IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol. x, No.x, July xx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx

APPLICATION OF DATA MINING USING THE C4.5 ALGORITHM AND THE K-NEAREST NEIGHBOR (KNN)

Nurmayanti[1], Supriyanto[2], Merri Parida[3], Sartika[4]
[1] Information systems, Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia, Candi mas, Abung Selatan, Lampung Utara,34511, Indonesia
[3]Department of Information Systems, Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia, Candi mas, Abung Selatan, Lampung Utara
e-mail: nurmayanti89@gmail.com[1], supriyanto@dcc.ac.id[3], merriparida27@gmail.com [3], sartika3639@gmail.com[4]

Abstrak

Bantuan langsung tunai merupakan bentuk intervensi pemerintah atau lembaga sosial yang memberikan bantuan keuangan langsung kepada individu atau keluarga yang membutuhkan. Untuk mengatasi hal tersebut diperlukan suatu sistem yang dapat mengolah data menjadi informasi yang dapat memprediksi siapa saja masyarakat yang berhak dan tidakberhak menjadi penerima bantuan langsung tunai. Sistem prediksi yang akan dikembangkan pada penelitian ini menggunakan metode algoritma C4.5 dan algoritma K. -TETANGGA TERDEKAT (KNN). Metode ini dipilih untuk memprediksi apakah penerima bantuan langsung tunai memenuhi syarat atau tidak berdasarkan status perumahan, pekerjaan, pendapatan, dan status kelayakan. Hasil perhitungan Microsoft Excel dari metode C4.5. Kelas Layak berjumlah 238, tergolong Layak, dan 62 diprediksi Tidak Layak dan tergolong Tidak Layak namun ternyata Layak, dengan total penerima bansos langsung tunai sebanyak 300 orang. Hasil perhitungan Rapid Miner menunjukkan nilai akurasi sebesar 93.00%. Hasil perhitungan Microsoft Excel dari metode k-Nearest Neighbor adalah 226 kelas yang layak, dan 74 kelas yang tidak layak, dengan jumlah penerima bantuan langsung tunai sebanyak 300 orang. Dari hasil perhitungan Rapid Miner terdapat 226 Kelas LAYAK dengan tingkat akurasi 76.55%, Kelas TIDAK LAYAK sebanyak 74 dengan tingkat akurasi 98.23%.
Kata Kunci *- Algoritma C4.5, K-Nearest Neighbor, Data mining, BLT*

Abstract

Direct cash assistance is a governmental or social institution intervention that provides financial aid directly to individuals or families in need. To streamline this process, a system is necessary to convert data into predictive information regarding eligibility for direct cash assistance. This research utilizes the C4.5 algorithm and the K-Nearest Neighbor algorithm for predicting eligibility based on factors such as housing status, employment, income, and eligibility status. Using the C4.5 algorithm, Microsoft Excel calculations identified 238 individuals as eligible and predicted 62 as ineligible who were eligible, out of a total of 300 recipients. The accuracy rate from RapidMiner calculations was 93.00%. Regarding the K-Nearest Neighbor method, Microsoft Excel calculations identified 226 eligible and 74 ineligible recipients out of 300. RapidMiner analysis showed an accuracy rate of 76.55% for the 226 eligible recipients and 98.23% for the 74 ineligible recipients.
Keyword *- Algoritma C4.5, K-Nearest Neighbor, Data mining, BLT*

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

1. INTRODUCTION

Based on population data from Pematang Kasih village, there are 557 families divided into 10 RTs and 8 RKs. From this data, it was recorded that the number of poor people was 380 families. From data on poor residents, 240 families received direct BLT assistance. This problem resulted in a lack of synchronization in recipients of direct BLT assistance. If BLT direct assistance recipients are well recorded and the eligibility requirements for BLT direct assistance recipients are well recorded, it will reduce the of BLT direct assistance recipients and it will also be easier for officers to distribute the BLT direct assistance [1].

To address this issue, a system is required to process data into predictive information that can determine eligibility for direct BLT assistance recipients in Pematang Kasih village. The research will focus on developing a prediction system using the C4.5 method. The study, titled "Application of Data Mining Using the C4.5 Algorithm and K-Nearest Neighbors (KNN) for Determining Receipt of Direct Cash Aid (Case Study of Pematang Kasih Village, West Abung District, North Lampung Regency)," aims to enhance accuracy in identifying eligible and ineligible recipients of direct cash aid [2].

2. METHODS

The flowchart illustrates the methodologies employed by the author, specifically the C4.5 and K-Nearest Neighbor Algorithm methods. In Fig. 2, the stages of the C4.5 algorithm are delineated, starting with data collection, followed by entropy calculation, gain calculation, computation of individual information values, calculation of gain ratio, creation of branches for each value, and iterative repetition of these processes until all nodes are partitioned [3].

The process begins by inputting each transformed data point, followed by determining the count of each neighbor and identifying the central neighbor. Distances from each data point to the cluster center are calculated, and the data is grouped based on the minimum distance to the central neighbor. This calculation process is repeated iteratively to refine and confirm the center point until consistent values are achieved. Finally, the process concludes once stable cluster centers or central points are established, ensuring they accurately represent the data distribution.

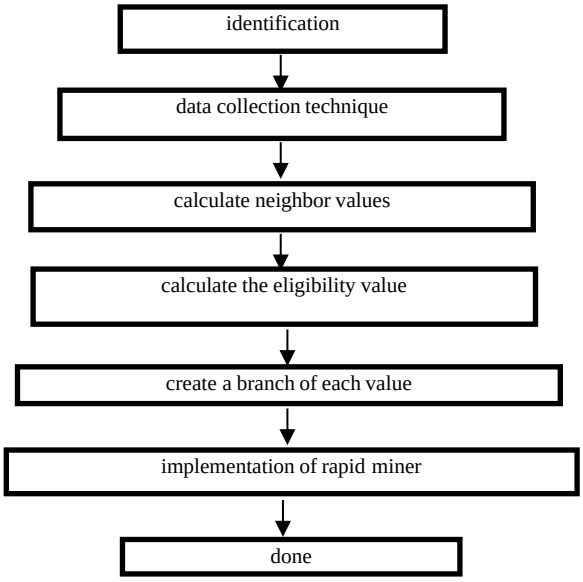

Fig 1 K-Nearest neighbor stages

2.1. Research Resources

The research for this case study was conducted in Pematang Kasih Village, West Abung District, North Lampung Regency. In addition to local data, journals or scientific publications discussing relevant research topics were also consulted. The application of the C4.5 algorithm includes the calculation of entropy and gain values based on specific criteria outlined in this study. These criteria include:

Table 1 Criteria for Direct Cash Assistance Recipients (Blt-Dd)

No.	Recipient' sname	Home status	work	Income	Eligibilit yStatus

2.2. Sample population

In a study, the selected recipients of aid have a close relationship with the problem under study, the recipient or universe is the total number of analysis units whose characteristics will be used. The population in this case is the people who receive direct cash assistance in Pematang Kasih village, sub-district. West Abung, North Lampung Regency in 2021-2023.

The data used is Direct Cash Assistance data from Pematang Kasih Village, West Abung District, North Lampung Regency, some of which are used for 2021-2023.

2.3. Object data

Data on recipients of direct cash assistance is updated every three months, and the total number of recipients in Pematang Kasih Village, West Abung District, North Lampung Regency, amounts to 300 individuals.

2.4. Mathematical Formulation

2.4.1 C4.5 Algorithm

The steps for calculating the Entropy and Gain values for each criterion with high and low information are outlined. The C4.5 algorithm is designed to assist in classifying vehicle test results based on influencing factors. By utilizing the C4.5 (Decision Tree) method and data mining techniques, the author can make informed decisions to predict sales quotas that align with the needs of outlet owners and support sales efforts [4].

Entropy Calculation

The initial step of the C4.5 algorithm involves calculating the entropy value. To determine the total entropy value in a given case, you can use the following formula :

$$Entropy\ (S) = i = 1n - \ pi * \log_2 pi \qquad (1)$$

In summary, *S* is the entire dataset or a subset of cases, *A* is the attribute used to partition *S*, n is the number of resulting partitions from the split, S_i are the subsets or partitions of *S*, and p is the proportion of cases in each subset *Si* relative to the total cases in *S*. These elements are fundamental in the process of constructing decision trees using algorithms like C4.5, where attribute selection and partitioning based on information gain play critical roles in classification tasks [5].

After calculating the entropy value, the next step in the C4.5 algorithm is to compute the gain value for each attribute to determine the optimal attribute for splitting the decision tree. The gain is calculated using the following formula [6] :

$$Gain\ S, A\ = entropyS - in\ = 1SiS \qquad (2)$$

In decision tree algorithms, *S* is progressively split based on different attributes *A*, creating subsets S_i where each subset corresponds to a different branch in the decision tree. The goal is to select

attributes that maximize information gain or reduce entropy, leading to more homogeneous subsets regarding the target variable being predicted [7].

Table 2. Calculation results in Microsoft Excel

Attribute	Mark	Number of cases	worthy	Not feasible	entropy	Gain
Total		300	240	60	0,217322	
Home status						0,01028883
	Rent	110	100	10	0,132302	
	Alone	190	140	50	0,250299	
Work						0,07565576
	Street vendors	126	68	58	0,299661	
	Farm workers	174	172	2	0,027256	
Income						0,156281719
	< 1 million	204	204	0	0	
	1-2 million	62	36	26	0,295356	
	> 2 million	34	0	34	0	

Based on the calculations above, it can be concluded that the accuracy of Testing data which contains 20 data has an accuracy of 65%. Based on the decision tree on the test data above,the criteria that most influence predicting whether it is feasible or not feasible shows that the Gaininformation in Criterion A3 (income) is 0, 19087 is greater than the other criteria.

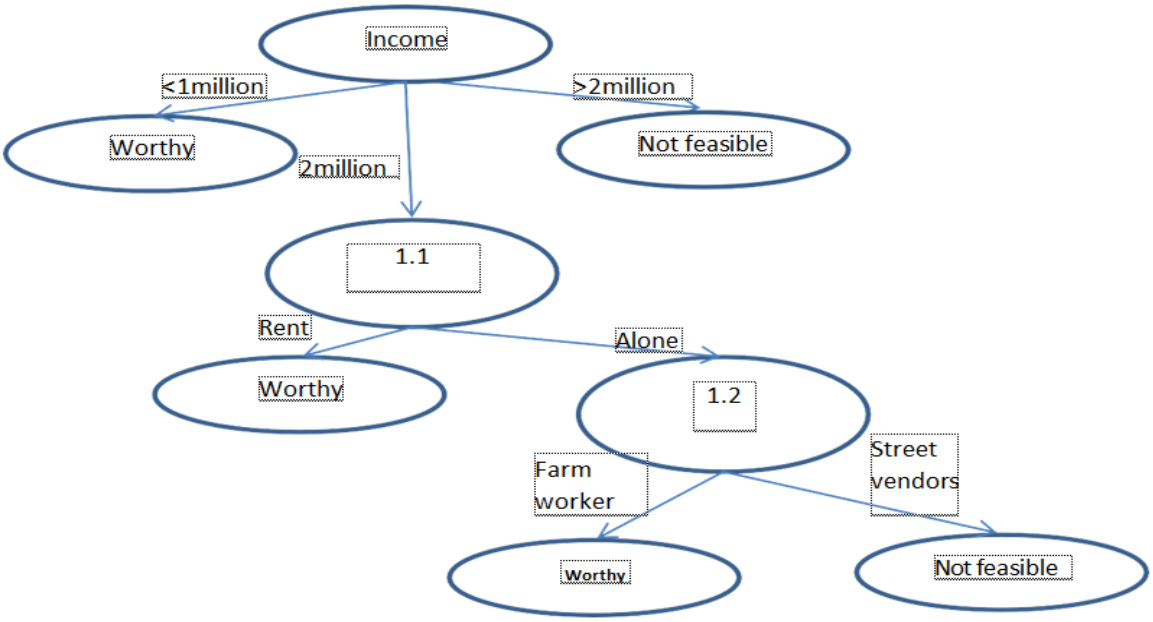

Fig 2 Decision Tree

Accuracy

In determining the percentage of accuracy of the processed data, the following formula is used :

$$\text{Accuracy percentage} = \frac{\text{Amount of Correct Prediction Results Data}}{\text{Number of Predictions Made}} \text{ X } 100\%$$

Calculations from Testing Data with total data from 20 BLT recipients and accuracy presentation as follows:

$$\text{Accuracy percentage} = \frac{240}{300} \text{ X } 100\% = 80\%$$

Tabel 3 Confusion Tabel

	Class	
Prediction	Worthy	Not feasible
Worthy	238	3
Not feasible	2	57

From the results of the data calculation conclusions on the Direct Cash Assistance above using Microsoft Excel, it is known that there are 238 eligible classes classified as Eligible and 2 predicted as Ineligible and classified as Not Eligible but Eligible, with a total of 300 data on direct Cash Assistance recipients. it is also known that each criterion has the following results:

1. The status of the house with the attribute "rental" is worth 100 and not worth 10, with an Entropy value of 0.132302, "Owned" is worth 190 and not worth 50, with an Entropy value of 0.250299, and the Gain value for the house status attributes 0.01028883.
2. Jobs with the attribute "street sword" are 68 Eligible and 58 are Not Eligible, with an Entropy value of 0.299661, "farm laborers" are 172 Eligible and 2 Not Eligible, with an Entropy value of0.027256, and the Gain Value for the Job Attribute is 0.07565576.
3. Income with the attribute "<1 million" is worth 204 and is not worth 0, with an entropy value of 0, "1-2 million" is worth 36 and not worth 26, with an entropy value of 0.295356, ">2 million" is worth 0 and Not Eligible 34, with an Entrpy value of 0, and a Gain Value for Job Attributes of 0.156281719.

2.4.2 K-NEAREST NEIGHBOR

Data used in research must be collected and organized well. It can be collected from various sources such as observations, surveys, or databases where the data is like[9] There are several ways to measure the proximity distance between new data and old data (Training Data), including Euclidean Distance and Manhattan Distance (City Block Distance). In this study, Euclidean Distance was used to measure distance. To measure distance. To measure Euclidean Distance, you can use the following equation:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_{2-y})^2}.....(3)$$

Information:
D: Proximity distance
X: Training data
Y: Testing data
N: Number of individual attributes between 1 and N
F: I attribute similarity function between cases X and Y
I: Individual attributes between 1 and N

CLASSIFICATION

The classification process begins by calculating the distance between the testing data and the training data. Then, test data with training data. Then, the test data is classified into the class that appears most frequently among the K nearest neighbors. From the K value provisions above, the independent variable (dis) value can be calculated as follows [10][11]:

Table 4 Criteria Weight

Number	Valuation Attributes		Assessment Range		Average
	Code	Information	Range	Value	
1.	A1	Home status	Rent	2	2,5
			Alone	1	
2.	A2	Work	Fram workers	2	2,5
			Trader	1	
3.	A3	Income	< 1 million	3	4
			1-2 million	2	
			> 2 million	1	
Average Weight (Eligibility Conditions)					**9**

Model Evaluation

Based on the calculation results from the Training Data for Direct Cash Assistance recipients in Pematangkasih village, North Lampung, using Microsoft Excel, it is determined that there are 226 recipients classified as ELIGIBLE and 74 recipients classified as UNELIGIBLE, making a total of 300 Direct Cash Assistance recipients.

Tabel 5 Counfusion Tabel

INFORMATION	AMOUNT
WORTHY	226
NOT FEASIBLE	74

3. RESULTS AND DISCUSSION

3.1 K-Nearest Neighbor

The following is a prediction calculation with 3 attributes, namely: income, employment, and housing status using Microsoft Excel with data on direct cash assistance recipients as follows:

1. The number of data on direct cash assistance recipients is 300 data
2. It is known that there are 226 eligible classes, and 64 ineligible classes, with a total of 300 direct cash assistance recipients.

From the data above, the author carried out prediction calculations using the Rapid Miner 10.1.2 application for recipients of government assistance. The k-nearest neighbor calculation for the ELIGIBLE class is 226 with an accuracy level of 76.55%, and the number of UNWORTHY classes is 74 with an accuracy level of 98.23% in Figure 3.

accuracy: 76.00%

	true Not feasible	true Worthy	class precision
pred. Not feasible	6	4	60.00%
pred. Worthy	68	222	76.55%
class recall	8.11%	98.23%	

Fig 3 the final result

From the final results of the recipients of assistance from the government of Pematangkasih village of North Lampung above using Rapid Miner 10.1.2, it is known that the DECENT Class is 224 with an accuracy rate of 76.55%, the Ineligible Class is 74 with an accuracy rate of 98.55%, with a total of 300 data. data on DirectCash Assistance recipients.

3.2 C4.5 Algorithm

From the training data predictions involving 300 data points, it is evident that the C4.5 Algorithm, computed using the Rapid Miner 10.1.002 software, achieves an accuracy of 93.00%. The "Income" criterion plays a significant role in determining the eligibility of Direct Cash Assistance recipients, influencing whether they qualify or do not qualify to receive BLT.

accuracy: 92.33%

	true Not feasible	true Worthy	class precision
pred. Not feasible	59	22	72.84%
pred. Worthy	1	218	99.54%
class recall	98.33%	90.83%	

Fig 4 Rapidminer Accuracy Display

The "income" category was created to assist researchers in pinpointing crucial benchmarks for this study. This category includes attributes labeled "Eligible" and "Ineligible," which will be featured in the knowledge tree after the image. Figure 7's code automates computations based on the original C4.5 algorithm. Each code is crafted to forecast outcomes using a confusion matrix, which determines the final decision count. The predictions are displayed as a decision tree, followed by a confusion matrix that shows the accuracy of the C4.5 algorithm as a percentage.

The code in Figure 8 then generates the knowledge tree visualization using RapidMiner. Accuracy metrics, obtained from the confusion matrix table, indicate the precision of the sample data. By differentiating between correct and incorrect classifications, accuracy is calculated until it reaches the set threshold of 93.00%. However, the author considers achieving this exact target to be non-essential based on the analyzed data. These real-world findings can be applied to case studies involving direct cash assistance recipients, providing practical insights from the C4.5 algorithm's calculations. The decision tree is shown in Figure 5.

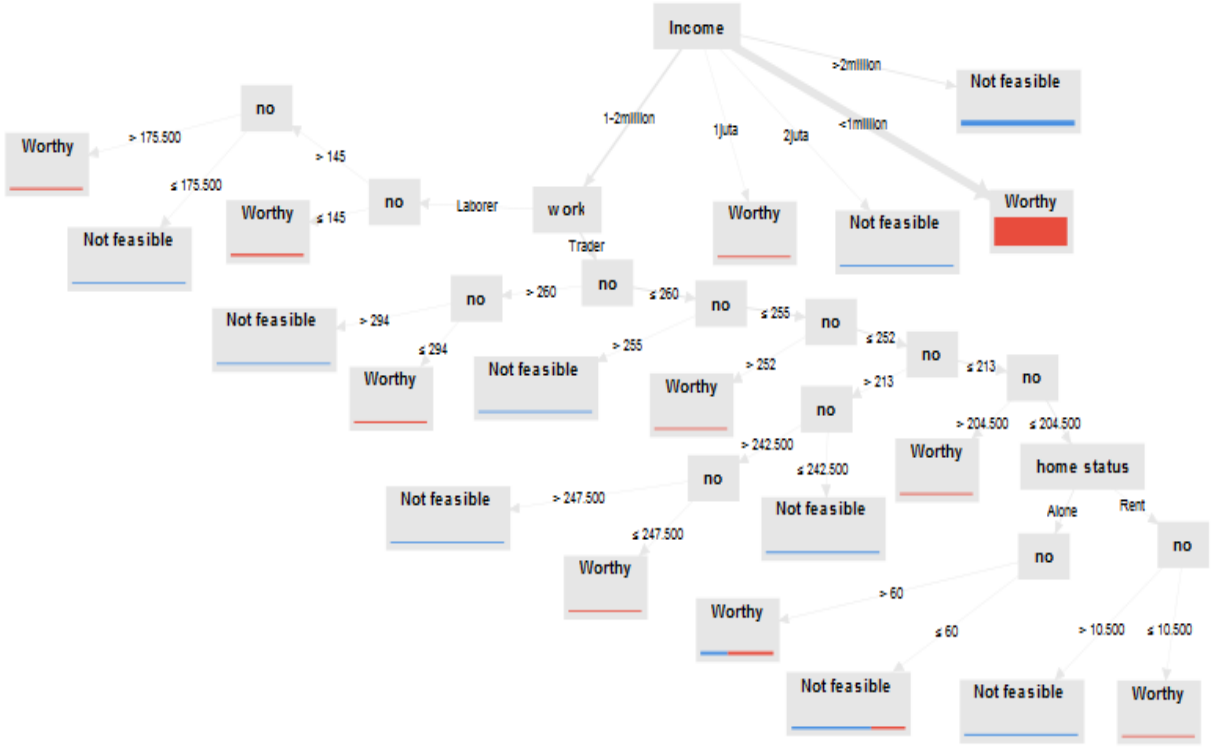

Fig 5 decision tree display

3.3 K-Nearest Neighbor

The following is a prediction calculation with 3 attributes, namely: income, employment, and housing status using Microsoft Excel with data on direct cash assistance recipients as follows:

1. The number of data on direct cash assistance recipients is 300 data
2. It is known that there are 226 eligible classes, and 64 ineligible classes, with a totalof 300 direct cash assistance recipients.

From the data above, the author carried out prediction calculations using the Rapid Miner 10.1.2 application for recipients of government assistance. The k-nearest neighbor calculation for the ELIGIBLE class is 226 with an accuracy level of 76.55%, and the number of UNWORTHY classes is 74 with an accuracy level of 98.23% in Figure 6.

accuracy: 76.00%

	true Not feasible	true Worthy	class precision
pred. Not feasible	6	4	60.00%
pred. Worthy	68	222	76.55%
class recall	8.11%	98.23%	

Fig 6 the final result

From the final results of the recipients of assistance from the government of Pematangkasih village of North Lampung above using Rapid Miner 10.1.2, it is known that the DECENT Class is 224 with an accuracy rate of 76.55%, the Ineligible Class is 74 with an accuracy rate of 98.55%, with a total of 300 data. data on DirectCash Assistance recipients.

4. CONCLUSIONS

From the research, the following conclusions were drawn:
After conducting extensive calculations using the C4.5 and K Nearest Neighbor algorithms, high accuracy scores of up to 100% were achieved. The results indicate that the C4.5 algorithm identifies a higher number of ineligible cases compared to the K-Nearest Neighbor algorithm. The C4.5 algorithm's approach involves centralized data processing when computing gain and entropy, contributing to its highly accurate predictions that favor viable products over non-viable ones. Consequently, this research provides a method for precisely identifying less feasible targets, thereby enhancing decision-making accuracy for the future. The government intends to improve data collection processes to enhance targeting capabilities as planned. Furthermore, this research aims to provide the public with accurate information, promoting greater satisfaction with government services.

Acknowledgments

The authors of this article extend their appreciation to ITBA-PSDKU Dian Cipta Cendikia Kotabumi for their assistance in the development of this research paper.

.

REFERENCES

[1] Y. S. Siregar, B. O. Sembiring, H. Hasdiana, A. R. Dewi, and H. Harahap, "Algortihm C4.5 in mapping the admission patterns of new Students in Engineering Computer," *SinkrOn*, vol. 6, no. 1, pp. 80–90, 2021, [Online]. Available: https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11154

[2] A. Primadewi, F. A. Kurniawan, and E. U. Artha, "Using Data Mining with C4.5 Algorithm for Student Department Selection at MTs N Kaliangkrik," *Borobudur Informatics Rev.*, vol. 1, no. 1, pp. 22–36, 2021, doi: 10.31603/binr.4989.

[3] J. R. M. Ledoh, F. E. Andreas, E. S. Y. Pandie, and C. E. Amos Pah, "C4.5 Algorithm Implementation to Predict Student Satisfaction Level of Lecturer's Performance in the Covid-19 Pandemic," *Komputasi J. Ilm. Ilmu Komput. dan Mat.*, vol. 20, no. 2, pp. 126–134, 2023, doi: 10.33751/komputasi.v20i2.8284.

[4] J. Riyono, A. L. R. Putri, and C. E. Pujiastuti, "Early Detection of COVID-19 Disease Based on Behavioral Parameters and Symptoms Using Algorithm-C5.0," *Indones. J. Artif. Intell. Data Min.*, vol. 6, no. 1, p. 47, 2023, doi: 10.24014/ijaidm.v6i1.22074.

[5] m s mauludin and l hermawanti, "Merger C4. 5 Algorithm and 3. Adaboost for Determining the Department Ipa Students Graduation in Sma Islam Sultan Fatah Wedung Demak," *Proceeding ...*, pp. 3–7, 2016, [Online]. Available: https://www.publikasiilmiah.unwahas.ac.id/index.php/isc/article/view/1664/0

[6] F. Riandari and H. T. Sihotang, "Implementation Of C4.5 Algorithm To Analyze Library Satisfaction Visitors," *Pelita Nusant. Medan Jln. Iskandar Muda*, vol. 4, no. 2, pp. 1076–1084, 2020, [Online]. Available: https://iocscience.org/ejournal/index.php/mantik

[7] E. B. Wijaya, A. Dharma, D. Heyneker, and J. Vanness, "Comparison of the K-Means Algorithm and C4.5 Against Sales Data," *SinkrOn*, vol. 8, no. 2, pp. 741–751, 2023, doi: 10.33395/sinkron.v8i2.12224.

[8] Y. Perwira, A. Sitohang, M. Pandjaitan, and K. Simamora, "Application of the Classification Decision Tree Method to Determine Student Satisfaction Factors for Student Services," vol. 13, no. 02, pp. 87–93, 2023.

[9] E. V. Astuti, A. Afandi, and D. M. Efendi, "Classification and Clustering of Internet Quota Sales Data Using C4.5 Algorithm and K-Means," *J. Ilm. Tek. Elektro Komput.dan Inform.*, vol. 9, no. 2, pp. 268–283, 2023, doi: 10.26555/jiteki.v9i2.25970.

[10] N. A. Prahastiwi, R. Andreswari, and R. Fauzi, "Students Graduation Prediction Based on Academic Data Record Using the Decision Tree Algorithm C4.5 Method,"

JURTEKSI (Jurnal Teknol. dan Sist. Informasi), vol. 8, no. 3, pp. 295–304, 2022, doi: 10.33330/jurteksi.v8i3.1680.

[11] A. T. Indal Karim and S. Sudianto, "Dominant Requirements for Student Graduation inthe Faculty of Informatics using the C4.5 Algorithm," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 3, no. 2, pp. 50–58, 2023, doi: 10.20895/dinda.v3i2.1040.