JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 5, No. 2 March 2023 DOI: https://doi.org/10.34288/jri.v5i2.516
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

THE IMPLEMENTATION OF THE C4.5 ALGORITHM FOR DETERMINING THE DEPARTMENT OF VOCATIONAL HIGH SCHOOL

Mirza Sutrisno[1], Jefri Kusuma Rambe[2], Asruddin[3], Ade Davy Wiranata[4]

Teknik Informatika, Universitas Muhammadiyah Jakarta
Jakarta Indonesia
mirza.sutrisno@umj.ac.id*)

Ilmu Komputer, Universitas Budi Luhur
Jakarta Indonesia
jefriekussuma@gmail.com

Sistem Komputer, Universitas Bung Karno
Jakarta Indonesia
asruddin69@gmail.com

Teknik Informatika, Universitas Muhammadiyah Prof. Dr. HAMKA
Jakarta Indonesia
adedavy@uhamka.ac.id

(*)Corresponding Author

Abstrak

Pemilihan jurusan di Sekolah Menengah Atas (SMK) adalah sebuah keharusan bagi peserta didik dalam menentukan konsentrasi peminatan belajar siswa selama tiga tahun di sekolah. Kurangnya pengetahuan siswa dan sosialisasi tentang jurusan ini menyebabkan tidak sedikit dari siswa menetukan pilihan berdasarkan pilihan terbanyak dari rekan sesama pelajar yang mengakibatkan kesulitan dalam mengikuti peminatan pembelajaran dan tidak sedikit yang gagal. Setiap siswa perlu menemukan jurusan yang sesuai dengan minat, kemampuan, dan bakat mereka. Dikarenakan setiap siswa memiliki kemampuan untuk berpikir serta bakat yang berbeda. Algoritma C4.5 dapat memberikan kemudahan dalam pengelompokan mahasiswa berdasarkan jurusan. Menggunakan metode decision tree dengan atribut-atribut yang digunakan seperti nilai matematika, Bahasa Inggris, minat dan bakat, sistem dapat merekomendasikan pilihan jurusan berdasarkan tingkat peminatan siswa. Hasil dari penelitian ini adalah menentukan jurusan dengan akurasi perhitungan menggunakan metode confusion matrix yang dengan tingkat akurasi 98,55% dan nilai recall rate 100.

Kata Kunci : SMK; Sistem Rekomendasi; Pemilihan Jurusan; Algoritma C4.5

Abstract

The selection of departments in vocational high schools (SMK) is a must for students to determine the concentration of student learning interest for three years in a school. The lack of student knowledge and outreach about this department caused many students to choose their majors by the most choices and following other students. This problem can cause some difficulties for the students to participate in learning, and most fail. Students must select their major based on their interests, abilities, and talents because every student has different abilities and talents. The C4.5 algorithm can provide convenience in grouping students based on majors. Using the decision tree method with attributes such as grades in mathematics, English, interests, and talents, the system can recommend majors based on students' interest levels. The results of this study are the determination of the departments with the accuracy of the calculation using the confusion matrix method with a 98,55% accuracy rate and 100% recall rate value.

Keywords : Vocational School; Recommendation System; Department Selection; C4.5 Algorithm

INTRODUCTION

The students who continue to Vocational High School (SMK) level often struggle to determine the department and concentration of study choice. The various departments are not offered to the students who want to continue to the vocational level. The appropriate selection in the vocational will give the students some motivation and interest in learning. The students' mistakes in determining a department will impact problems such as failure and lost time, energy, and mind. The students must choose a department that suits their interests, abilities, and talent. Students have different thinking skills and talents to do something (Khairina et al., 2015). For selecting study programs in Senior High School (SHS), they have also developed a system for helping students select study programs. The cases used in the study include results of the intelligence test, students' interests, and grades in several subjects (Mulyana et al., 2015). An intelligent knowledge-based system was also developed to provide appropriate and accurate recommendations for determining student learning levels based on the assessment criteria in English Language Course (Sutrisno & Budiyanto, 2019).

This research implements a systematic method for providing departmental recommendations for Vocational High School (SMK) students based on the specified criteria. Data mining uses statistics, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases (Turban et al., 2007). Data mining has some functions for processing in several applications, such as description, estimation, prediction, classification, clustering, and association (Larose, 2005). Data mining is a series of processes to extract added value in the form of information that has not been known manually from a database. The resulting information is obtained by extracting and recognizing important or interesting patterns from the data contained in the database (Soufitri et al., 2021).

The Algorithm of C4.5 determines the students who take the department according to their educational background, interests, and abilities of students. The major's selection parameter is a Grade Passing Academy (GPA) in Semesters 1 and 2. This research produces the experiments and evaluations showing that The accurate Decision Tree C4.5 algorithm is applied for determining the suitability of student majors with 93.31% accuracy and departmental recommendations of 82.64% (Swastina, 2018). The other fields of education also use the C4.5 algorithm to classify the students' successful predicates. The analysis used Data Mining using the C4.5 method, and the process used Rapidminer software to make decision trees (Luvia et al., 2017).

The decision tree model uses the C4.5 algorithm to develop an effective selection system for vocational schools. The input variables were: interest, academic talent, National Exam score, and gender. The input variables are interest, academic talent, National Exam score, and gender. The C4.5 algorithm was used to build decision trees to describe the relationship between the input variables and the target variable in patterns. The patterns were used to classify the input variables into the target variable. The system results provide appropriate recommendations for up to 83.33% of the 48 tested data (Prabowo & Subiyanto, 2017).

The Algorithm of C4.5 with the decision tree method can provide predictive rule information to describe the association process with the predictions of the students who repeat their studies. The characteristics of classified data can be obtained through decisions and the rule of tree structures, so the testing phase with WEKA software can help predict that students will repeat the course (Azwanti, 2018). The C4.5 algorithm can change a considerable fact into a decision tree representing the rule to determine prospective students' prediction retirement. The result of the research is that the application can classify the new students in a tree structure to produce a rule and predict the possibility of the retirement of new students (Darmawan, 2018).

The Algorithm of C4.5 is used to classify students in determining majors by looking for patterns of rules based on supporting variables in the form of junior high school (SMP) average report cards, academic test scores such as Natural Sciences (IPA) grades, Social Sciences (IPS) grades, and Language scores. The results of this study are in the form of a data mining application with the C4.5 algorithm to predict majors in science, social studies, or language. The level of accuracy obtained is 97.42% (Kurniasari & Fatmawati, 2019).

Classification with the C4.5 Algorithm and the Forward selection method to determine factors of late coming to school. The sample used was questionnaire data for class VIII (eight) State Junior High School students 271, totalling 270 students. Using training data, specific attributes are determined to form a classifier model. The results of this study are the results of the accuracy of the C4.5 method of 60.74%, with the results of the tree showing congestion is a factor of school delay and the results accuracy of 65.93% for Forward

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i2.516
Accredited rank 3 (SINTA 3), excerpts from the decision of the Minister of RISTEK-BRIN No. 200/M/KPT/2020

selection and getting the three best attributes (Puspitasari, 2020).

Implementing the decision tree method can be used in determining student majors using the C4.5 algorithm. Data mining is a gain ratio of student report cards, interests, and talents. Testing the C4.5 decision tree algorithm results can make more accurate predictions in research on department management and department recommendations for students (Baktiar, 2022).

RESEARCH METHODS

C 4.5 Algorithm

The Algorithm of C4.5 is one of the algorithms applied in the data mining process. The C4.5 algorithm is an extension of Quinlan's own ID3 algorithm to generate a decision tree. Like CART, the C4.5 algorithm recursively visits each decision node, and chooses optimal separation, until there is no further separation (Larose, 2005).

Decision Tree

A decision tree is a very well-known method of classification and prediction. The decision tree method converts facts into decision trees that represent rules. The rules can be easily understood with natural language and expressed in database languages such as SQL (Structured Query Language) to find records in specific categories (Luvia et al., 2017). The provisions of the C4.5 algorithm for building a decision tree are as follows:

a. Determining the highest gain value as the root
b. Creating the branches for each attribute
c. Sharing the cases in the branches
d. Repeat the process for each branch until all the cases in the branch have the same class.

The stages calculation of the C4.5 decision tree algorithm has several stages :

1. Preparing the data training.
2. The tree's root was determined from the highest gain value.
3. Calculating the entropy value (Larose, 2005)

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\ pi \quad \text{..... (1)}$$

4. Calculating the gain value.

$$Gain(S, A) = S - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Si \quad \text{..... (2)}$$

5. After the gain value is found, it will continue in the decision tree process.

The research method consists of some steps from Figure 1.

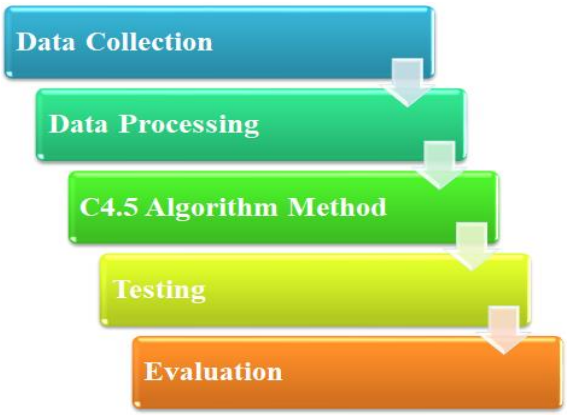

Figure 1. Research's step

Data Collection

The data collection technique uses secondary data from a study of the journal of determining majors using the Naïve Bayes method (Khairina et al., 2015). The authors develop the data into 2000 data sets. Data sets are presented in table 1.

Table 1. Data set

No	MTK	ENG	MINAT	BAKAT	JURUSAN
1	>75	>75	RPL	Multimedia	RPL
2	>75	>75	RPL	Programing	RPL
3	>75	>75	RPL	Teknik Komputer	RPL
4	>75	>75	MM	Multimedia	MM
5	>75	>75	MM	Programing	MM
6	>75	>75	MM	Teknik Komputer	MM
7	>75	>75	TKJ	Multimedia	TKJ
8	>75	>75	TKJ	Programing	TKJ
9	>75	>75	TKJ	Teknik Komputer	TKJ
10	>75	>75	RPL	No	RPL
11	>75	>75	MM	No	MM
12	>75	>75	TKJ	No	TKJ
13	>75	70-75	RPL	Multimedia	RPL
14	>75	70-75	RPL	Programing	RPL
15	>75	70-75	RPL	Teknik Komputer	RPL
16	>75	70-75	MM	Multimedia	MM

No	MTK	ENG	MINAT	BAKAT	JURUSAN
17	>75	70-75	MM	Programing	RPL
18	>75	70-75	MM	Teknik Komputer	MM
19	>75	70-75	TKJ	Multimedia	TKJ
20	>75	70-75	TKJ	Programing	TKJ
21	>75	70-75	TKJ	Teknik Komputer	TKJ
22	>75	70-75	RPL	No	RPL
23	>75	70-75	MM	No	MM
24	>75	70-75	TKJ	No	TKJ
25	70-75	>75	RPL	Multimedia	RPL
26	70-75	>75	RPL	Programing	RPL
27	70-75	>75	RPL	Teknik Komputer	TKJ
28	70-75	>75	MM	Multimedia	MM
29	70-75	>75	MM	Programing	MM
30	70-75	>75	MM	Teknik	MM
31	70-75	>75	TKJ	Multimedia	TKJ
32	70-75	>75	TKJ	Programing	TKJ
33	70-75	>75	TKJ	Teknik	TKJ
34	70-75	>75	RPL	No	RPL
35	70-75	>75	MM	No	MM
36	70-75	>75	TKJ	No	TKJ
37	70-75	70-75	RPL	Multimedia	MM
38	70-75	70-75	RPL	Programing	RPL
39	70-75	70-75	RPL	Teknik	TKJ
40	70-75	70-75	MM	Multimedia	MM
41	70-75	70-75	MM	Programing	MM
42	70-75	70-75	MM	Teknik Komputer	MM
43	70-75	70-75	TKJ	Multimedia	TKJ
44	70-75	70-75	TKJ	Programing	TKJ
45	70-75	70-75	TKJ	Teknik Komputer	TKJ
46	70-75	70-75	RPL	No	RPL
47	70-75	70-75	MM	No	MM
48	70-75	70-75	TKJ	No	TKJ

Data processing

Data is processed and classified based on four criteria to calculate the entropy value and gain—data criteria in table 2.

Table 2. Criteria

Criteria	Description
MTK	Value of Mathematics
ENG	Value of English
Minat	Interest of Students
Bakat	Talent of Students

RESULTS AND DISCUSSION

The modelling of the C4.5 algorithm uses several stages, first calculating the entropy value and then the gain values from the training data. After obtaining the highest gain value, it will be converted into the decision tree. The calculation of the value is represented in Node 1 Table 3.

Table 3. Counting from Node 1

Criteria	Sub criteria	Total Case	RPL	MM	TKJ	Entropy	Gain
	Total	2000	586	668	746	1,57801	
MTK	Value > 75	1008	377	295	336	1,57778	0,0235157
	Value 70-75	992	209	373	410	1,53083	
ENG	Value > 75	1008	292	338	378	1,57703	1,1156631
	Value 70-75	992	294	330	368	1,57893	
MINAT	RPL	670	534	48	88	0,91799	1,2704865
	MM	666	47	615	4	0	
	TKJ	664	5	5	654	0	
BAKAT	Multimedia	502	125	211	166	1,55295	0,1189128
	Programing	500	210	124	166	1,55265	
	Teknik Komputer	374	84	124	166	1,5321	
	NO	498	167	167	164	1,58491	
	Teknik	126	0	42	84	0	

All criteria can represent the calculations with the calculations:

Entropy total = (-Total RPL / Total Case) * IMLOG2 (Total RPL/ Total Case) + (-MM / Total Case) *IMLOG2 (Total MM / Total Case) + (-TKJ/Total Case) * IMLOG2 (Total TKJ / Total Case)

(-586/2000)*IMLOG2(586/2000) + (-668/2000) *IMLOG2 (668/2000) + (-668/2000) * IMLOG2 (746/2000) = 1,57801

Gain criteria value of MTK:
(Entropy Total)-(Total Case Criteria value > 75 / Total Case)* (Entropy value > 75) - ((Total case Value 70-75/Total case)* Entropy value 70-75)
= (1,57801)-(1008/2000)*1,57778) - ((992/2000)* 1,53083)= 0,0235157

Gain criteria value of ENG:
(Entropy Total)-(Total Case Criteria value > 75 / Total Case)* (Entropy value > 75) - ((Total case Value 70-75/Total case)* Entropy value 70-75)
= (1,57801)-(292/2000)*1,57703) - ((294/2000)* 1,57893)= 1,1156631

Gain criteria value of MINAT:
(Entropy Total)-(Total Case Criteria value RPL/ Total Case)* (Entropy RPL) - ((Total case value MM/Total case)* Entropy value MM))-(Total Case Criteria value TKJ/ Total Case)* (Entropy TKJ)
= (1,57801)-(670/2000)*0,91799) - ((666/2000)*0) -(664/2000)* 0) = 1,2704865

Gain criteria value of BAKAT:
(Entropy Total)-(Total Case Criteria value Multimedia/ Total Case)* (Entropy Multimedia) - ((Total case value Programming/Total case)* Entropy value Programming))-(Total Case Criteria value Teknik Komputer/ Total Case)* (Entropy Teknik Komputer) -(Total Case Criteria value NO/ Total Case)* (Entropy NO) -(Total Case Criteria value Teknik / Total Case)* (Entropy Teknik)
= (1,57801)-(502/2000)*1,55295) - ((500/2000)* 1,55265) -(374/2000)*1,5321) - (498/2000)* 1,58491) -(126/2000)* 0)= 0,1189128

It can be seen that the Gain value of the 4 Attributes is:

1. MTK : 0,0235157
2. ENG : 1,1156631
3. MINAT : 1,2704865
4. BAKAT : 0,1189128

Moreover, the highest Gain value is the MINAT criteria of 1,2704865, representing a decision tree.

Decision Tree

The decision tree formed from node one is represented in Figure 2 by using rapid miner software.

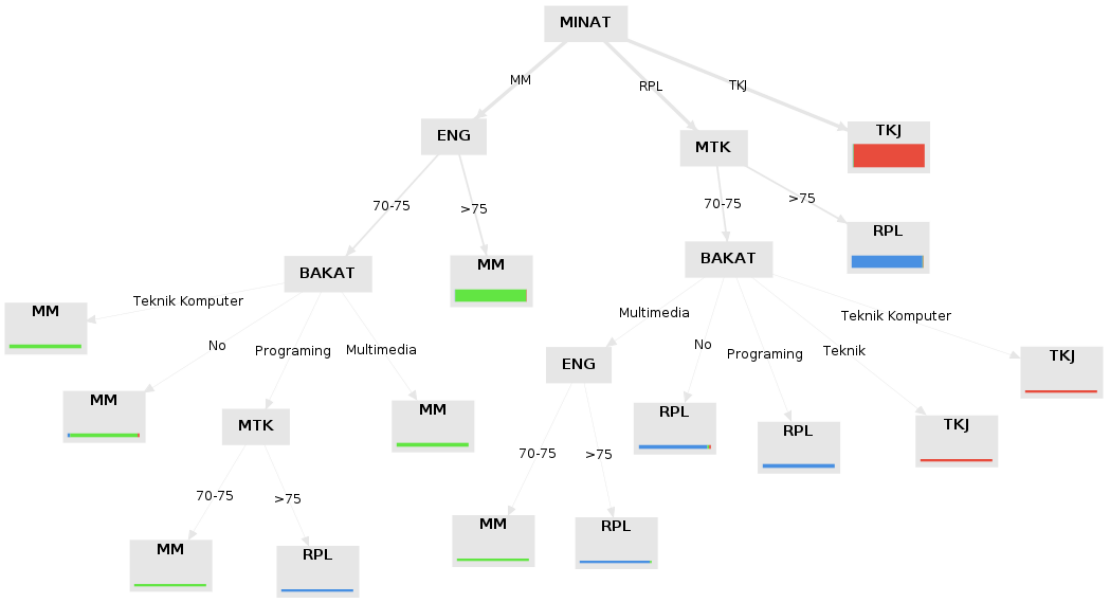

Figure 2. The decision tree

In addition to graphs, the Decision Tree can also be described as follows:

MINAT = MM
| ENG = 70-75
| | BAKAT = Multimedia: MM {RPL=0, MM=84, TKJ=0}
| | BAKAT = No: MM {RPL=3, MM=77, TKJ=2}
| | BAKAT = Programing
| | | MTK = 70-75: MM {RPL=0, MM=40, TKJ=0}
| | | MTK = >75: RPL {RPL=42, MM=0, TKJ=0}
| | BAKAT = Teknik Komputer: MM {RPL=0, MM=82, TKJ=0}
| ENG = >75: MM {RPL=2, MM=332, TKJ=2}
MINAT = RPL
| MTK = 70-75
| | BAKAT = Multimedia
| | | ENG = 70-75: MM {RPL=0, MM=42, TKJ=0}
| | | ENG = >75: RPL {RPL=41, MM=1, TKJ=0}
| | BAKAT = No: RPL {RPL=78, MM=2, TKJ=2}
| | BAKAT = Programing: RPL {RPL=84, MM=0, TKJ=0}
| | BAKAT = Teknik: TKJ {RPL=0, MM=0, TKJ=42}
| | BAKAT = Teknik Komputer: TKJ {RPL=0, MM=0, TKJ=42}
| MTK = >75: RPL {RPL=331, MM=3, TKJ=2}
MINAT = TKJ: TKJ {RPL=5, MM=5, TKJ=654}

The "MINAT" criteria produce the highest gain value with a result of 1,2704865, calculated in Table 4 node 1.2.

Table 4. Counting from Node 1.2

Criteria	Jurusan	Total Case	RPL	MM	TKJ	Entropy	Gain
MINAT	RPL	670	534	48	88	0,91799	1,2704865
	MM	666	47	615	4	0	
	TKJ	664	5	5	654	0	

The decision tree formed from node 1.1 is represented in Figure 3 using RapidMiner software.

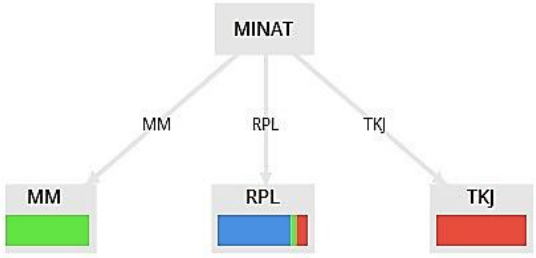

Figure 3. Decision Tree Node 1.1

In addition to graphs, the decision tree can also be described as follows:

TREE
MINAT = MM: MM {RPL=47, MM=615, TKJ=4}
MINAT = RPL: RPL {RPL=534, MM=48, TKJ=88}
MINAT = TKJ: TKJ {RPL=5, MM=5, TKJ=654}

Confusion Matrix Testing

A confusion matrix is a table that states the classification of the amount of data correct test and the number of incorrect test data (Normawati & Prayogi, 2021). The tests of data sets used 2000 data with the confusion matrix method tested using WEKA software to calculate the accuracy and recall value. The calculation of the confusion matrix is in Table 5.

Table 5 Confusion Matrix Classification

Actual Value		Prediction: True	Prediction: False
Actual Value	True	1971 (TP)	29 (FP)
	False	0 (FP)	0 (TN)

$$\boldsymbol{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}\ X\ 100\%$$

$$= \frac{1971 + 0}{1971 + 0 + 29 + 0}\ X\ 100\% = \mathbf{98,55}\%$$

$$\boldsymbol{Recall} = \frac{tp + tn}{fn + tp}\ X\ 100\%$$

$$= \frac{2000}{0 + 2000}\ X\ 100\% = \mathbf{100}\%$$

The results of confusion matrix testing using WEKA software with 10-fold validation can be seen in Figure 4.

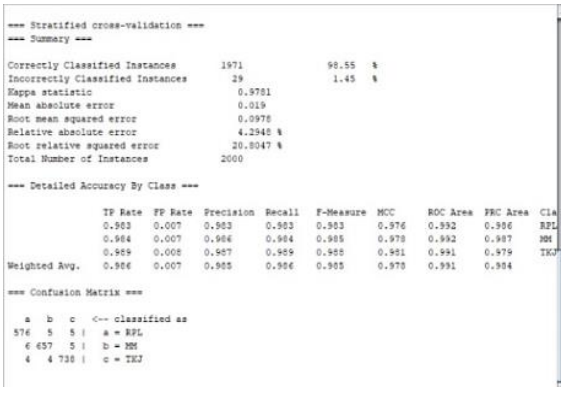

Figure 4. WEKA Testing Result

The test results are also represented as Area Under Curve (AUC) in Figure 5. The curve shows that of the 2000 data tested, 1971 data (98,55%) are correctly classified.

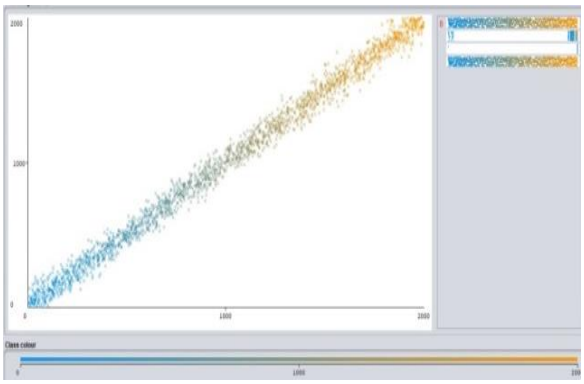

Figure 5. Area Under Curve (AUC)

CONCLUSIONS AND SUGGESTIONS

Conclusion

Based on the description, explanation, and testing, the conclusions are that the C4.5 algorithm method can provide convenience for grouping students based on the departments. Using the decision tree method with the attributes used, such as the value of mathematics, English, interests, and talents, produce a TKJ department with the highest level of specialization. Determining the departments in vocational high school can use the RapidMiner application using the decision tree method and the C4.5 algorithm with the calculation accuracy using the confusion matrix method that has been done by using WEKA software with a 98,55% accuracy rate—and 100% recall rate value.

Suggestion

The following research can try to add other criteria in determining student majors. The next researcher can use more collections of data sets and then test the result with different testing methods on more varied users.

REFERENCES

Azwanti, N. (2018). Algoritma C4.5 Untuk Memprediksi Mahasiswa Yang Mengulang Mata Kuliah (Studi Kasus Di Amik Labuhan Batu). *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *9*(1), 11–22. https://doi.org/10.24176/simet.v9i1.1627

Baktiar, A. (2022). Decission Tree Sebagai Metode Penentuan Penjurusan Perguruan Tinggi Berdasarkan Minat Dan Bakat Melalui Data Raport Dengan Uji Algoritma C4 . 5. *Jurnal Pilar Teknologi*, *7*(1), 40–45. https://doi.org/10.33319/piltek.v7i1.110

Darmawan, E. (2018). C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education. *Jurnal Online Informatika*, *3*(1), 22. https://doi.org/10.15575/join.v3i1.171

Khairina, D. M., Ramadhani, F., Maharani, S., & Hatta, H. R. (2015). Department Recommendations for Prospective Students Vocational High School of Information Technology with Naïve Bayes Method. *2nd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 92–96. https://doi.org/10.1109/ICITACEE.2015.7437777

Kurniasari, R., & Fatmawati, A. (2019). Penerapan Algoritma C4.5 Untuk Penjurusan Siswa Sekolah Menengah Atas. *Jurnal Ilmiah Komputer Dan Informatika (KOMPUTA)*, *8*(1), 19–27. https://doi.org/10.34010/KOMPUTA.V8I1.3045

Larose, D. T. (2005). Discovering Knowledge in Data: An Introduction to Data Mining. In *Discovering Knowledge in Data: An Introduction to Data Mining* (2nd ed., pp. 1–222). John Willey & Sons Inc. https://doi.org/10.1002/0471687545

Luvia, Y. S., Windarto, A. P., Solikhun, S., & Hartama, D. (2017). Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Keberhasilan Mahasiswa Di AMIK Tunas Bangsa. *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika)*, *1*(1), 75–79. https://doi.org/10.30645/jurasik.v1i1.12

Mulyana, S., Hartati, S., Wardoyo, R., & Winarko, E. (2015). Case-Based Reasoning for Selecting Study Program in Senior High School. *International Journal of Advanced Computer Science and Applications*, *6*(4), 136–140. https://doi.org/10.14569/ijacsa.2015.060418

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 697–711. http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

Prabowo, I. M., & Subiyanto, S. (2017). Sistem Rekomendasi Penjurusan Sekolah Menengah Kejuruan Dengan Algoritma C4.5. *Jurnal Kependidikan*, *1*(1), 139–149. https://doi.org/10.21831/jk.v1i1.8964

Puspitasari, C. (2020). Implementation of C4.5 Method To Determine the Factor of Being Late for Coming To School. *Jurnal Riset Informatika*, *2*(3), 115–120. https://doi.org/10.34288/jri.v2i3.132

Soufitri, F., Purwawijaya, E., Hasibuan, E. H., & Singarimbun, R. N. (2021). Testing C4.5 Algorithm Using RapidMiner Applications in Determining Customer Satisfaction Levels. *Jurnal INFOKUM*, *9*(2), 510–517. https://infor.seaninstitute.org/index.php/infokum/article/view/198

Sutrisno, M., & Budiyanto, U. (2019). Intelligent System for Recommending Study Level in English Language Course Using CBR Method. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 153–158. https://doi.org/10.23919/EECSI48112.2019.8977047

Swastina, L. (2018). Penerapan Algoritma C4 . 5 Untuk Penentuan Jurusan Mahasiswa. *Gema Aktualita*, *2*(1), 93–98. https://doi.org/10.24252/insypro.v6i2.7912

Turban, E., E. Aronson, J., & Liang, T.-P. (2007). Decision Support Systems and Business Intelligence. *Decision Support and Business Intelligence Systems, 7/E*, 1–35. https://doi.org/10.1017/CBO9781107415324.004