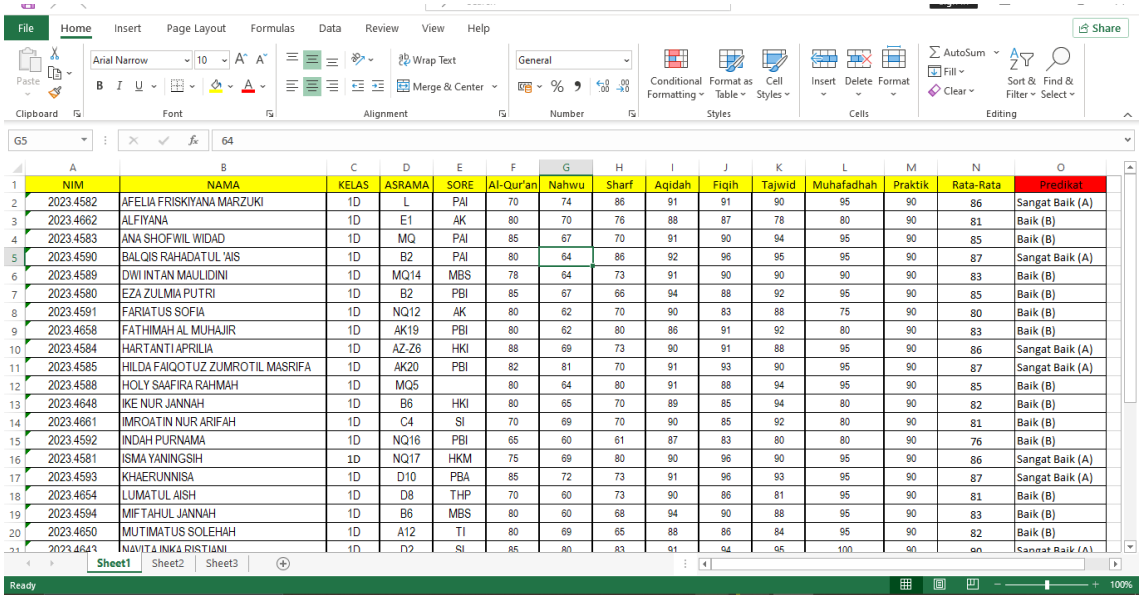
The image illustrates the steps for building a decision tree model using the C4.5 algorithm. The process begins with data collection, where all necessary information is gathered from various sources. Once the data is collected, it undergoes preprocessing, during which the data is cleaned to remove errors, such as missing or irrelevant values, to ensure it is ready for use.

Next, the processed data moves to the data selection and transformation stage. At this step, only important and relevant data is selected, and the data is transformed into a suitable format for the algorithm to process. Once the data is prepared, the model is built using the C4.5 algorithm, which creates a decision tree based on the processed data.

After the decision tree model is built, it goes through testing using test data to evaluate whether the model can make accurate predictions. Finally, the results of the testing are assessed in the evaluation stage, where the model's performance is measured using various metrics, such as accuracy. This evaluation determines whether the model is ready for use or requires further improvement. These steps aim to create a reliable model that can assist in making accurate data-driven decisions.

3. Literature Study

The initial stage in this study is data collection. The data we get is data in the form of *Dataset* which is in the form of *spreedshet* at *Excel*. This is the data that we have obtained.



NIM	NAMA	KELAS	ASRAMA	SORE	Al-Qur'an	Nahwu	Sharf	Aqidah	Fiqih	Tajwid	Muhafadhah	Praktik	Rata-Rata	Prediket
2023.4582	AFELIA FRISKIYANA MARZUKI	1D	L	PAI	70	74	86	91	91	90	95	90	86	Sangat Baik (A)
2023.4662	ALFIYANA	1D	E1	AK	80	70	76	88	87	78	80	90	81	Baik (B)
2023.4583	ANA SHOFWIL WIDAD	1D	MQ	PAI	85	67	70	91	90	94	95	90	85	Baik (B)
2023.4590	BALQIS RAHADATUL 'AIS	1D	B2	PAI	80	64	86	92	96	95	95	90	87	Sangat Baik (A)
2023.4589	DWI INTAN MAULIDINI	1D	MQ14	MBS	78	64	73	91	90	90	90	90	83	Baik (B)
2023.4580	EZA ZULMIA PUTRI	1D	B2	PBI	85	67	66	94	88	92	95	90	85	Baik (B)
2023.4591	FARIATUS SOFIA	1D	NQ12	AK	80	62	70	90	83	88	75	90	80	Baik (B)
2023.4658	FATHIMAH AL MUHAJIR	1D	AK19	PBI	80	62	80	86	91	92	80	90	83	Baik (B)
2023.4584	HARTANTI APRILIA	1D	AZ-Z6	HKI	88	69	73	90	91	88	95	90	86	Sangat Baik (A)
2023.4585	HILDA FAIQOTUZ ZUMROTIL MASRIFA	1D	AK20	PBI	82	81	70	91	93	90	95	90	87	Sangat Baik (A)
2023.4588	HOLY SAAFIRA RAHMAH	1D	MQ5		80	64	80	91	88	94	95	90	85	Baik (B)
2023.4648	IKE NUR JANNAH	1D	B6	HKI	80	65	70	89	85	94	80	90	82	Baik (B)
2023.4661	IMROATIN NUR ARIFAH	1D	C4	SI	70	69	70	90	85	92	80	90	81	Baik (B)
2023.4592	INDAH PURNAMA	1D	NQ16	PBI	65	60	61	87	83	80	80	90	76	Baik (B)
2023.4581	ISMA YANINGSIH	1D	NQ17	HKM	75	69	80	90	96	90	95	90	86	Sangat Baik (A)
2023.4593	KHAERUNNISA	1D	D10	PBA	85	72	73	91	96	93	95	90	87	Sangat Baik (A)
2023.4654	LUMATUL AISH	1D	D8	THP	70	60	73	90	86	81	95	90	81	Baik (B)
2023.4594	MIFTAHUL JANNAH	1D	B6	MBS	80	60	68	94	90	88	95	90	83	Baik (B)
2023.4650	MUTIMATUS SOLEHAH	1D	A12	TI	80	69	65	88	86	84	95	90	82	Baik (B)
2023.4643	NAVITA INKA RISTIANI	1D	D2	SI	85	80	83	91	94	95	100	90	90	Sangat Baik (A)

Figure 2. Student Grade Data

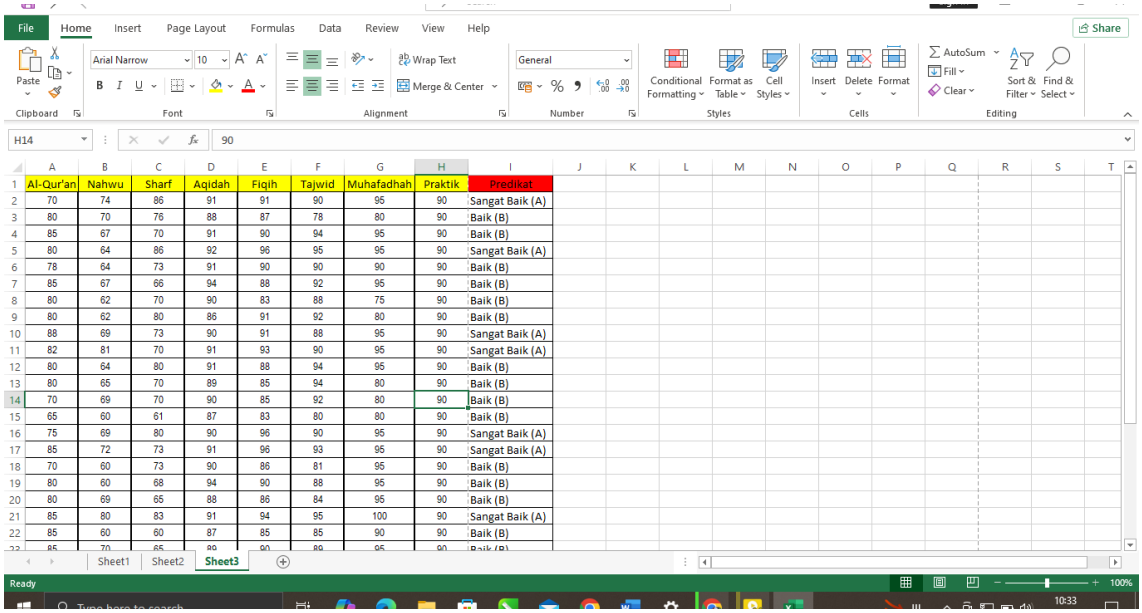
The displayed data is a student dataset containing comprehensive information such as NIM, student name, class, dormitory, afternoon program (main subject), scores from various subjects (such as Al-Qur'an, Nahwu, Sharaf, Aqidah, Fiqh, Tajwid, Muhafadhah, and Practice), average scores, and student performance grades. This data will be processed to analyze student performance based on the obtained scores. The processing involves several stages, starting from preprocessing to ensure the data is clean and complete, data selection to extract relevant attributes, and statistical processing such as calculating overall average scores or ranking students. Additionally, this data can be used for modeling, such as classification or prediction using specific algorithms like decision trees. The purpose of this processing is to gain deeper insights into students' academic performance and the factors influencing it.

3.1. Preprocessing

Text pre-processing is a series of important steps to clean and prepare text data before further analysis.[9] Before the data mining process is carried out, the researcher conducts a data preprocessing process, which is a stage in the data mining process. Before the dataset is processed to produce the expected output.[4]. The following are the stages *prepocessing.*

a. Prepoccessing

This initial stage is an important step to ensure that the data used in the KDD process is *(Knowledge Discovery in Database*) is accurate and quality data.[10] From the data that has been obtained, then the data will be filtered according to the selection criteria that we have set. Eliminating duplicate data and unimportant data is something that needs to be done to maintain the consistency of the data that will be used in the analysis.



	A	B	C	D	E	F	G	H	I
1	Al-Qur'an	Nahwu	Sharf	Aqidah	Fiqih	Tajwid	Muhafadhah	Praktik	Predikat
2	70	74	86	91	91	90	95	90	Sangat Baik (A)
3	80	70	76	88	87	78	80	90	Baik (B)
4	85	67	70	91	90	94	95	90	Baik (B)
5	80	64	86	92	96	95	95	90	Sangat Baik (A)
6	78	64	73	91	90	90	90	90	Baik (B)
7	85	67	66	94	88	92	95	90	Baik (B)
8	80	62	70	90	83	88	75	90	Baik (B)
9	80	62	80	86	91	92	80	90	Baik (B)
10	88	69	73	90	91	88	95	90	Sangat Baik (A)
11	82	81	70	91	93	90	95	90	Sangat Baik (A)
12	80	64	80	91	88	94	95	90	Baik (B)
13	80	65	70	89	85	94	80	90	Baik (B)
14	70	69	70	90	85	92	80	90	Baik (B)
15	65	60	61	87	83	80	80	90	Baik (B)
16	75	69	80	90	96	90	95	90	Sangat Baik (A)
17	85	72	73	91	96	93	95	90	Sangat Baik (A)
18	70	60	73	90	86	81	95	90	Baik (B)
19	80	60	68	94	90	88	95	90	Baik (B)
20	80	69	65	88	86	84	95	90	Baik (B)
21	85	80	83	91	94	95	100	90	Sangat Baik (A)
22	85	60	60	87	85	85	90	90	Baik (B)

Figure 3. Dataset Images after preproccessing

b. Transformation

Transformation is the stage of connecting attributes/variables that will later be used to predict.[11] In this stage we will change the data in the form of categories to numerical. Below is the data that has been transformed.

Table 1. Value Transformation

It	Value	Predicate
1	>85	Excellent
2	>70	Good
3	>55	Enough
4	>40	Less
5	>0	Very Less

3.2. C4.5 Algorithm Modeling

The C4.5 algorithm is a well-known algorithm used to group data with numerical and categorical characteristics. The grouping process generates rules that can be used to predict the value of discrete typical attributes of a new record. The C4.5 algorithm is also an ID3 algorithm developed, which is designed to address lost data, to address continuous and truncated data.[12]

This technique consists of a collection of decision nodes, and connected by branches, moving down from the root node until it ends in a leaf node.[13]

To build a *decision tree*, we first need to select the attribute that we will use as *the root*, then create a branch for each value, then distribute the cases on the branch, go through the process of creating each branch until all *instances* on the branch have the same class. The highest gain value of the existing attribute will be used for the selection of *the root* attribute according to the Formula according to the Equation in calculating the gain.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} . \frac{|S_v|}{S} . Entroph\,y(S_v)$$

Where:

S : the dataset on the *current* node

A : the attribute that is being calculated

Values(A) : the possible set of values of *the A attribute*

Sv : a subset of *S* where the attribute *A* has *a value of v*
Entropy(S): measures the uncertainty in the group of daa *S* and is calculated as follows:

$$Entropy(S) = -\sum_{i=1}^{n} p_i . \log_2(p_i)$$

Where:
p_i : proportion of examples in class i in dataset *S*

3.3. Testing

The testing phase will later produce a decision tree. Decision trees are a well-known classification and prediction technique. Decision trees are capable of turning very large problems into decision trees with rules. With a decision tree, it is easy to identify the relationships between the factors that affect the problem and find a good solution by taking those factors into account.[14]

3.4. Evaluation

After producing accuracy, the model is evaluated against the existing prediction results. This was done to find out how much accuracy in the form of accuracy results from prediction results using the C4.5 algorithm using the Confusion matrix.
Confusion Matrix is a method whose use is to perform accurate calculations on concepts in data mining, evaluation using the confusion matrix method produces accuracy, precision and recall values.[15]

In this data, there are already several variables that are useful for finding out the provisions of the predicate criteria that will be obtained by students. Here we plan to categorize the predicate of student exam results with several parts, namely, very good, good, enough, less, very less.

4. Results and Discussion

From the data we obtained, we will group the exam scores of Madrasah Ta'hiliyah Ibrahimy students, totaling 32 students, based on the following predicate.

Table 2. Value Transformation

It	Predicate	Sum
1	Excellent	9
2	Good	23

Then we process or analyze the data with an algorithm or a C.45 method or commonly called *a decision tree* using rapid maner software as follows.

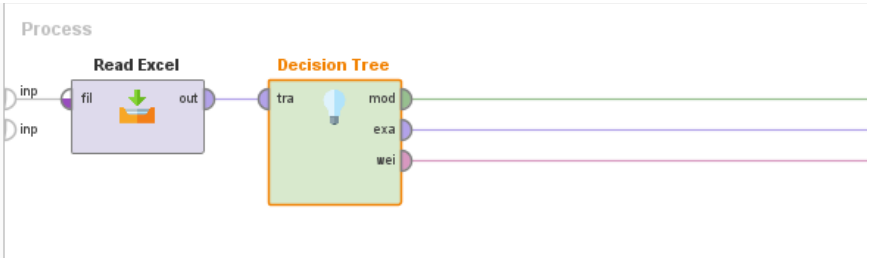


Figure 4. Decision Tree Experiment Process

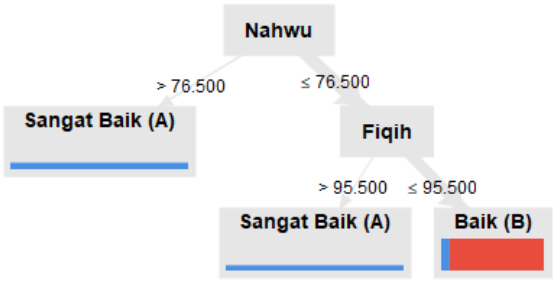


Figure 5. Results of the Decision Tree Model of the Best Predicate of Madrasah Ta'hiliyah Ibrahimy

From the image above we can see the decision tree model and the experimental process using *the decision tree*, the main purpose of this study is to conduct experiments with the C4.5 algorithm in order to produce patterns or rules and in this study 3 rules or patterns are obtained, namely:

1. If the nahwu > 76,500 then Very Good (A).
2. If nahwu ≤ 76,500 and fiqh > 95,500, then Very Good (A).
3. If nahwu ≤ 76,500 and fiqh ≤ 95,500, then Good (B).

Then the last stage is to evaluate the decision tree model against the prediction of the student's best predicate using the Confussion Matrix and ROC Curva.

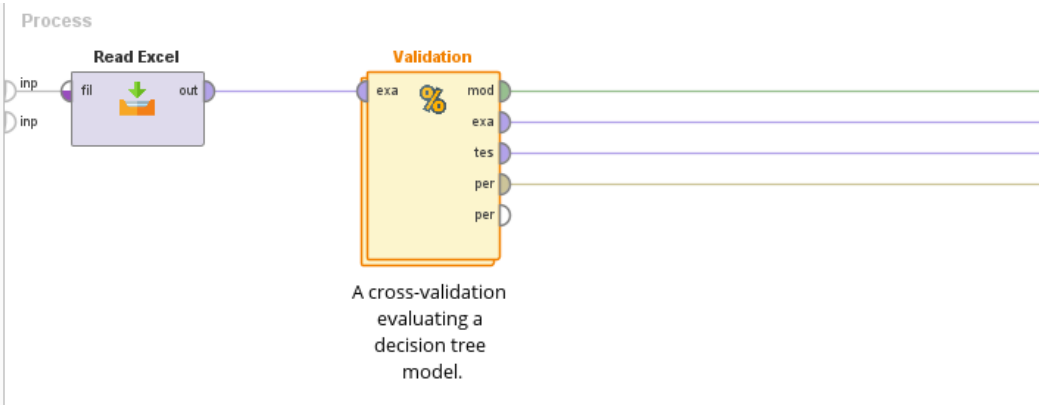


Figure 6. *Decision Tree* C4.5 Experiment Process

From the processing of *the cross validation* evaluation above, it produced an accuracy value of 74.17% and the following is the *confusion matrix*.

accuracy: 74.17% +/- 20.95% (micro average: 74.19%)

	true Sangat Baik (A)	true Baik (B)	class precision
pred. Sangat Baik (A)	4	3	57.14%
pred. Baik (B)	5	19	79.17%
class recall	44.44%	86.36%	

Gambar 7. Model *Confusion Matrix*

5. Conclusion

From the results of my research, which aims to determine the development of students in the learning process so that it can be used as evaluation material in the future so that it can be better, therefore I predict the best predicate for students at Madrasah Ta'hiliyah Ibrahimy in

2024 using the Decision Tree algorithm which produces a decision tree based on the student's exam score index and factors that can influence the best predicate for students, with an accuracy level of 74.17%.

Reference

[1] B. Q. Husaini, "Penerapan Algoritma Decision Tree C45 untuk Klasifikasi Penjurusan Siswa," vol. 9, no. 1, pp. 455–470, 2023.

[2] M. M. Prof. DR. H. A. Rusdiana and M. S. Dr. H. Aep Saepuloh, *SOSIOLOGI PENDIDIKAN: Menuju Pendidikan Unggul dan Kompetitif*. MDP, 2022. [Online]. Available: https://books.google.co.id/books?id=xUBpEAAAQBAJ

[3] J. Hamdayama, *Metodologi Pengajaran*. Bumi Aksara, 2022. [Online]. Available: https://books.google.co.id/books?id=ywFjEAAAQBAJ

[4] S. Sains, P. Kelulusan, M. Di, P. Kampar, A. Saputra, and T. A. Fitri, "Penerapan Data Mining Algortima C4 . 5 Dalam Memprediksi," 2023.

[5] M. S. Iskandar and Z. Fatah, "Gudang Jurnal Multidisiplin Ilmu Implementasi Metode Algoritma K-Means Clustering Untuk Menentukan Penerima Program Indonesia Pintar (PIP)," vol. 2, no. November, pp. 1–8, 2024.

[6] S. T. M. Yessy Asri, M. K. Dr. Dra. Dwina Kuswardani, S. T. M. C. S. Dr. Widya Nita Suliyanti, and S. T. Chrystyna Monica Tambunan, *ALGORITMA C4.5 : KLASIFIKASI TITIK DAN JENIS GANGGUAN PADA JARINGAN DISTRIBUSI PENYULANG*. Uwais Inspirasi Indonesia , 2023. [Online]. Available: https://books.google.co.id/books?id=5FzrEAAAQBAJ

[7] I. Nawawi and Z. Fatah, "Penerapan Decision Trees dalam Mendeteksi Pola Tidur Sehat Berdasarkan Kebiasaan Gaya Hidup," vol. 2, no. 4, pp. 34–41, 2024.

[8] S. Kasus, D. I. Smk, N. Lintau, and D. N. Yoliadi, "PERATURAN DISIPLIN SISWA," vol. 11, no. 01, pp. 50–62, 2022.

[9] A. Muzakir and U. Suriani, "Model Deteksi Berita Palsu Menggunakan Pendekatan Bidirectional Long Short-Term Memory (BiLSTM)," vol. 4, no. 2, pp. 93–105, 2023.

[10] U. Suriani, "Penerapan Data Mining untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunakan Algoritma," vol. 3, no. 2, pp. 55–66, 2023.

[11] P. Algoritma, C. Dalam, V. S. Ginting, and E. T. Luthfi, "KETERLAMBATAN PEMBAYARAN UANG SEKOLAH MENGGUNAKAN PYTHON," vol. 4, no. 1, 2020.

[12] C. Pada, U. Syarif, and H. Jakarta, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma," vol. 6, pp. 61–74, 2023.

[13] R. H. Pambudi and B. D. Setiawan, "Penerapan Algoritma C4 . 5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal," vol. 2, no. 7, pp. 2637–2643, 2018.

[14] R. Musfikar, H. Apriadinata, and B. Yusuf, "Aplikasi Prediksi Prestasi pada Siswa Menggunakan Algoritma C4 . 5 Student Achievement Prediction Application Using C4 . 5," vol. 13, pp. 148–162, 2023.

[15] J. S. Komputer, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," vol. 5, no. November 2019, pp. 697–711, 2021.