Application of The C4.5 Algorithm for New Student Admission Selection via Achievement Track at MAN 1 Samarinda

Ramadiani Ramadiani[1*], Hanie Sri Romandani[1], Ghubta Mahendra Putra[1], Muhammad Labib Jundillah[1] and Azainil Azaini

[1] Mulawarman University, Samarinda, East Kalimantan 75119, Indonesia
ramadiani@unmul.ac.id

Abstract. Participating in the New Student Admissions selection process is one method to pursue education. Schools use this selection to sift and select students who are academically suited to benefit from educational opportunities at the institution. The algorithm is employed at MAN 1 Samarinda To aid in classifying data from prospective students. This algorithm, which generates a decision tree, is a renowned method for classification and prediction. Participating in the New Student Admission selection process is one way to attain secondary education. Schools conduct the PPDB selection annually at the start of each academic year. MAN 1 Samarinda is a school that accepts new students based on their achievements rather than their residential zones. This situation allows students from any area to apply without concern for zoning restrictions. A Decision Support System is needed to assist decision-making based on these issues. DSS is an information system designed to help management make decisions related to semi-structured problems. This system can support the decision-making process for prospective new students based on predetermined criteria. The system's workflow includes all stages of problem-solving, selecting relevant data, and determining the approach used in the decision-making process up to problem resolution and solution implementation. MAN 1 Samarinda helps decide whether or not they are accepted or rejected in classifying the data of prospective new students. C4.5 is a suitable algorithm for classification problems in machine learning and data mining. The admissions process is an educational service designed to ensure citizens' fundamental right to quality and fair education by adhering to objectivity, accountability, transparency, and non-discrimination principles, thus promoting greater access to high-quality educational services.

Keywords: Decision, Data, Tree, Education, Algorithm.

1 Introduction

Education is a learning process that allows every individual to attain higher knowledge and understanding about specific subjects. According to Indonesia's Law No. 20 of 2003, Article 14, the educational levels include elementary, secondary, and higher education. Additionally, preschool education can also be offered, which is not a prerequisite for entering elementary education. As stated in the Republic of Indonesia

M. R. Septyandy et al. (eds.), *Proceedings of the International Conference on Tropical Studies and Its Application (ICTROPS 2024)*, Advances in Engineering Research 263,
https://doi.org/10.2991/978-94-6463-732-8_29

Number 20 of 2003, article 17, secondary education is a continuation of basic education. It includes general secondary education and vocational secondary education. The forms of secondary education are Senior High Schools (SMA), Islamic Senior High Schools (MA), Vocational High Schools (SMK), and Vocational Islamic Senior High Schools (MAK), along with other equivalent institutions. Participating in the New Student Admission (PPDB) selection process is one way to attain secondary education. Schools conduct the PPDB selection annually at the start of each academic year. According to the Minister of Education and Culture Regulation Article 21 No 44 of 2019, which governs New Student Admissions for Kindergarten, Elementary, Junior High, Senior High, and Vocational Schools, the PPDB process involves several stages: announcing candidate registration, the registration process, selection of candidates, announcing the results, and re-registration.

MAN 1 Samarinda is a school that accepts new students based on their achievements rather than their residential zones. This situation allows students from any area to apply without concern for zoning restrictions. The school offers three main academic tracks: Science, Social Studies, and Religion. Additionally, it provides four vocational skill options: Culinary Arts, Fashion Design, Automotive Welding, and Network Computer Engineering. The use of technology in student admissions (PPDB) at MAN 1 Samarinda is relatively new. Due to limited time and experience, the admissions process is considered less than optimal. For instance, there are often issues when uploading grades and images as proof. As a writer, I aim to investigate the problems faced by MAN 1 Samarinda.

Based on these issues, a Decision Support System (DSS) is needed to assist decision-making. A DSS is an information system designed to help management make decisions related to semi-structured problems [1]. This system can support the decision-making process for prospective new students based on predetermined criteria. The system's workflow includes all stages of problem-solving, selecting relevant data, and determining the approach used in the decision-making process up to problem resolution and solution implementation [2]. The C4.5 algorithm is proposed to assist MAN 1 Samarinda in classifying the data of prospective new students, helping determine whether they are accepted or rejected. C4.5 is a suitable algorithm for classification problems in machine learning and data mining. Its main advantages include the ability to generate quickly interpretable models in the form of trees or rules, acceptable accuracy levels, and the capability to handle both discrete and numeric attributes [3].

According to previous research, the random forest method for New Student Admissions (PPDB), using five attributes and data from all junior high schools in Cimahi over two years (2017 and 2018), achieving an accuracy of 98.7% [4]. In another study used the C4.5 Algorithm to select Smart Indonesia Card (KIP) recipients, utilizing nine predictor attributes and one target attribute, resulting in training and testing data accuracy of 90.24% [3]. This condition demonstrated the algorithm's effectiveness in selecting KIP recipients [3]. The C4.5 algorithm to predict PPDB outcomes at SMK Hasanah, considering four criteria: region of origin, test scores, parent's occupation, and significance [4]. Test scores, having the highest gain, were chosen as the root of the decision tree [5]. Lastly, the C4.5 algorithm to determine the placement and admission of new students at SMK Plus PGRI 1 Cibinong, using three attributes: written test, national exam scores, and aptitude test, with an accuracy rate of 90%. Based on the discussion above, the author has applied the C4.5 method to identify the best students for admission through the achievement track at Madrasah Aliyah Negeri 1 Samarinda.

2 Materials and Methods

2.1 New Student Admissions

According to the Director General of Islamic Education's Decree Number 7292 of 2020, which provides technical guidelines for New Student Admissions at Raudhatul Athfal, Madrasah Ibtidaiyah, Madrasah Tsanawiyah, Madrasah Aliyah, and Madrasah Aliyah Kejuruan for the 2021/2022 academic year, madrasahs are a form of general education distinguished by their Islamic character, overseen by the Ministry of Religion. One mission of the Ministry of Religion is to "improve access to and the quality of general education with Islamic characteristics, religious education in general educational institutions, and religious education." The New Student Admissions process is an educational service designed to ensure citizens' fundamental right to quality and fair education by adhering to objectivity, accountability, transparency, and non-discrimination principles, thus promoting greater access to high-quality educational services.

2.2 New Student Admissions via Achievement Track

The criteria for applicants for the Achievement Track of PPDB, as outlined in the 2024 Technical Guidelines for MAN 1 Samarinda are:

1. Muslim
2. Age should not exceed 21 years as of July 1 of the current year.
3. Must have an average grade of at least 83 for the 3rd, 4th, and 5th semesters on the report card.
4. Have a certificate of achievement at least at the city or district level (optional).
5. Achieve a minimum score of 83 on the interview test.

2.3 Decision Support Systems (DSS)

The concept of Decision Support Systems was first introduced in 1971 by Michael Scoot Morton with the term Management Decision System. Subsequently, various companies, research institutions, and universities began researching and developing Decision Support Systems. From the resulting products, it can be concluded that this system is a computer-based system designed to assist decision-making by utilizing specific data and models to solve various unstructured problems [6-11]. Decision Support System (DSS) is an interactive computer-based system that uses data and models to address unstructured and semi-structured difficulties [6-11].

2.4 Decision Tree

A decision tree model is a classification technique used in data mining, a field that involves extracting knowledge from data. Decision trees utilize parameters to measure the diversity or difference within a dataset [12]. In a decision tree, each node represents a tested attribute, and each branch represents a division based on the test results, and leaf nodes represent specific class groups. The topmost node is the root node, typically an attribute with the most significant impact on a particular class. Generally, decision trees use a top-down search strategy to find solutions [13]. Data in a decision tree is represented in a table format consisting of attributes and records [14]. An example of a decision tree is shown in Fig. 1.

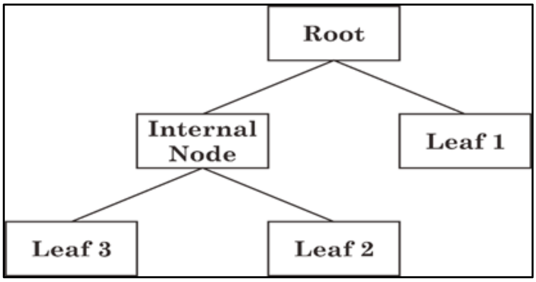

Fig. 1. Example of a Decision Tree

The branches of a decision tree represent classification questions, while the leaves represent classes or groups. Therefore, the primary benefit of using a decision tree is its ability to simplify complex decision-making processes, making it easier for decision-makers to interpret solutions to problems.

2.5 C4.5 Algorithm

C4.5 is an algorithm used to create a decision tree. The C4.5 algorithm is an influential and well-known classification and prediction method [3]. Decision trees are valuable for analyzing data and uncovering concealed connections among prospective and target input variables. The approach involves converting tabular data into a tree structure, translating this tree structure into rules, and then streamlining these rules. The C4.5 algorithm for constructing decision trees in the following steps [12]:

1. Select an attribute as the root.
2. Generate branches for each value.
3. Distribute cases across these branches.
4. Repeat the process for each branch until all cases within a branch are of the same class.

The selection of the root is based on the highest gain among the attributes. Additionally, the gain is calculated using the following formula:

$$Gain(S, A) = Entropy(S) \sum_{1=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (1)$$

where S is the set of cases, A is the attribute, n is the number of partitions of attribute A, $|S_i|$ is the number of cases in the i-th partition, |S| is the total number of cases in S, and I is the iteration index

We first need to calculate the Entropy value to determine the Gain value. Entropy is crucial for selecting the next node splitter, as a higher Entropy value indicates a more significant potential for classification. The formula for calculating entropy is as follows:

$$Entropy(S) = \sum_{1=1}^{n} -p_i * \log_2 p_i \qquad (2)$$

where S is the set of cases, n is the number of partitions of attribute S, and p_i is the ratio of Si relative to S

2.6 Confusion Matrix

The Confusion Matrix is a test method employed to assess the effectiveness or accuracy of classification processes. It helps evaluate a classifier's capability to identify records belonging to various classes [13]. The table for the Confusion Matrix is presented in Fig. 2.

CONFUSION MATRIX		ACTUAL	
		CORRECT	WRONG
PREDICTION	CORRECT	TP	FP
	WRONG	FN	TN

Fig. 2. Confusion Matrix Table

Details:

1. TP (True Positive), the number of positive data correctly classified by the system.
2. TN (True Negative), the number of negative data correctly classified by the system.
3. FP (False Positive), the number of positive data misclassified by the system.
4. FN (False Negative), the number of negative data misclassified by the system-Select an attribute as the root.

The accuracy value, which indicates how precisely the system classifies data correctly, can be determined from the TP, TN, FP, and FN values. This accuracy is calculated by comparing the amount of data correctly classified to the total data, using the equation provided below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

2.7 Entity Relationship Diagram (ERD)

Fig. 3 displays the Entity-Relationship Diagram developed for the system's design.

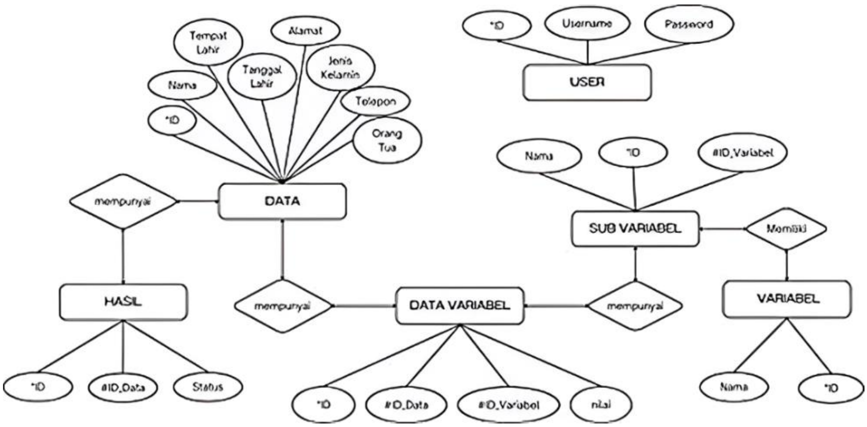

Fig. 3. Entity Relationship Diagram

2.8 Data Flow Diagram (DFD)

Below is the Data Flow Diagram planned for use in developing the system.

- DFD Level 0 / Diagram Context

Fig. 4. DFD Level 0

- DFD Level 1
  Fig. 5 displays the Entity-Relationship Diagram (ERD) developed for the system's design.

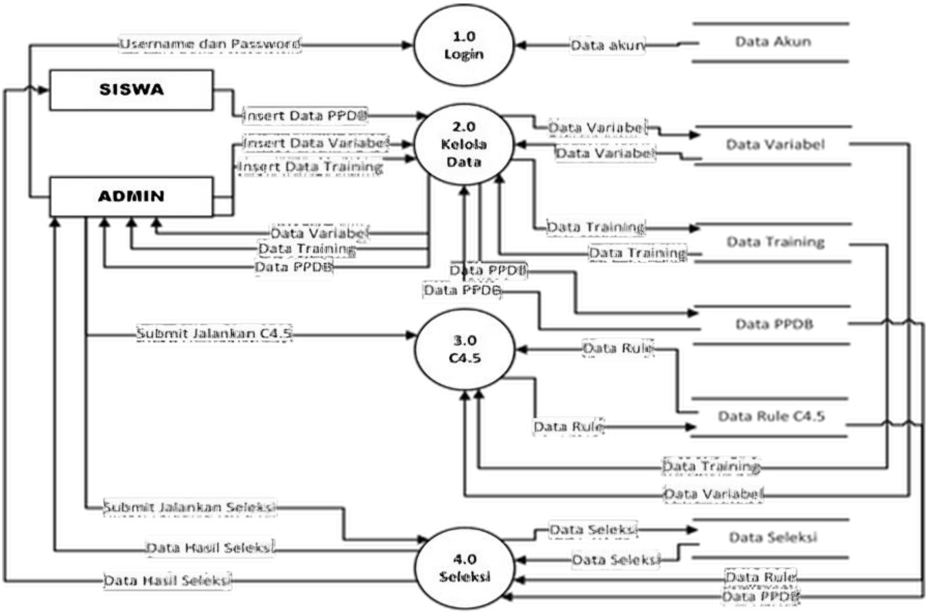

Fig. 5. DFD Level 1

2.9 Entity Relationship Diagram (ERD)

Below are the research implementation phases that will be utilized in the development of the system.

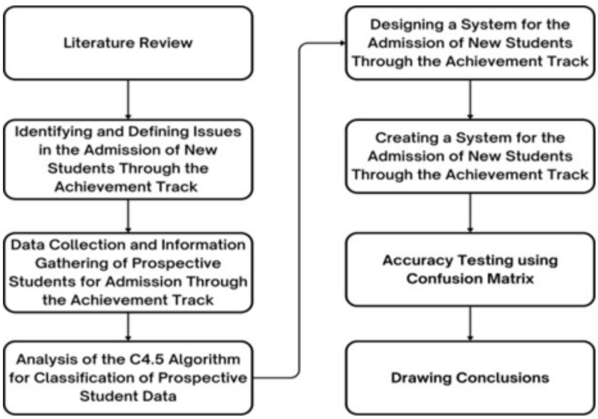

Fig. 6. Stages of Research Implementation

3 Results and Discussion

Preparation of Training and Test Data, Attributes, and Sub-Attributes. We utilize the 2023 PPDB data encompassing 138 students for the training dataset. After the decision tree has been created from the training data, its accuracy will be evaluated using the 2023 PPDB data, which includes 138 students, as shown in Table 1.

Table 1. Training Dataset

SEQ	FULL NAME
A1	Adila Rahmah
A2	Adilla Nur Hidayah
A3	Aditya Dwi Saputra
….	….
A136	Yudina Audiah
A137	Zahirah Alazizah
A138	Zaskia Marwah

The research incorporates three attributes and their corresponding sub-attributes according to the Technical Guidelines for PPDB MAN 1 Samarinda for 2024, as outlined in Table 2.

Table 2. Attribute and Sub-Attribute Data

SEQ	Attribute	Sub Attribute	Score
1	Report Card Score	Good	>=83
		Poor	<83
2	Certificate	Available	1
		Not Available	0
3	Interview Score	Good	>=83
		Poor	<83
4	Graduation Status	Pass	
		Fail	

Below is the compatibility data of the 2023 PPDB data with the sub-attributes according to the predefined attributes, detailed in Table 3.

Table 3. Alternative and Criteria Matching Data

SEQ	Attribute			
	Report Card Score	Certificate	Interview Score	Graduation Status
A1	Good (90)	Available	Good (88)	Pass
A2	Good (86)	Not Available	Poor (79)	Fail
A3	Poor (82)	Not Available	Good (90)	Pass
…	…	…	…	…
A136	Good (88)	Not Available	Poor (76)	Fail
A137	Good (87)	Not Available	Good (99)	Pass
A138	Good (88)	Not Available	Good (89)	Pass

Application of the C4.5 Algorithm and Decision Tree. The steps involved in creating a decision tree using the C4.5 Algorithm.

3.1 The C4.5 algorithm is outlined as follows

1. Initial Stage: Identifying the Root (Iteration 1)
   Determine the number of cases and count the number of pass and fail decisions.

Table 4. Overall Cases

Attribute	Sub Attribute	Total (P + F)	Pass(P)	Fail(F)	Entropy	Gain
Total	Total	138	115	23		
ReportCard Score	>= 83 (Good)	120	101	19		
	< 83 (Poor)	18	14	4		
Certificate	Available	16	16	0		
	Not Available	122	99	23		
Interview Score	>= 83 (Good)	109	109	0		
	< 83 (Poor)	29	6	23		

- Calculating the entropy value
  Use equation (1) to calculate the entropy value Entropy

$$(\mathbf{Total}) = -\left(\frac{\mathbf{115}}{\mathbf{138}} * \mathbf{log_2}\frac{\mathbf{115}}{\mathbf{138}}\right) + -\left(\frac{\mathbf{23}}{\mathbf{138}} * \mathbf{log_2}\frac{\mathbf{23}}{\mathbf{138}}\right)$$

=-(0.833333 * -0.263034) + -(0.166667 * -2.584963)

= 0.219195 + 0.430827

= 0.650022

Attribute 1 (Report Card Score) Entropy =

$$(\mathbf{Good}) = -\left(\frac{\mathbf{101}}{\mathbf{120}} * \mathbf{log_2}\frac{\mathbf{101}}{\mathbf{120}}\right) + -\left(\frac{\mathbf{19}}{\mathbf{120}} * \mathbf{log_2}\frac{\mathbf{19}}{\mathbf{120}}\right)$$

=-(0.841667 * -0.248679) + -(0.158333 * -2.658963)

= 0.209305 + 0.421002

= 0.630307

$$\mathbf{(Poor)} = -\left(\frac{\mathbf{14}}{\mathbf{18}} * \mathbf{log_2}\frac{\mathbf{14}}{\mathbf{18}}\right) + -\left(\frac{\mathbf{4}}{\mathbf{18}} * \mathbf{log_2}\frac{\mathbf{4}}{\mathbf{18}}\right)$$

=-(0.777778 * -0.362570) + -(0.222222 * -2.169925)

= 0.281999 + 0.482206

= 0.764205

Attribute 2 (Certificate) =

$$\mathbf{(Entropy)} = -\left(\frac{\mathbf{16}}{\mathbf{16}} * \mathbf{log_2}\frac{\mathbf{16}}{\mathbf{16}}\right) + -\left(\frac{\mathbf{0}}{\mathbf{16}} * \mathbf{log_2}\frac{\mathbf{0}}{\mathbf{16}}\right)$$

=-(1 * 0) + -(0 * 0)

= 0 + 0

= 0

The resulting table displays the attributes and Entropy values for each sub-attribute following the calculations. These Entropy values will be utilized to determine the gain for each attribute.

Table 5. Total Entropy Value

Attribute	Sub Attribute	Total (P + F)	Pass (P)	Fail (F)	Entropy	Gain
Total	Total	138	115	23	0.650022	
Report Card Score	>= 83 (Good)	120	101	19	0.630307	
	< 83 (Poor)	18	14	4	0.764205	
Certificate	Available	16	16	0	0	
	Not Available	122	99	23	0.698375	
Interview Score	>= 83 (Good)	109	109	0	0	
	< 83 (Poor)	29	6	23	0.735509	

- Calculating the gain value for each attribute
  Use equation (2) to determine the gain value for each attribute
  Atrribute 1 (Report Card Score)

$$\mathbf{Gain} = \mathbf{0.650022} - \left(\frac{\mathbf{120}}{\mathbf{138}} * \mathbf{0.630307}\right) + \left(\frac{\mathbf{18}}{\mathbf{138}} * \mathbf{0.764205}\right)$$

$$=0.650022 - (0.548093 + 0.099679)$$

$$= 0.\ 650022 - 0.647772$$

$$= 0.002250$$

Atrribute 2 (Certificate)

$$\textbf{Gain} = \mathbf{0.650022} - \left(\frac{\mathbf{16}}{\mathbf{138}} * \mathbf{0}\right) + \left(\frac{\mathbf{122}}{\mathbf{138}} * \mathbf{0.698375}\right)$$

$$=0.650022 - (0 + 0.617404)$$

$$= 0.\ 650022 - 0.617404$$

$$= 0.032619$$

Atrribute 3 (Interview Score)

$$\textbf{Gain} = \mathbf{0.650022} - \left(\frac{\mathbf{109}}{\mathbf{138}} * \mathbf{0}\right) + \left(\frac{\mathbf{29}}{\mathbf{138}} * \mathbf{0.735509}\right)$$

$$=0.650022 - (0 + 0.154563)$$

$$= 0.\ 650022 - 0.154563$$

$$= 0.\ 495459$$

Upon completing the calculations, the resulting table displays the gain values for each attribute. The attribute with the highest gain value will be selected as the node.

Table 6. Total Gain Value

Attribute	**Sub Attribute**	**Total (P + F)**	**Pass (P)**	**Fail (F)**	**Entropy**	**Gain**
Total	Total	138	115	23	0.650022	0.002250
ReportCard Score	>= 83 (Good)	120	101	19	0.630307	
	< 83 (Poor)	18	14	4	0.764205	
Certificate	Available	16	16	0	0	0.032619
	Not Available	122	99	23	0.698375	
Interview Score	>= 83 (Good)	109	109	0	0	0.495459
	< 83 (Poor)	29	6	23	0.735509	

- Determining the node and the next branch

The Certificate attribute, with a gain value of 0.081910, is chosen as the next node in the decision tree. The Present sub-attribute of the Certificate, which previously had a Poor Interview Score, has an Entropy of 0, thus ending the calculations. The Present sub-attribute records one student passing and none failing, making it a branch labeled pass. The Absent sub-attribute of the Certificate, which also had a Poor Interview Score, has an Entropy of 0.676942, prompting further calculations with the Absent sub-attribute as the next branch. The decision tree is shown in Fig. 7.

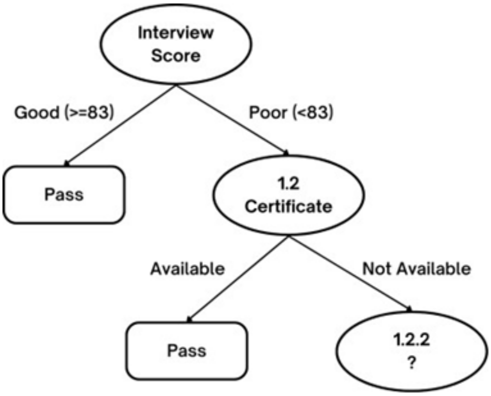

Fig. 7. Decision Tree with Certificate Node.

2. Accuracy Testing Using Confusion Matrix
   Accuracy is calculated using Equation 3, as shown in Table 7.

Table 7. Confusion Matrix Table

CONFUSION MATRIX		**ACTUAL**	
		CORRECT	**WRONG**
PREDICTION	**CORRECT**	115	0
	WRONG	5	23

$$\textbf{Accuracy} = \frac{\mathbf{115 + 23}}{\mathbf{115 + 0 + 5 + 23}}$$

$$= \frac{\mathbf{133}}{\mathbf{138}}$$

= 96,38%

As a result of the calculations above, an accuracy rate of 96.38% was achieved.

4 Conclusions

This research concludes that the C4.5 method is effective for the PPDB student selection process at MAN 1 Samarinda, achieving an accuracy rate of 96.38% as tested by the confusion matrix.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Nurjaman, J., Rosyid, H., Devi, P.A.R.: Sistem Pendukung Keputusan Dengan Metode Topsis Untuk Penyeleksian Penerimaan Siswa Baru. Indexia **3**, 23 (2021).
2. Oktavia, P.: Sistem Pendukung Keputusan Seleksi Penerima Beasiswa dengan Metode Weighted Product pada SMP Negeri 1 Parung Berbasis Web. J. Inform. Univ. Pamulang **3**, 80 (2018).
3. Yunus, M., Ramadhan, H., Aji, D.R., Yulianto, A.: Penerapan Metode Data Mining C4.5 Untuk Pemilihan Penerima Kartu Indonesia Pintar (KIP). Paradig. - J. Komput. dan Inform. **23**, (2021).
4. Winanjaya, R., Amir, F., Doni, R.: Penerapan Data Mining Untuk Memprediksi Penerimaan Peserta Didik Baru Menggunakan Algoritma C4.5. In: Pros. Semin. Nas. Ris. Inf. Sci. 1, 1 (2019).
5. Sumpena, Akbar, Y., Nirat: Penerimaan Calon Siswabaru dan Penentuan Penjurusan Dengan Algoritma C 4.5 SMK PLUS PGRI 1 CIBINONG. Selection and peer-review under responsibility of the 11th STIKOM CKI on SPOT. in CKI On SPOT **11** (2018).
6. Ramadiani, Hatta, H.R., Novita, N., Azainil: Comparison of two methods between TOPSIS and MAUT in determining BIDIKMISI scholarship. In: Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018 1–6 (2018). doi:10.1109/IAC.2018.8780455.
7. Turban, E., Aronson, J.E., Liang T- P.: Decision Support Systems and Intelligent Systems 7th Edition. Pearson/Prentice Hall, New Jersey (2020).
8. Ramadiani, Rani, F.P., Khairina, D.M., Hatta, H.R.: Sistem Pendukung Keputusan Pemilihan Pramuka Pandega Berprestasi Menggunakan Metode Multi Objective Optimization on the Basis of Ratio Analysis. J. Teknol. Inf. dan Ilmu Komput. **6**, 155 (2019).
9. Ramadiani, Arifin, Z., Jundillah, M.L., Azainil: Decision support system for determining Chili land using weighted product method. Bull. Electr. Eng. Informatics **9**, 1229–1237 (2020)
10. Ramadiani, Rahmana, A.R., Islamiyah, Balfas, M.D., Rahman, T & Yunianta, A.: Decision Support System of Direct Cash-Village Fund Recipients Using Multi Attribute Utility Theory. In: Proceedings - International Conference on Informatics and Computational Sciences vols 2021-November 232–237 (2021).
11. Ramadiani, Setiawan, D.A., Jundillah, M.L., Amatullah, D.H., Azainil, Agus, F., Dengen, N., Patulak, I.M., Widians, J.A.: Decision System for Beneficiaries of the Family Hope Program Using the Rank Order Centroid Method and Weighted Product Method. Migr. Lett. **20**, 1170–1182 (2023).

12. Rizki, N.A., Mumtaza, M., Dewi, C.F., Syahlafandi, D., Fendiyanto, P.: Implementasi Pohon Keputusan untuk Klasifikasi Cara Belajar Mahasiswa Mandiri dan Berkelompok Berdasarkan Sumber Belajarnya. Pros. Semin. Nas. Pendidik. Mat. Univ. Mulawarman **3**, 101–106 (2023).
13. Sidette, J.A., Eko, E., Nurhayati, O.D.: Pendekatan Metode Pohon Keputusan Menggunakan Algoritma ID3 Untuk Sistem Informasi Pengukuran Kinerja PNS. J. Sist. Inf. Bisnis **4**, 75–86 (2014).
14. Setio, P.B.N., Saputro, D.R.S., Winarno, B.: Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. Prism. Pros. Semin. Nas. Mat. **3**, 64–71 (2020).