Hybrid Data Mining with the Combination of K-Means Algorithm and C4.5 to Predict Student Achievement

by Agung Ramadhanu

Submission date: 25-Jun-2022 11:26AM (UTC+0700)

Submission ID: 1862598447

File name: of_K-Means_Algorithm_and_C4.5_to_Predict_Student_Achievement.pdf (468.8K)

Word count: 5017

Character count: 25019

Hybrid Data Mining with the Combination of K-Means Algorithm and C4.5 to Predict Student Achievement

Agung Ramadhanu[a,1,*], Sarjon Defit[b,2] Shahab Wahhab Kareem[c]

[a, b] Universitas Putra Indonesia YPTK Padang, Jl. Raya Lubuk Begalung, Padang, 25221, Indonesia
[c]Erbil Polytechnic university Iraq
[1] agung_ramadhanu@upiyptk.ac.id; [2] sarjond@yahoo.co.uk*
** corresponding author*

ARTICLE INFO

Article history:
Received 26 May 2021
Revised : 07 Sept 2021
Accepted : 29 Oct 2021

Keywords:
Hybrid Data Mining
K-Means Algorithm
C4.5 Algorithm
Student Achievement

ABSTRACT

Getting academic achievement is the dream of every student who studies at higher education, especially undergraduate level. Undergraduate students aspire to the highest achievement (champion) at the last achievement of their studies. However, students cannot predict whether these students with the habits that have been done and the current conditions will make them excel or not. Apart from that, of course, students also want to know what factors and conditions influence the achievement the most. The objective to be achieved in this research is how to predict which number of students among them are predicted to excel (champion) at the end of the semester with a combination of the K-Means and C4.5 methods. Besides, the purpose of this study reveals how the K-Means algorithm performs data clustering of student data who will excel or not and how the C4.5 algorithm predicts students who have been grouped. Data processing in this study uses the Rapid Miner software version 9.7.002. The result of this research is that it is easier to group data in numerical form than data in polynomial form. Other results in this study were that out of 100 students, 27 students (27%) were predicted to excel (winners) and 73 (73%) did not achieve (not winners).

I. Introduction

A country is said to be a developed country which is indicated by the education in the country is developed or not. If education is not a special concern for a country, that country will not become a developed country. Education greatly influences other fields. For example in finance, defense, government, and so on. Education is also very important for the sustainability of the country in the future because the children who are the successor of the country are not well prepared so that it threatens the sustainability of the country in the future [1]–[3]. Education is a basic right for every citizen of all over the world, including Indonesia. The state must provide institutions that provide education. The implementation of this education is held at various levels from elementary, middle, and high levels. Primary level education is education that must first be pursued then followed by secondary education and then higher education. Higher education is also often known as higher education [4], [5].

Higher education is the highest educational institution in Indonesia. Everyone who wants to continue their education in higher education must complete their previous education, namely high school education. In higher education, students are called students, and educators are called lecturers [6], [7]. Higher education consists of various forms including Academies, Polytechnics, Institutes, Colleges, and Universities. Each type of tertiary education institution has different educational strata, namely D3, S1, S2, and S3. Every student who takes education certainly wants to get an achievement. Achievement is a matter of pride for the students themselves, the family, and the institution. This includes students as long as they take their undergraduate education (S1), which is 4 years when a student wants to get an achievement. The achievements can be in academic and non-academic forms [8], [9]. Besides, achievements can also be obtained on campus and outside the campus.

Academic achievements on campus are the main achievements that students must obtain. This achievement is like a class champion. Every parent who finances their children to study in tertiary institutions expects their children to get the highest achievement such as champions in class every semester. This achievement is also an indicator that indicates the success of a student in pursuing education [10]. To achieve a champion's achievement is certainly not as easy as turning your palm. But it needs a very big sacrifice, be it time, effort thought, and cost. Many factors can influence students to get achievements. Can be parents' income, distance from the house to campus, number of siblings, completeness of learning tools, student socialization in the campus environment, student socialization in the home environment, GPA, and so on [11], [12].

Many calculation methods can be used to predict whether a student will get an achievement at the end of the semester or not. Predictions are made based on predetermined factors and the value of the student is based on these factors. Two methods are often used, namely the K-Means method [13] and the C4.5 method. The K-Means method is used to classify student data into several groups or clusters while the C4.5 method is used to make predictions. The combination of the two methods will produce a more precise and accurate output [14]. The goal of this research is how to predict which number of students among them are predicted to excel at the end of the semester with a combination of the K-Means and C4.5 methods. Besides, the purpose of this study reveals how the K-Means algorithm performs data clustering [15] data on students who will excel or not and how the C4.5 algorithm predicts students who have been grouped [16].

The campus which is carried out as the object of this research is the students of the Putra Indonesia University YPTK Padang. This campus was chosen because there was a policy from the campus that outstanding students would be given an award in the form of free tuition for the next semester and if in the next semester they also excel, they would be allowed to study comparative studies in Malaysia and Singapore. Many previous studies have been carried out with the K-Means and C4.5 methods, including by Wiga Maulana Baihaqi who researched Predicting Sara's Elements in Tweets, further research opportunities are for improvement in the proposed method to obtain more accurate results, both in grouping and classification. Twitter data that contains SARA elements and does not contain SARA elements. Besides. Therefore, it is best to build a system that can be applied to analyze Twitter content.

II. Methodology

A. Research Framework

In conducting this research the authors followed the research framework that had been prepared. The following is the research framework that the authors conducted:

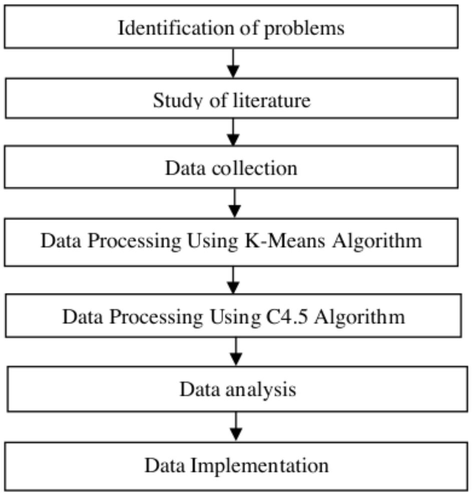
Identification of problems
Study of literature
Data collection
Data Processing Using K-Means Algorithm
Data Processing Using C4.5 Algorithm
Data analysis
Data Implementation

Fig. 1. Research Framework

B. Research Framework Details

Problem identification is the process of finding out what are the important and main problems in the field that must be solved in this research. Literature study is the process of seeking or studying previous sciences and research related to solving previously identified problems. Data collection is the process of collecting field data that will be used in problem-solving. In this study, the data were sourced from the Academic Bureau (PDE and SISFO) of Putra Indonesia University YPTK Padang and filling out the questionnaire by students. The number of test data records to be tested that have been collected is 50 data records or 50 students. The data collected consists of 8 data attributes as in table 1:

Table 1. Collected Data Attributes

No	Data Attributes	Data Type
1	BP Number	Text
2	Name	Text
3	Parents' Income	Integer
4	Home (Boarding) distance to campus	Integer
5	Complete Learning Tools	Integer
6	Student socialization with the campus environment	Integer
7	Student socialization with their living environment	Integer
8	Grade Point Average (GPA)	Integer

Data that has been collected in the field is processed using the K-Means method. The following are the steps for the K-Means method [17], [18], namely:

1. Determine how many groups (clusters) will be created which is called the value k
2. Determine the mean (centroid) random value (random) for each predetermined cluster.
3. Determine the nearest cluster center on each data record with the centroid value using the formula:

   $$d_{Euclidean}(x,y) = \sqrt{\sum(x_i - y_i)^2} \quad (1)$$

   Information: $d_{Euclidean}(x,y)$ = the distance value for each record with the centroid value, $x = x1, x2, x3, etc, y = y1, y2, y3, etc$
4. Determine the closest cluster for each data record by comparing the closest distance value that has been obtained previously and then updating the cluster center value (centroid) using the formula:

   $$Cluster\ Center = \sum \frac{a_i}{n} \quad (2)$$

   Information: $Cluster\ Center$ = The cluster center value, a_i = Value on each cluster, n = Number of $cluster$
5. Repeating steps 3 to 5 until there is no data transfer from one cluster to another.

The data that has been collected in the field is not only processed by the K-Means method but also processed using the C4.5 method. Below are the steps for the C4.5 method [19], [20], namely:

1. Select data attributes that will be used as root or prediction nodes in the decision tree and calculate the number of YES and NO values for each data record.
2. Make a branch for each value after obtaining the root of the decision tree by calculating the Gain value using the following formula:

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

   Information: Gain (S, A) = total gain value with attributes, Entropy (S) = total entropy value, Entropy (S_i) = Entropy value for each attribute, n = number of clusters
   While the formula for calculating the entropy value is:

   $$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\, pi \quad (4)$$

   Information: $Entropy(S)$ = total entropy value, pi = proportion of S_i to S
3. Divide cases into branches
4. Repeat the process for each branch until all cases on the branch have the same class.

The data that has been processed and generated using the K-Means and C4.5 methods are then analyzed and conclusions are drawn from the results of the analysis. How much data from the clustering results and who is included in the cluster have been processed using the K-Means method [21] then the clustering data is analyzed so that it is known which students are predicted to excel (winners) and who It is predicted that they will not perform well (not winning) which have been processed using the C4.5 method [22] among all the processed data. Data processing uses Rapid Miner software version 9.7.002.

The data that has been analyzed and concluded will then be implemented in the field which will later be useful and very helpful for an institution, especially an educational institution in predicting that its students will excel (win) or not (not win).

III. Result and Discussion

A. *Results of Clustering Data Processing Using the K-Means Algorithm*

The initial data that has been collected consisting of 8 attributes and 50 data records can be seen in Table 2 below:

Table 2. Initial Data

No	BP Number	Name	Parents' Income (Rp./mountly)	Home (Boarding) distance to campus (Km)	Complete Learning Tools	Socialization with the campus environment	Socialization with the neighborhood	Grade-Point Average (GPA)
1	18101152600003	ALDO RAZZAQ	10.000.000	4	3	1	0	4,00
2	18101152600007	CHAIRUL HAMID	6.000.000	6	2	0	1	3,65
3	18101152600018	STEFANI PRATIWI	8.000.000	5	3	0	0	3,65
4	18101152600044	FENTI PURWANTI	8.000.000	5	3	1	1	3,80
5	18101152600050	PUTRI WAHYUNI	6.000.000	1	2	1	1	3,70
….	…	…	…	…	…	…	…	…
50	16101152610036	REZA FAHLEFI	10.000.000	5	3	0	1	4,00

Note:

- Number 1 in the learning completeness attribute means that the student has incomplete learning tools, number 2 means having complete learning tools, and number 3 means having very complete learning tools.
- The number 0 in the attributes of the student's socialization with the campus environment and the environment in which they live means that the student is not active socializing. Number 1 means that the student is actively socializing.

After processing data on the data in table 2 above with the K-Means method using the Rapid Miner application (Mardalius, 2018), data clustering is obtained. Data clustering is carried out on all data attributes, namely attributes of parents' income, distance from house to campus, completeness of learning tools, socialization of students with the campus environment, socialization of students with their living environment, and cumulative grade point average (GPA). Before grouping, the next step is to convert the data in table 2 into values 1, 2, 3, and 4 using table 3 of the conversion of raw data values below:

Table 3. Converting Raw Data Values

No	Parents' Income			Range Distance	
4	5,100,000	High	4	>6	Very Far
3	3100000 - 5100000	Medium	3	4 until 6	Far
2	990000 - 3100000	Low	2	2 until 4	Close
1	990,000	Very Low	1	<2	Very Close

Below are the results of data conversion based on table 3 above as in table 4 below:

Table 4. Results of Data Conversion

No	BP Number	PO	JAR	AB	SLK	SLT	IPK
M1	18101152600003	4	4	3	1	0	4.00
M2	18101152600007	2	6	2	0	1	3.65
M3	18101152600018	2	5	3	0	0	3.65
M4	18101152600044	4	5	3	1	1	3.80
M5	18101152600050	2	1	2	1	1	3.70
….	….	….	….	….	….	….	….
M50	18101152610602	2	3	1	0	0	3.65

After obtaining the converted data as in table 3 above, the next step is to cluster the data using the K-Means algorithm. The following are the steps for the K-Means algorithm:

a. Step 1. Determine the number of data clusters to be created, the number of data clusters is called the value k. The number of data groups determined in this study derived from the data in table 4 is as many as 2 data clusters, namely cluster 1 and cluster 2.

b. Step 2. Determine the centroid value randomly for each predetermined group. Below is the centroid value for each cluster is shown in table 5 below:

Tabel 5. Nilai Tengah (*Centroid Value*) Hasil *Clustering*

No	Attribute	cluster_0	cluster_1
1	Parents' Income (Rp./mounthly)	2.412	2.060
2	Home (Boarding) distance to campus (Km)	5.941	2.364
3	Complete Learning Tools	2.412	1.939
4	Socialization with the campus environment	0.647	0.515
5	Socialization with the neighborhood	0.471	0.455
6	Grade-Point Average (GPA)	3.792	3.745

c. Step 3. Determine the closest cluster center on each row of data with the centroid value, to determine this value using the formula:

$$d_{Euclidean}(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

$$d_{Euclidean}(x_1, y_1) = \sqrt{(3.33 - 3.1)^2 + (3.27 - 3.1)^2 + \cdots + (3.1 - 3.1)^2} = \sqrt{4.0529} = 2.013$$

….

$$d_{100}(x_{100}, y_{100}) = \sqrt{(3 - 1)^2 + (3 - 1)^2 + \cdots + (3 - 1)^2} = \sqrt{4.0289} = 2.007$$

d. Step 4. Determine the closest cluster for each row of data by comparing the closest distance values that have been obtained in the previous process than updating the center value of the group using the formula:

$$Cluster\ Center = \sum \frac{a_i}{n} = \frac{171.559}{100} = 1.72$$

e. Step 5. Repeating steps 3 to step 5 until there is no transfer of data for each row of data from one group to another.

The following table 6 shows the results of grouping data using the K-Means algorithm:

Table 6 Results of Data Clustering

No	Cluster	Item	Naming
1	Cluster 1	17 Items	Champion
2	Cluster 2	33 Items	Not Champion
3	Total Item	50 Items	-

To produce data as seen in table 3 and table 4 above, of course, requires a data processing block design using the Rapid Miner application. The block design can be seen in Figure 2 below [23]:

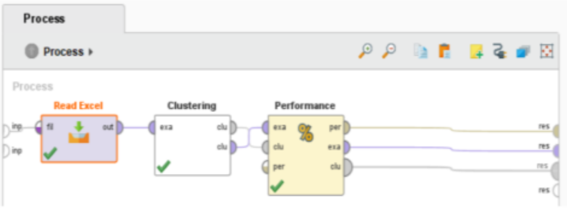

Fig. 2. Block Design Method K-Means Rapid Miner application

So that the data can be seen visually the results of the clustering so that we can find out what the grouping looks like, it can be seen in Figure 3 below:

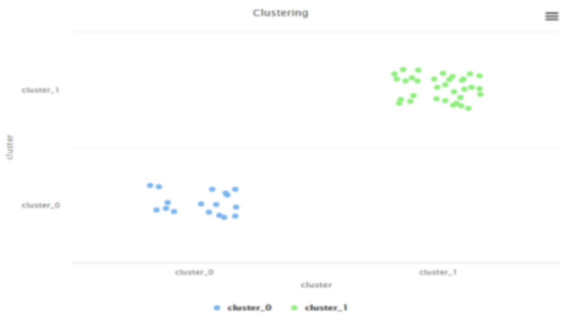

Fig. 3. Visualization Data View

B. Result of Prediction Data Processing Using The C4.5 Algorithm

The data that the researchers collected were 50 rows of data or 50 students who won and did not win the previous year. Data that have been grouped into clusters are given names that describe the type of cluster. In analyzing the data, the C4.5 method. Following in table 7 below the data that has been given the name and the addition of attributes.

Step 1. Determine the data attribute that will be used as the root node or prediction in the decision tree and calculate the number of YES and NO values for each data row. Below, in table 7 the data has been named YES and NO.

Table 7. Naming Data According to Clusters

No	PO	JAR	AB	SLK	SLT	GPA
1	Hight	Far	Complete	Yes	No	4.00
2	Low	Far	Complete	No	Yes	3.65
3	Low	Far	Complete	No	No	3.65
4	Hight	Far	Complete	Yes	Yes	3.80
5	Low	Very Near	Less	Yes	Yes	3.70
….	….	….	….	….	….	….
50	Low	Very Far	Complete	No	Yes	3.86

After naming each data record on each attribute, the prediction results are obtained using the C4.5 method using the Rapid Miner application. Below table 8 shows the prediction results.

Table 8. Prediction Result Data

no		True (YES)	True (NO)	Class Precission
1	pred. NO	1	0	100.00%
2	pred. Champion	0	4	100.00%
3	class recall	100.00%	100.00%	
4	pred. NO	1	0	100.00%

Accuracy 100%

The prediction result data above can be presented in graphical form (plot view) as shown in Fig. 4 below:

ExampleSet

Fig 4. Graph form (plot view) of the predicted data

To produce data as seen in table 7, table 8, and figure 4, the data processing block design using the Rapid Miner application can be seen in Figure 5 below:

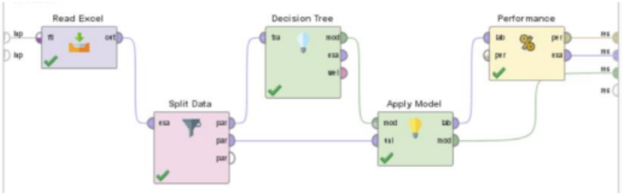

Fig. 5. Block Design Method C4.5 Rapid Miner application

The results of the predictions in Table 8 can be in the form of statistics in the form of data conclusions. Below is the presentation of the data in the form of a description decision tree text view.

Tree

```
Distance from house (boarding) to campus (km) = close
|   Socializing with the campus environment = No
|   |   Socialization with the neighborhood = No
|   |   |   Parents' Income (Rp./mounthly) = Rendah
|   |   |   |   Grade Point Average (GPA) > 3.950: NO {NO=1, NO=0}
|   |   |   |   Grade Point Average (GPA) ≤ 3,950: CHAMPION {NO = 0,
CHAMPION = 4}
|   |   |   Parents' income (Rp./mounthly) = Very Low: NO {NO=1, JUARA=0}
|   |   Socialization with the neighborhood = YES: NO{NO=2, CHAMPION=0}
|   Socialization with the campus environment = Yes: CHAMPION {NO=0,
CHAMPION=8}
Distance from house (boarding) to campus (Km) = Far: NO {NO=12,
CHAMPION=0}
Distance from house (boarding) to campus (Km) = Very Near
|   Grade Point Average (GPA) > 3.660: CHAMPION {NO=0, CHAMPION=9}
|   Grade Point Average (GPA) ≤ 3.660
|   |   Socializing with the campus environment = No: NO {NO=1,
CHAMPION=0}
|   |   Socializing with the campus environment = Yes: CHAMPION {NO=0,
CHAMPION=2}
   Distance from house (boarding) to campus (Km) = Very Far: NO {NO=5,
   CHAMPION=0}
```

The shape of the decision tree from the prediction results can be seen in Figure 6 below:

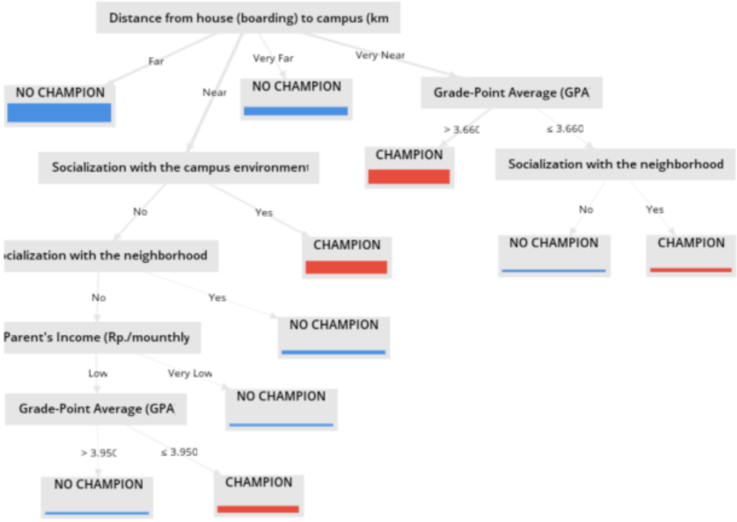

Fig. 6. Decision Tree for Prediction Results

C. *Discussion*

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Based on the results of the K-Means method clustering data processing using the Rapid Miner application according to the data in table 3, it can be seen that from a total of 50 student data into several clusters based on their respective attributes. This was done so that the data could be processed easily into predictive data using the C4.5 method. using the RapidMiner application. In Fig. 2 is a design view, which is a display block design of clustering data processing for each attribute. Each data attribute is clustered using the K-Means method in the Rapid Miner application. From the far left is the read excel block, then the clustering block using the C4.5 method, then the performance block to see the data capabilities.

Based on the results of processing the predictive data using the C4.5 method using the Rapid Miner application according to the data in table 5, it can be seen that from a total of 50 student data that became testing data, there were 17 students who would excel (Champion) and 17 people who would not with achievement (Champion) there are as many as 33 people who have NO. Thus the predicted achievement (champion) was 34% and the predicted non-achievement (not champion) was 66% of the 50 students who became the testing data. In Fig. 5, we can see the decision tree on the results of the predictions that have been made. The image is obtained from the Rapid Miner menu graph view application. In the decision tree, it can be seen that the root (root) of the decision is the attribute of the cumulative achievement index (GPA) then followed by the attribute of the distance from the house (boarding house) to the campus, then the attributes of the parent's income, then the attribute of the number of siblings, then the attributes of learning tools completeness, attribute of student socialization with the campus environment and finally student socialization with the environment in which they live.

IV. Conclusion and Suggestion

Based on the research that has been done, it can be concluded several things including Data Clustering will produce good results if the correct number of k values (number of clusters) is selected. If the number of clusters is too large, the results of the clustering will not be good. In making data predictions, the data should be converted into polynomial data or given a name first according to the data group (cluster) so that the resulting decision tree is easy to see and analyze. The results of the predictions made will be more accurate if the training data entered into the C4.5 method has the same number of records (≥) than the number of data testing records. Of the 100 students whose data were processed, 27 students (27%) were predicted to excel (winners) and 73 students (73%) did not achieve (not winners).

Based on the research that has been done, it can be concluded several things including We recommend that you use more attributes than we have done so that the results of clustering and predicting are better. We recommend that the data attribute for clustering with the K-Means method is in the form of numerical data and then the results are used as polynomial data so that it can be used for the prediction of the C4.5 method. It is recommended that the number of training data records be greater than the number of testing data records so that the predictions made are more precise and accurate.

References

[1] I. Anugraheni, "The Effect of Learning Problem Solving Model Polya on the Ability to Solve Mathematical Problems in Indri Anugraheni Students," *J. Pendidik.*, vol. 4, no. 1, pp. 1–6, 2013.

[2] W. G. Abdisara, S. Patmanthara, and D. U. Soraya, "Contribution of Student Independence and Availability of Infrastructure to Learning Outcomes of Wireless Network Subjects," *J. Pendidik.*, vol. 04, no. 2, pp. 55–62, 2019.

[3] I. Amaliah and E. Sudihartinih, "Development of Multimedia Assisted Fraction Concept Teaching Materials to Improve Students' Mathematical Comprehension Ability in Inclusive Schools," *J. Pendidik.*, vol. 4, no. 2, pp. 6–10, 2019.

[4] E. F. Rusydiyah, E. Purwati, and A. Prabowo, "How to use digital literacy as a learning resource for

teacher candidates in Indonesia," *Cakrawala Pendidik.*, vol. 39, no. 2, pp. 305–318, 2020, doi: 10.21831/cp.v39i2.30551.

[5] S. A. Nulhaqim, D. H. Heryadi, R. Pancasilawan, and M. Ferdryansyah, "The Role of Higher Education in Improving the Quality of Education in Indonesia to Face the 2015 Asean Community Case Study: University of Indonesia, Padjadjaran University, Bandung Institute of Technology," *Share Soc. Work J.*, vol. 6, no. 2, p. 197, 2016, doi: 10.24198/share.v6i2.13209.

[6] S. Haryati, A. Sudarsono, and E. Suryana, "Implementation of Data Mining to Predict Student's Study Period Using the C4.5 Algorithm (Case Study: Bengkulu Dehasen University)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.

[7] L. P. Sinambela, "Lecturer Professionalism and Quality of Higher Education," *Populis*, vol. 2, no. 4, pp. 579–596, 2017.

[8] Y. Yuzarion, "Factors Affecting Students' Learning Achievement," *Ilmu Pendidik. J. Kaji. Teor. dan Prakt. Kependidikan*, vol. 2, no. 1, pp. 107–117, 2017, doi: 10.17977/um027v2i12017p107.

[9] A. Syafi'i, T. Marfiyanto, and S. K. Rodiyah, "Study of Student Achievement in Various Aspects and Affecting Factors," *J. Komun. Pendidik.*, vol. 2, no. 2, p. 115, 2018, doi: 10.32585/jkp.v2i2.114.

[10] R. Rasmawan, "Students' Critical Thinking Skills Profile and Their Correlation with Academic Achievement Index," *EduChemia (Jurnal Kim. dan Pendidikan)*, vol. 2, no. 2, p. 130, 2017, doi: 10.30870/educhemia.v2i2.1101.

[11] M. Shaleh, "The Influence of Motivation, Family Factors, Campus Environment and Active Organizations on Academic Achievement," *Phenom. J. Pendidik. MIPA*, vol. 4, no. 2, pp. 109–141, 2016, doi: 10.21580/phen.2014.4.2.122.

[12] Nurhasanah, Purwati, and H. Ahmad, "The Influence of the College Entrance Selection System on the Achievement Index of Students of the Department of Mathematics Education, University of Papua (UNIPA)," *Pros. Semin. Nas.*, vol. 03, pp. 114–120, 2015.

[13] F. Nur, M. Zarlis, and B. B. Nasution, "Application of the K-Means Algorithm to New Vocational High School Students for Department Clustering," *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 1, no. 2, pp. 100–105, 2017, doi: 10.30743/infotekjar.v1i2.70.

[14] E. Elisa, "Analysis and Application of the C4.5 Algorithm in Data Mining to Identify Factors that Cause PT.Arupadhatu Adisesanti Construction Accidents," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017, doi: 10.15575/join.v2i1.71.

[15] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Application of the K-Means Algorithm for Clustering Drug Data," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[16] R. Rismayanti, "IImplementation of the C4.5 Algorithm to Determine Scholarship Recipients at Stt Harapan Medan," *J. Media Infotama*, vol. 12, no. 2, pp. 116–120, 2017, doi: 10.37676/jmi.v12i2.413.

[17] Jaroji, Danuri, and F. P. Putra, "K-Means To Determine Candidates," *J. Inovtekpolbeng- Seri Inform.*, vol. 1, no. 1, pp. 87–94, 2016.

[18] W. Dhuhita, "Clustering Using the K-Mean Method to Determine the Nutritional Status of Toddlers," *J. Inform. Darmajaya*, vol. 15, no. 2, pp. 160–174, 2015.

[19] A. H. Nasrullah, "Application of the C4.5 Method for Classification of Students with the Potential to Drop Out," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 244–250, 2018, doi: 10.33096/ilkom.v10i2.300.244-250.

[20] N. Azwanti, "C4.5 Algorithm to Predict Students Who Repeat a Course (Case Study at Amik Labuhan Batu)," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 11–22, 2018, doi: 10.24176/simet.v9i1.1627.

[21] M. L. Sibuea and A. Safta, "Mapping Student Achievement Using the K-Means Clustering Method," *Jurteksi*, vol. 4, no. 1, pp. 85–92, 2017, doi: 10.33330/jurteksi.v4i1.28.

[22] T. Tukino, "Application of the C4.5 Algorithm to Predict Profits at PT SMOE Indonesia," *J. Sist. Inf. Bisnis*, vol. 9, no. 1, p. 39, 2019, doi: 10.21456/vol9iss1pp39-46.

[23] A. F. Sallaby and E. Suryana, "Application of Data Mining to Determine the Number of Registered Job Seekers by Age and Education Using K-Means Clustering (Case Study at the Bengkulu Province Manpower and Transmigration Office)," *J. Technopreneursh. Inf. Syst.*, vol. 1, no. 1, pp. 35–38, 2018, doi: 10.36085/jtis.v1i2.28.

Hybrid Data Mining with the Combination of K-Means Algorithm and C4.5 to Predict Student Achievement

ORIGINALITY REPORT

18%	17%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijair.id Internet Source	12%
2	www.atlantis-press.com Internet Source	2%
3	www.researchgate.net Internet Source	1%
4	Dicky Nofriansyah. "Application to Determination of Scholarship Worthiness Using Simple Multi Attribute Rating Technique and Merkle Hellman Method", International Journal of Artificial Intelligence Research, 2017 Publication	1%
5	journal.trunojoyo.ac.id Internet Source	1%
6	Rina Novita, Supratman Zakir, Agus Nur Khomarudin, Efmi Maiyana, Hamimah Hasyim. "Use of the C4.5 Algorithm in Determining Scholarship Recipients", Journal of Physics: Conference Series, 2021 Publication	1%

7	d.researchbib.com Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		