Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Prediction of Student Scholarship Recipients Using the K-Means Algorithm and C4.5

Rizky Wandri[1*], Yudhi Arta[2], Anggi Hanafiah[3], Rizka Oktaviani[4]
rizkywandri@eng.uir.ac.id[1*], yudhiarta@eng.uir.ac.id[2], anggihanafiah@eng.uir.ac.id[3], rizkaoktaviaani@student.uir.ac.id[4]
[1234] Informatics Engineering Study Program, Riau Islamic University

Article Information

Submitted : 29 Jan 2023
Reviewed: 15 Feb 2023
Accepted : 27 Feb 2023

Keywords

Data Mining, Clustering Method, Classification Method, K-Means Algorithm, and C4.5 Algorithm.

Abstract

The government has a program called "Indonesia Smart Card" to assist students in financing education. Where private college makes the selection manually, with the Data Mining technique, a process will be carried out to speed up the manual process. This research will apply the clustering method with the K-Means algorithm and the classification method with the C4.5 algorithm, as well as a test using the RapidMiner application, which utilizes applicant data in 2022 with a total of 1298 participants. The test results found that 327 participants were in the highest cluster, "cluster_0", then the results of the c4.5 algorithm test obtained a decision tree if the participant has a Indonesia Smart Card and the value obtained from the Income Amount criterion is more than 70 points, then the participant concerned is entitled to scholarship where 317 participants meet the criteria, and the university only has to choose participants from the results obtained in accordance with a predetermined quota.

A. Introduction

PIP (Smart Indonesia Program) is one of the Ministry of Education and Culture's assistance programs in meeting the education needs in Indonesia [1]. PIP is a solution to increase access to education to reduce the number of children dropping out of school [2]. Assistance (cash and study opportunities) is given to students from underprivileged families to finance their education [3]. Scholarships are one of the most needed programs at this time, both for the upper and lower middle class. At private tertiary institutions whose registration will be opened after registration at state tertiary institutions is complete, where private tertiary institutions conduct selection to select prospective participants who are entitled to receive scholarships manually and carry out selections by looking at data per participant in the system that has been provided. From the manual method used by the manager, can the selection process be carried out with data quickly? This is an obstacle that must be resolved.

The process of extracting significant information from huge datasets is known as data mining. The process of data mining makes extensive use of mathematical techniques, statistics, and artificial intelligence [4]. In data mining techniques, there are several methods, namely Classification, Association, Clustering, Regression, Forecasting, Sequencing, and Descriptive. With the application of data mining, an analysis related to predictions will be carried out at the selection stage for receiving the smart Indonesia card scholarship, which will become a reference for universities in the acceptance selection stage. The expected result is to speed up the manual process carried out during the participant selection process. In this study of the many mining methods, the mining process will be carried out using the clustering method using the K-means algorithm and the Classification method using the C4.5 algorithm. This research will utilize the 2021 applicant data to obtain a decision analysis of prospective students who meet the requirements to get a scholarship. Attribute data for clustering will be carried out using the k-means algorithm and continued with making a decision tree with the C4.5 algorithm, and the results obtained will be tested using the RapidMiner application to test whether the results are valid.

Technically, the K-means algorithm groups data into clusters based on the shortest distance (initial centroid value). Distance is calculated using the Euclidean Distance formula [5]. The aim is to create groups on the existing dataset so that the data is sorted according to their respective group's participants whose requirements are complete and will be grouped with complete participants and vice versa. Participants whose requirements are incomplete will be groups with incomplete participants. After the clustering process, the results of the K-means algorithm will become a dataset in the C4.5 algorithm process to create a decision tree from the existing dataset. The classification process is carried out by determining the similarity of the characteristics of a particular class [6], where the results of the decision tree will become a college reference for decision-making.

B. Research Method

a. Research Stages

The method used is the Knowledge Discovery in Databases (KDD) process, which is divided into stages, namely selection, cleaning, transformation, data

mining, and interpretation. The data mining stage will use the k-means and c4.5 algorithms, which are expected to produce better results. Algorithm C4.5 has a weakness in dealing with continuous attributes, which greatly reduces classifier accuracy. Using K-means clustering changes continuous values from attributes to categorical values while building a decision tree [7] is expected to significantly increase classification accuracy. The following diagram depicts the flow of the research process:

Figure 1. Research Flow

Knowledge Discovery in Databases has been widely used alternately with the aim of explaining the process of data development in a large data set [8]. In the KDD process, apply Data Mining Techniques to find knowledge in the following way:

1) Selection: the process of selecting the data used.
   The object of this research is prospective students who are applying for scholarships in 2022. The research was carried out at Riau Islamic University, then the data obtained was qualitative in nature and contained information about each variable that has been determined by the Ministry of Education and Culture in receiving scholarships.
2) Cleaning: removing incomplete and inconsistent entries from the data in order to produce a normalized dataset.
   The stage where data in 2021 will be carried out. The process of selecting or selecting data that is considered relevant to the analysis and irrelevant data will be eliminated.
3) Transformation: Transform data to apply data mining methods to discover previously unknown patterns [9].
   At this stage, the initials will be assigned, and the specified data transformation technique will be converted into a mining procedure.
4) Data Mining: the extraction of a pattern from data using certain techniques and algorithms.

In this research, we will use clustering and classification techniques. The K-Means and C4.5 algorithms are used for clustering and classification, respectively.

5) Interpretation: the process of evaluating a pattern that has been determined. The purpose of the evaluation is to see the suitability of the results that have been found and used for decision-making [10].

b. Data Mining

Data Mining is the process of mining large data using certain techniques or methods [11] that can produce effective and valuable knowledge [12]. Another definition of data mining is a sequence of computer procedures that automatically extract previously unknown knowledge from a data source [13]. The three steps of data mining consist of data, analysis, and decision-making. Data mining aims to extract pertinent information from data, to help better decision-making [14].

c. Clustering

Clustering is a technique of data mining, grouping data into clusters so that the cluster contains similar/similar data that is different from data in other clusters [15]. K-means is the clustering algorithm, and the following sections detail each stage of the procedure.

Step 1 : Take K as the number of clusters.

Step 2 : Any of the K data points can be chosen at random as the cluster center. The greater the distance between cluster centers, the better [16].

Step 3 : Determine the distance between the center of each cluster and each data point. To calculate the distance, you can use the Euclidean distance method [17].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

where:

$d(x,y)$= dissimilarity size

x_i = $(x_1, x_2 \cdots, x_i)$ namely data variables

y_i = $(y_1, y_2 \cdots, y_i)$ i.e., the variable at the central point

Each data point must be assigned to a cluster. Find the cluster with the nearest center and add data points to the cluster.

Step 4 : If a new cluster occurs, recalculate the center. The cluster center is calculated by averaging all data points inside the cluster.

Step 5 : Continue alternating between Steps 3 and 5 until one of the following conditions is met: The newly created cluster retains its original center. Every point is still in the original cluster. The iteration ends at this point.

d. Classification

Classification is a data mining approach for predicting group membership within a data set. [18]. C4.5 is an evolution of ID3 and is the algorithm for data mining classification [19]. The C4.5 algorithm evolves into a machine-learning classification algorithm capable of making accurate predictions [20].

There are various factors to look for while solving cases using the C4.5 method, including:

Step 1 : Entropy (S) - Parameters that are applied to a data collection to determine how diverse each attribute value is for the decision attribute. The lower the entropy value, the lower the level of diversity of data. Conversely, the higher the entropy value, the higher the diversity value.

$$Entropy(S) = \sum_{i=n}^{n} -P_i * Log_2(P_i) \quad (2)$$

S = Sampling
n = Number of partitions S
P_i = The proportion against S

Step 2 : Gain (S, A) - The gain value serves to measure the efficiency of each characteristic in classifying data.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (3)$$

S = Sampling
A = Attribute
n = Number of attribute partitions S
$|S_i|$ = Number of cases on the partition to i
$|S|$ = Number of cases in S

Step 3 : Split Information is a new term before calculating the gain ratio, with the following formula [21]:

$$SplitInformation = -\sum_{t=1}^{c} \frac{|S_i|}{|S|} \ Log_2 \ \frac{|S_i|}{|S|} \quad (4)$$

Step 4 : Then calculate the gain ratio with the following formula:

$$SplitInformation = \frac{Gain(S.A)}{SplitInformation(S.A)} \quad (5)$$

Step 5 : Repeat the second step until all records are complete, where the process ends when each record has the same N class, the record has no attributes that must be partitioned, and the branch is empty [22].

C. Result and Discussion

This study aims to find predictions of prospective scholarship recipients using the K-Means and C4.5 algorithms, where the findings from this study are expected to help educators get predictions during the selection process. In this section, the researchers used data from applicants in 2022, with as many as 1298 participants. This data will be processed through Knowledge Discovery in Database (KDD). Then after the data is ready to be processed at the data mining stage, K-Means and the C4.5 algorithm process will be applied at a later stage.

Table 1. Data Set (Transformation Process)

Initials	Social Welfare Integrated Data Status	Indonesia Smart Card	Prosperous Family Card	Father status	Mother Status	Income Amount
A-1	100	0	0	50	50	100
A-2	0	0	0	75	50	40
...	...	...	...	...	...	...
A-1297	0	0	0	50	50	40
A-1298	0	100	100	50	50	80

The first mining process employs the k-means algorithm to reduce data into groups or clusters of all data.

Step 1 : Determine how many clusters there are in the dataset, which is three.

Step 2 : Determine the central value (centroid).

Table 2. Initial Centroid Value

Initials	Social Welfare Integrated Data Status	Indonesia Smart Card	Prosperous Family Card	Father status	Mother Status	Income Amount
C1	100	100	100	100	50	100
C2	0	0	0	50	50	0
C3	0	0	0	100	50	100

Step 3 : Determine the object's closest proximity to the centroid (Euclidean Distance).

$$C(1.1) = \sqrt{\begin{matrix}(a_1 - c_1)^2 + (b_1 - c_1)^2 + (c_1 - c_1)^2 \\ +(d_1 - c_1)^2 + (e_1 - c_1)^2 + (f_1 - c_1)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100 - 100)^2 + (0 - 100)^2 + \\ (0 - 100)^2 + (50 - 100)^2 + \\ (50 - 50)^2 + (100 - 100)^2\end{matrix}}$$

$$= 150$$

Do it until the last dataset (C 1.1298)

$$C(2.1) = \sqrt{\begin{matrix}(a_1 - c_2)^2 + (b_1 - c_2)^2 + (c_1 - c_2)^2 \\ +(d_1 - c_2)^2 + (e_1 - c_2)^2 + (f_1 - c_2)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-0)^2+(0-0)^2+\\(0-0)^2+(50-50)^2+\\(50-50)^2+(100-0)^2\end{matrix}}$$

$$= 141{,}42$$

Do it until the last dataset (C 2.1298)

C (3.1) $$= \sqrt{\begin{matrix}(a_1-c_1)^2+(b_1-c_1)^2+(c_1-c_1)^2\\+(d_1-c_1)^2+(e_1-c_1)^2+(f_1-c_1)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-0)^2+(0-0)^2+\\(0-0)^2+(50-100)^2+\\(50-50)^2+(100-100)^2\end{matrix}}$$

$$= 111{,}8$$

Do it until the last dataset (C 3.1298)

Step 4 : Sort items by their distance from the nearest Centroid.
The results of calculating the data at the centroid center point for each existing cluster are shown below.

Table 3. Centroid Calculation Results (Each Cluster) And Shortest Distance

Initials	C1	C2	C3	Shortest Distance
A-1	150	141,42	111,8	C3
A-2	185	47,17	65	C2
A-3	181,38	80	53,85	C3
…	…	…	…	…
A-1297	190,00	40,00	78,1	C2
A-1298	113,58	162,48	151,3	C1

The following is the result of the first iteration of calculating the data distance to the cluster center point.

Table 4. First Iteration Cluster Results

Cluster	Results
C1	317
C2	239
C3	742

Step 5 : Repeat steps 3–4 until the centroid is optimal.
Calculate the new center point based on the results of each cluster member.

$$\text{C1 x1} = \frac{(100+100+100+100+100+\cdots n)}{317} = 98{,}74$$

$$C1\ x6 = \frac{(100+100+100+100+\cdots n)}{317} = 98{,}11$$
$$C2\ x1 = \frac{(0+0+0+0+0+\cdots n)}{239} = 1{,}67$$
$$C2\ x6 = \frac{(40+60+40+40+60+\cdots n)}{239} = 49,\ 54$$
$$C3\ x1 = \frac{(100+0+0+100+0+\cdots n)}{742} = 29{,}92$$
$$C3\ x6 = \frac{(100+80+100+100+\cdots n)}{742} = 91{,}78$$

Table 5. New Centroid Values

Initials	Social Welfare Integrated Data Status	Indonesia Smart Card	Prosperous Family Card	Father status	Mother Status	Income Amount
C1	98,74	98,11	17,03	57,02	51,26	98,11
C2	1,67	2,93	0,42	50,73	50,42	49,54
C3	29,92	2,29	0,13	60,11	52,70	91,78

Next, do a calculation of the new centroid value with the data set value and calculate the shortest distance of the data in the second iteration. And calculate the data distance to the cluster center point in the second iteration. The iteration process is carried out until the centroid value resulting from the previous iteration has the same value or the centroid value is optimal (does not change); then, the iteration process stops.

After repeated iterations, the process stops at the fifth iteration and produces the following cluster result values:

Table 6. Fifth Iteration Cluster Results

Cluster	Results
C1	327
C2	739
C3	232

In the next stage, the results of the first data mining will be processed using the c4.5 algorithm to create a decision tree from the data that has been processed.

Step 1 : Look for the Entropy (S) value for each criterion value. By calculating the Entropy and Gain values, create the first root node.
Step 2 : Determine the Gain value (S, A) for each value on the smart Indonesia card Attribute.

Table 7. Node 1 Decision Tree Root Calculation

Node		Information	Number of Cases (S)	C0	C1	C2	Entropy	Information Gain	Split Information	Gain Ratio
1	**Total**		1298	327	739	232	1,41			
	Social Welfare Integrated Data Status							0,90	0,98	0,91
		100	539	307	0	232	0,99			
		0	759	20	739	0	0,18			
	Indonesia Smart Card							0,77	0,82	**0,94**
		100	335	327	8	0	0,16			
		0	963	0	731	232	0,80			
	Prosperous Family Card							0,06	0,26	0,23
		100	56	48	2	6	0,71			
		0	1242	279	737	226	1,38			
	Father status							0,01	0,82	0,01
		100	167	43	89	25	1,46			
		75	62	7	44	11	1,15			
		50	1069	277	606	186	1,41			
	Mother Status							0	0,24	0,01
		100	50	8	32	10	1,30			
		50	1248	319	707	222	1,41			
	Income Amount							0,36	1,68	0,21
		100	744	303	218	223	1,57			
		80	296	13	277	5	0,40			
		60	165	10	152	3	0,46			
		40	65	0	65	0	0,00			
		20	23	0	22	1	0,26			
		0	5	0	5	0	0,00			

Step 3 : Making nodes with branching depending on the greatest Gain value.

Because the highest gain value is the smart Indonesia card Attribute, the initial node is formed from the smart Indonesia card Attribute. There are two criteria for the smart Indonesia card Attribute, namely 100 and 0. Then we will look for nodes from the following two criteria.

Figure 2. Initial root node decision tree

Next, do the process of calculating node 1.1, namely smart Indonesia card criteria with a value of 100.

Table 8. Node 1.1 Decision Tree Root Calculation (Smart Indonesia Card - 100)

Node		Information	Number of Cases (S)	C0	C1	C2	Entropy	Information Gain	Split Information	Gain Ratio
1	**Total**	**100**	335	327	8	0	0,16			
	Social Welfare Integrated Data Status							0,09	0,41	0,22
		100	307	307	0	0	0			
		0	28	20	8	0	0,86			
	Prosperous Family Card							0,01	0,59	0,01
		100	48	48	0	0	0			
		0	287	279	8	0	0,18			
	Father status							0	0,70	0
		100	44	43	1	0	0,16			
		75	7	7	0	0	0			
		50	284	277	7	0	0,17			
	Mother Status							0	0,16	0,01
		100	8	8	0	0	0			
		50	327	319	8	0	0,17			
	Income Amount							0,11	0,32	**0,34**
		100	303	303	0	0	0			
		80	14	14	0	0	0			
		60	18	10	8	0	0,99			
		40	0	0	0	0	0			
		20	0	0	0	0	0			
		0	0	0	0	0	0			

After getting the results from node 1.1, then do the same process on node 1.2, namely the smart Indonesia card criteria with a value of 0.

Table 9. Node 1.2 Decision Tree Root Calculation (Smart Indonesia Card - 0)

Node		Information	Number of Cases (S)	C0	C1	C2	Entropy	Information Gain	Split Information	Gain Ratio
1	**Total**	**0**	963	0	731	232	0,80			
	Social Welfare Integrated Data Status							0,80	0,80	**1**
		100	232	0	0	232	0			
		0	731	0	731	0	0			
	Prosperous Family Card							0,01	0,07	0,10
		100	8	0	2	6	0,81			
		0	955	0	729	226	0,79			
	Father status							0	0,86	0
		100	123	0	88	35	0,86			
		75	55	0	44	11	0,72			
		50	785	0	599	186	0,79			
	Mother Status							0	0,26	0
		100	42	0	32	10	0,79			
		50	921	0	600	222	0,80			
	Income Amount							0,27	1,03	0,26
		100	441	0	218	223	1			
		80	282	0	277	5	0,13			
		60	147	0	144	3	0,14			
		40	65	0	65	0	0			
		20	23	0	22	1	0,26			
		0	5	0	5	0	0			

After calculating at node 1.1 and node 1.2, the highest gain value is obtained, namely Total Income at node 1.1 and social welfare integrated data status at node 1.2. Then continue the decision tree that was created earlier.

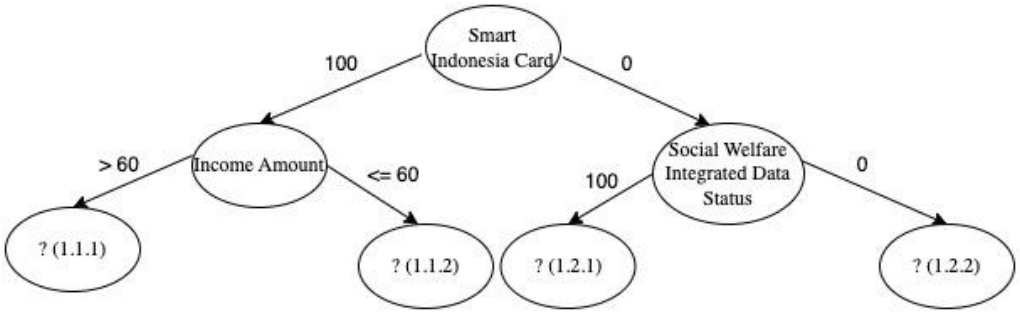

Figure 3. The root node of the next decision tree

Step 4 : Repeat the process for each of the existing branches.

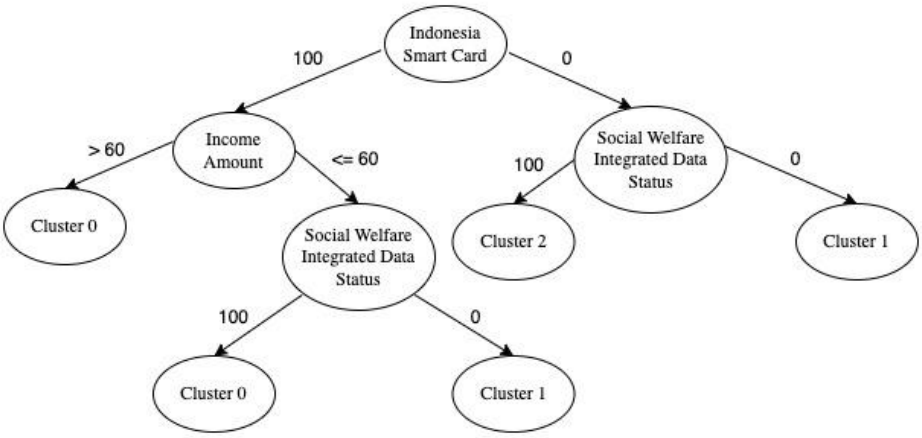

Figure 4. The root node of the final decision tree

In the next process, we will process using the rapid miner studio.

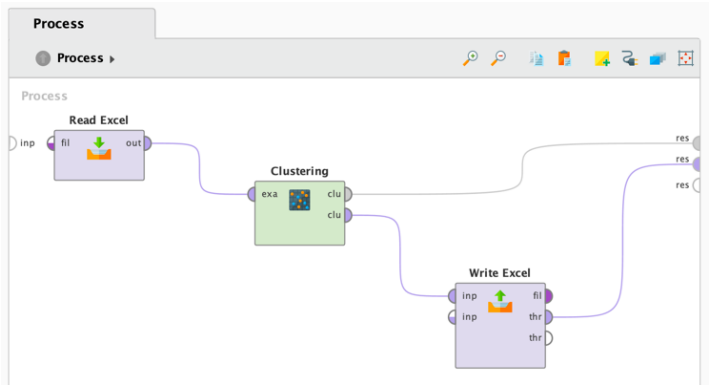

Figure 5. Process the K-Means Algorithm in Rapidminer Studio

a) Select the Excel Read operator to export an existing dataset.
b) To perform the K-Means method, choose the K-Means operator (Clustering).
c) When executing the k-means algorithm, use the Write Excel operator to import the results.

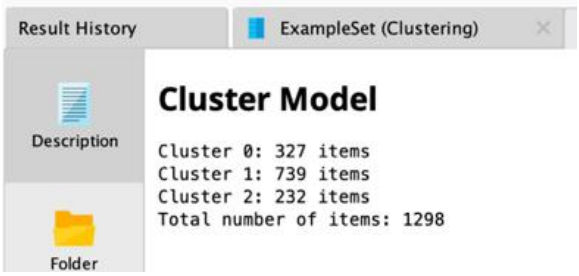

Figure 6. K-Means Algorithm Results in Rapidminer Studio

Furthermore, the C4.5 method will be used in Rapidminer Studio to process the K-Means algorithm results from the imported dataset.

a) Select the Excel Read operator to export the data set.
b) In order to perform the C4.5 algorithm, choose the Decision Tree operator.

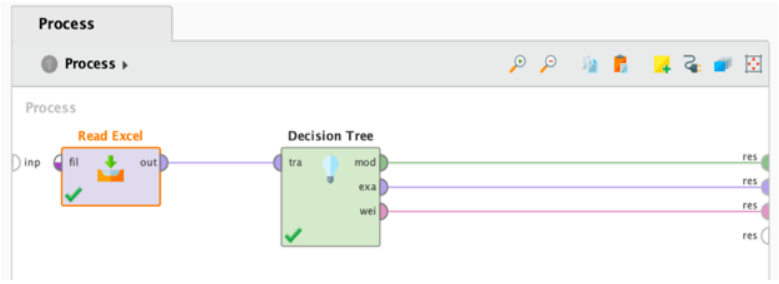

Figure 7. Process the C4.5 Algorithm in Rapidminer Studio

c) The results of the decision tree from the rapid miner application can be used as a reference in making decisions.

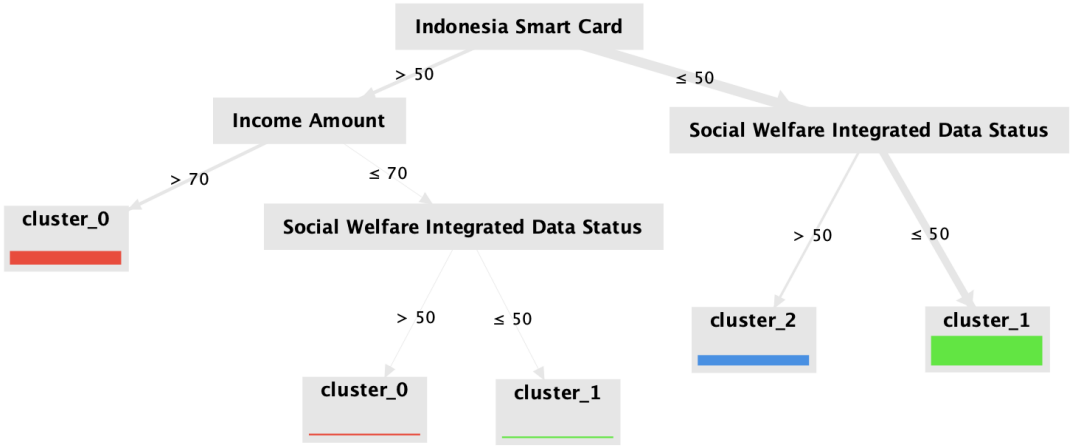

Figure 8. Decision Tree of Algorithm Processing Results from C4.5

Tree

```
Indonesia Smart Card > 50
|   Income Amount > 70: cluster_0 {cluster_2=0, cluster_1=0, cluster_0=317}
|   Income Amount ≤ 70
|   |   Social Welfare Integrated Data Status > 50: cluster_0 {cluster_2=0, cluster_1=0, cluster_0=10}
|   |   Social Welfare Integrated Data Status ≤ 50: cluster_1 {cluster_2=0, cluster_1=8, cluster_0=0}
Indonesia Smart Card ≤ 50
|   Social Welfare Integrated Data Status > 50: cluster_2 {cluster_2=232, cluster_1=0, cluster_0=0}
|   Social Welfare Integrated Data Status ≤ 50: cluster_1 {cluster_2=0, cluster_1=731, cluster_0=0}
```

Figure 9. Description of the Processed Decision Tree Algorithm C4.5

D. Conclusion

Processing data using two data mining algorithms produce findings that can help institutions make decisions about whether participants deserve a scholarship or not. From the processing that was carried out, there were 1289 participants then grouped into 3 clusters, and the results obtained were that 327 participants were in cluster 0, where this cluster was occupied by students who had high point scores. The decision tree made shows the following pattern: If the participant has a Indonesia Smart Card and the value obtained from the "Income Amount" criterion is greater than 70 points, then the participant concerned is entitled to receive a scholarship, and the university only needs to select participants from the results obtained according to the quota set for the tertiary institution.

E. Acknowledgment

The author wishes to express gratitude to the Islamic University of Riau Directorate of Research and Community Service (DPPM) for supporting this activity internally.

F. References

[1] M. D. V. Elvira, I. Muda, and A. Suharyanto, "Implementation of The Regulation Indonesian Ministry of Education and Culture Number 10 Of 2020 Concerning The Indonesia Pintar Program at SMAN 4 Kisaran in The Asahan District," *Strukturasi: Jurnal Ilmiah Magister Administrasi Publik*, vol. 4, no. 1, pp. 87–95, 2022, doi: 10.31289/strukturasi.v4i1.1187.

[2] Edrial, R. A. Putrama, and A. Sujastiawan, "Evaluasi Kebijakan Program Indonesia Pintar (PIP) di SMA Negeri 1 Utan Tahun 2019-2020," *Jurnal Kapita Selekta Administrasi Publik (JKSAP)*, pp. 109–116, 2022, [Online]. Available: http://e-journallppmunsa.ac.id/index.php/ksap

[3] Kemendikbud Ristek RI, "Pedoman Pendaftaran Kartu Indonesia Pintar Kuliah (KIP Kuliah) 2021," 2021.

[4] H. Hendri, "Implementasi Data Mining Dengan Metode C4.5 Untuk Prediksi Mahasiswa Penerima Beasiswa," *Indonesian Journal of Computer Science Attribution-ShareAlike*, vol. 4, no. 2, pp. 2021–312, 2021.

[5] Priati and A. Fauzi Sistem Informasi, *Data Mining dengan Teknik Clustering Menggunakan Algoritma K-Means pada Data Transaksi Superstore*. 2017. [Online]. Available: http://community.tableau.com.

[6] E. Hasmin and S. Aisa, "Penerapan Algoritma C4.5 Untuk Penentuan Penerima Beasiswa Mahasiswa Application of C4.5 Algorithm For Determining Student Scholarship Recipients," *Cogito Smart Journal |*, vol. 5, no. 2, 2019.

[7] D. Rajeshinigo, J. Patricia, and A. Jebamalar, "Accuracy Improvement of C4.5 using K means Clustering," *International Journal of Science and Research (IJSR) ISSN*, vol. 6, no. 6, pp. 2755–2758, 2017, [Online]. Available: www.ijsr.net

[8] C. Anam and H. B. Santoso, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa," 2018.

[9] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," *IEEE Access*, vol. 10. Institute of Electrical and Electronics Engineers Inc., pp. 72480–72503, 2022. doi: 10.1109/ACCESS.2022.3188767.

[10] N. Salsabila, N. Sulistiyowati, and T. N. Padilah, "Pencarian Pola Pemakaian Obat Menggunakan Algoritma FP-Growth," 2022. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] M. Isra, "Behavior Analysis and Prediction of Civil Services Staff in Occupational Functional Positions Using C4.5 Algorithm," *Jurnal Informasi dan Teknologi*, pp. 58–63, Feb. 2022, doi: 10.37034/jidt.v4i1.186.

[12] W. T. Wu *et al.*, "Data mining in clinical big data: the frequently used databases, steps, and methodological models," *Military Medical Research*, vol. 8, no. 1. BioMed Central Ltd, Dec. 01, 2021. doi: 10.1186/s40779-021-00338-z.

[13] N. Azwanti, "Algoritma C4.5 Untuk Memprediksi Mahasiswa Yang Mengulang Mata Kuliah (Studi Kasus Di Amik Labuhan Batu)," *Jurnal SIMETRIS*, vol. 9, no. 1, 2018.

[14] T. Rak and R. Żyła, "Using Data Mining Techniques for Detecting Dependencies in the Outcoming Data of a Web-Based System," *Applied Sciences (Switzerland)*, vol. 12, no. 12, Jun. 2022, doi: 10.3390/app12126115.

[15] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency," in *Journal of Physics: Conference Series*, Jan. 2021, vol. 1751, no. 1. doi: 10.1088/1742-6596/1751/1/012038.

[16] N. Yogeesh, "Mathematical Approach to Representation of Locations Using K-Means Clustering Algorithm," *International Journal of Mathematics And its Applications*, vol. 9, no. 1, pp. 127–136, 2021, [Online]. Available: http://ijmaa.in/

[17] M. Wahyudi, M. Masitha, R. Saragih, and S. Solikhun, *Data Mining: Penerapan Algoritma K-Means Clustering dan K-Medoids Clustering*. Yayasan Kita Menulis, 2020.

[18] E. Indra, K. Ho, Arlinanda, R. Hakim, D. Sitanggang, and O. Sihombing, "Application of C4.5 Algorithm for Cattle Disease Classification," in *Journal of Physics: Conference Series*, Sep. 2019, vol. 1230, no. 1. doi: 10.1088/1742-6596/1230/1/012070.

[19] M. A. Muslim, A. J. Herowati, E. Sugiharti, and B. Prasetiyo, "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease," in *Journal of Physics: Conference Series*, Apr. 2018, vol. 983, no. 1. doi: 10.1088/1742-6596/983/1/012062.

[20] R. Jothikumar and S. R. Balan, "C4.5 Classification Algorithm with Back-Track Pruning for Accurate Prediction of Heart Disease," *Biomedical Research*, 2016, [Online]. Available: www.biomedres.info

[21] A. Purwanto *et al.*, "ANALISA PERBANDINGAN KINERJA ALGORITMA C4.5 DAN ALGORITMA K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENERIMA BEASISWA," 2023. [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

[22] G. Sonia, A. Indriyani, and P. I. Komputer, "Analisis Penerapan Data Mining Dengan Metode Algoritma C4.5 Untuk Pendataan Karyawan Tetap Di Koni Sumatera Utara Analysis Of The Application Of Data Mining Using The C4.5 Algorithm Method For Data Collection On Permanent Employees In North Sumatera Coni," *JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING (CoSIE)*, vol. 02, no. 1, p. 2023, 2023, [Online]. Available: http://creativecommons.org/licenses/by-sa/4.0/