HASIL CEK_Comparison of K-Medoids Method and Analytical Hierarchy

by Lisna Zahrotun, Utaminingsih Linarti, Banu Harli, Herri Kurnia, Liya Yusrina Sabila

Submission date: 11-Aug-2023 02:49PM (UTC+0700)
Submission ID: 2144340814
File name: Comparison_of_K-Medoids_Method_and_Analytical_Hierarchy.pdf (3.67M)
Word count: 5679
Character count: 26138

JOIV : Int. J. Inform. Visualization, 7(2) - June 2023 446-454

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Comparison of K-Medoids Method and Analytical Hierarchy Clustering on Students' Data Grouping

Lisna Zahrotun [a,*], Utaminingsih Linarti [b], Banu Harli Trimulya Suandi As [a], Herri Kurnia [a], Liya Yusrina Sabila [c]

[a] *Informatics Department, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[b] *Industrial Engineering Department, Faculty of Industrial Technology, Universitas, Ahmad Dahlan, Yogyakarta, Indonesia*
[c] *Electrical Engineering Department, Faculty of Industrial Technology, Universitas, Ahmad Dahlan, Yogyakarta, Indonesia*

*Corresponding author: *lisna.zahrotun@tif.uad.ac.id*

***Abstract*— One sign of how successfully the educational process is carried out on campus in a university is the timely graduation of students. This study compares the Analytic Hierarchy Clustering (AHC) approach with the K-Medoids method, a data mining technique for categorizing student data based on school origin, region of origin, average math score, TOEFL, GPA, and length study. This study was carried out at University X, which contains a variety of architectural styles. The R department, the S department, the T department, and the U department make up one of them. K-Medoids and AHC techniques Utilize the number of clusters 2, 3, and 4 and the silhouette coefficient approach. The evaluation's findings indicate a value. Although there is a linear silhouette between the AHC and K-Medoids methods, the AHC approach (departments R: 0.88, S: 0.87, T: 0.88, and U: 0.88) has a more excellent Silhouette value than K-Medoids (department R: 0.35, department S: 0.65 number of cluster 2, department T: 0.67 number of cluster 2 and program Study U: 0,52). The results of the second approach, which includes the K-Medoids and AHC procedures, are determined by the data distribution to be clustered rather than by the quantity of data or clusters. Based on this methodology, University X can refer to the grouping outcomes for the four departments with two achievements to receive results on schedule.**

***Keywords*— Grouping; K-medoids; silhouette coefficient; analytical hierarchy clustering.**

Manuscript received 12 Sep. 2022; revised 8 Dec. 2022; accepted 20 Feb. 2023. Date of publication 30 Jun. 2023.

I. Introduction

The main metric used in universities to assess academic achievement is GPA [1]–[3]. The average grade received in all courses taken from the first semester to the last is known as the GPA. So, each semester's GPA will be determined. The GPA will be determined between the first and last semesters, though [4], [5]. Each student's GPA impacts how successful or successful they become at a university. Graduation rates or length of study are further indicators of a university's success besides GPA. University X has various faculties, and faculty Y is one of them. The R department, S department, T department, and U department are the four departments that make up Faculty Y. There are many students at Faculty Y. All departments in Faculty Y see a growth in the number of new students each year. This impacts a university's performance or success since there is an imbalance between new students and students who are graduating, which leads to poor evaluations. The ratio of instructors to students is still relevant because some factors, including the imbalance in the number of lecturers and students, can contribute to this imbalance. Another concern is the sheer volume of pupils, which will restrict how much time can be spent in places like labs. Analyzing student achievement is essential to figuring out how effective the current educational system [6]. Student graduation data, such as period of study, TOEFL, and GPA, with new student data, such as report score, school origin, and district origin, have never been combined together before now. The Faculty of Y will consider this, especially when choosing new students, by grouping the data from each department.

Clustering is a technique in data mining. Clustering is a method for organizing data that can be utilized [7]–[9]. Based on the values of their attributes, clustering is a technique for identifying homogeneous object groups. Clustering can assist in analyzing the variables influencing student learning results [10], [11]. Data in the form of text or numeric data can be grouped using clustering. Clustering has been employed in texts before [12], [13]. The Analytical Hierarchy Clustering

(AHC) approach and the K-Medoids method are two examples of the numerous techniques that can be applied. The accuracy of the AHP approach in this study's relationship management system research on electronic clients is 66.6%, and it performs better in terms of time complexity [14]. There has been an investigation into the center point using the K-Medoids technique [15]–[17]. In order to group data, this study contrasts the K-Medoids method with the Analytical Hierarchy Clustering (AHC) method. The data used in this study are the school's name, region of origin, math score, and GPA. The Silhouette Coefficient approach was used to conduct the test.

II. Material and Method

This study aims to examine the classification of student data using the K-Medoids method and the analytical Hierarchy Clustering (AHC) approach. The use of clustering approaches to analyze student academic performance has been studied in a number of publications over the past few years [6]. K-Means clustering is used to analyze student learning outcomes and performance [18]–[20] . K-Medoids and Analytical Hierarchy Clustering (AHC) are the methods applied in this study. Partition grouping is done using the K-Medoids approach, which is popular due to its effectiveness, simplicity, and convenience of usage [21]–[24]. The AHC approach, in contrast, uses hierarchical clustering to generate grouping of the individual data points inside a cluster in the shape of a tree [6]. These two grouping techniques, however, are appropriate for data having categorical data types [21]. As a result, the K-Medoids approach and the Analytical Hierarchy Clustering (AHC) method were used in this study.

A. K-Medoids

Data will be taken randomly to be used as central data in the cluster, each data has the opportunity to become central data, but most middle data is used as central data in a cluster based on the conditions of the K-Medoids Algorithm [25]. The steps of the K-Medoids Algorithm are as follows:

- Initialization of cluster centers as much as k (number of clusters).
- Group each data into the closest cluster using the Euclidean Distance approach to calculate the distance between data with the equation (1):

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

Explanation:
$x(i)$ = the first i data.
$y(i)$ = the second i data.
n = amount of data

- Then select the data randomly in each cluster used as a candidate for a new medoid.
- After that, calculate the distance of each data in each cluster with the new medoid candidate.
- Then calculate the total deviation (S) by calculating the new total distance value - the old total distance. If S < 0, replace objects with cluster data to form a new set of k objects as medoids.
- Repeat steps 3 to 5 until there is no medoid change so that clusters and their respective cluster members are obtained.

B. Analytical Hierarchy Clustering (AHC)

This grouping is a Hierarchical Grouping which allows having two main approaches, namely the hierarchical approach and the split approach [21], [26], [27]. AHC Algorithm Steps:

- The distance between data is calculated using the Euclidean formula at this stage. Euclidean Distance Formula (6):

$$\|U - V\| = \sqrt{\sum i}\,(Ui - Vi)^2 \quad (6)$$

where:
Ui = U value on training data
Vi = value of V on test data

- Based on the distance matrix, then the data is grouped using Agglomerative Hierarchical Clustering (AHC) using the single linkage method in equation (7).

$$d_{data} = min\{d_{data}\}, d_{data} \in D \quad (7)$$

where:
d_{data} = the distance between the nearest/smallest neighbor of the data group
D= Euclidean distance proximity matrix

C. Silhouette Coefficient

This method will calculate the level of proximity between data or objects in a cluster. For example, the steps in the silhouette coefficient process [16], [28], [29] are as follows :

- Calculate the average distance from a document, i with all other documents in one cluster [17].

$$a(i) = \frac{1}{[A]-1} \sum j \epsilon C\, d(i,j) \quad (8)$$

where
$a(i)$ = The mean difference of object (i) to all other objects of A
$d(i,j)$ = Distance between data i to j
A = cluster

- Calculate the average distance from document i to all documents in other clusters, and take the smallest value [17].

$$d(i,C) = \frac{1}{[A]} \sum j \epsilon\, C\, d(i,j) \quad (9)$$

$$b(i) = min C \neq A\, d(i,C) \quad (10)$$

where
$d(i,C)$ = the average distance of document, i with all objects in another cluster C
C = Other clusters other than cluster A or C are not the same as cluster A.
$b(i)$ = The average distance of the object with all other objects that are different in the other clusters.

- Calculating the value of the Silhouette Coefficient

$$s(i) = \frac{b(i) - a(i)}{\min(a(i), b(i))} \quad (11)$$

Explain:
$s(i)$) = Silhouette Coefficient Value.
$a(i)$ = The average distance i to all objects in cluster
$b(i)$ = Average distance i to all other cluster objects

III. Result and Discussion

Data on the number of incoming students and graduate students in a Y faculty at an X university are listed in Table 1. This investigation used the Python programming language to carry out the grouping process.

A. Calculation Process

This study uses student data from faculty Y's R, S, T, and U departments, with a total of 90, 87, 76, and 30 students. The information is comprised of student records from the classes of 2014 and 2015, and it includes a number of characteristics, including college class, ID, name, department, entrance path, school, name of the school, district of origin, mathematics score, length of study, GPA, and TOEFL. The data load stage precedes the data processing stage. As illustrated in Table 2, the S department dataset, for instance, is loaded first before processing.

TABLE I
DATA ON THE NUMBER OF STUDENTS AND GRADUATES OF Y FACULTY

Year	Number of Students	Number of Graduates
2012	360	14 or 4%
2013	340	22 or 6%
2014	591	55 or 9%
2015	837	181 or 22%

TABLE II
DATA OF STUDENTS

College class		ID	Name	Department	Entrance	School Type	Name of School	County Origin	Mathematics Score	Length of Study	GPA	TOEFL
0	2014	1400019002	Muhammad Andrean Pratama	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMK N 2 Yogyakarta	Yogyakarta	42.55	4 years, 9 months	3.42	466
1	2014	1400019006	Dedi Mustaal	Industry Engineering	Achivement path –Raport	Vocational High School (SMK)	SMK Kharya Dharma 1 Kotabumi	Lampung	56	4 years, 9 months	2.91	470
2	2014	1400019012	Novrawan	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMA N 12 Merangin	Jambi	57.67	4 years, 1 months	3.09	413
3	2014	1400019014	Bangun Sajiwo Prihatmoko	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMK N 3 Yogyakarta	Yogyakarta	80.33	4 years, 1 months	3.34	410
4	2014	1400019017	Muhammad Khrisna Putra	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMA Perintis 2 Bandar Lampung	Lampung	86.33	5 years, 0 months	3.22	456
					Achivement path – Mathematics Score							
71	2015	1500019163	Wahdi Luthfi Ramadhan	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 5 Tebo	Jambi	88.33	3 years, 11 months	3.48	406
72	2015	1500019165	Intan Pratiwi	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 1 Bandongan	Central Java	84.67	3 years, 11 months	3.66	413
73	2015	1500019166	Rama Yudhi Fernando	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA Budi Utomo, Perak	East Java	84	4 years, 0 months	3.30	463
74	2015	1500019206	Sava Luna Wahyu Ellenora	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 1 Temblahan Hulu	Riau	85	3 years, 11 months	3.44	406
75	2015	1500019207	Dea Arivah Avelia	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 2 Cirebon	West Java	87	3 years, 11 months	3.57	436

The next step is to tidy up the data, although the ID and School Name characteristics have already been saved. The One Hot Encoding method is then applied to the School attribute during data transformation. The data is translated into three categories for the district origin attribute, with origin one, grey, comprising North Maluku and Central Kalimantan. According to Figure 1, origin two, which is brown, is made up of Java Island, Sumatra Island, Sulawesi Island, West Kalimantan, South Kalimantan, Bali, NTT, and NTB; origin three, which is green, is made up of Papua, East Kalimantan, and Maluku. Based on a set of criteria, this classification is based on the caliber of the education received.

In Table 3, the categories are listed. As a consequence of the transformation findings based on mapping the quality of education in Indonesia, the school origin attribute is then changed into three new attributes, namely region I, region II, and region III, and placed into the dataset along with the values. Enter the K-Medoids Algorithm step after receiving the processing results, where each data is measured in relation to other data using the Euclidean Distance method, as indicated in Table 4.

The K-Medoids technique is used to process the data from the Euclidean Distance computation, and the result is some clusters with members that are similar to the medoid or the central data. The single linkage approach is used in the equation to perform calculations for the AHC Process from the Euclidean Table (3). Table 5 displays the results of the single link calculation from the first to the ninth student with the deletion of the rows and columns of the matrix in groups of students five and student ten and the addition of rows and columns for the group (school student 5, student school 10). The next step is to choose the smallest distance from the group to calculate the distance between the fifth and tenth students and the remaining groups. To achieve the results of clusters or grouping utilizing the K-Medoids and AHC methods, the single linkage step is carried out until there is only one cluster or grouping, as shown in Table 7.

TABLE III
DATA OF TRANSFORMATION RESULT

	Region I	Region II	Region III	Math Score	Length of Study	GPA	TOEFL	State Madrasah (MA)	Senior High School (SMA)	Vocational High School (SMK)
0	0	1	0	42.55	1748	3.42	466	0	0	1
1	0	1	0	56	1749	2.91	470	0	0	1
2	0	1	0	87.67	1513	3.09	413	0	1	0
3	0	1	0	60.33	1513	3.34	410	0	0	1
4	0	1	0	86.33	1842	3.22	456	0	1	0
71	0	1	0	88.33	1455	3.48	406	0	1	0
72	0	1	0	84.67	1455	3.66	413	0	1	0
73	0	1	0	84	1478	3.30	463	0	1	0
74	0	1	0	85	1455	3.44	406	0	1	0
75	0	1	0	87	1455	3.57	436	0	1	0

TABLE IV
EUCLIDEAN DISTANCE RESULT

	Euclidean Distance									
Data	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7	Student 8	Student 9	Student 10
Student 1	0	332.47	692.37	333.47	344.1	362.05	327.97	348.23	342.29	344.38
Student 2	332.47	0	362.18	11.95	32.9	29.62	12.29	26.35	44.12	31.04
Student 3	692.37	362.18	0	362.02	348.3	362.2	365.27	362.12	349.71	348.05
Student 4	333.47	11.95	362.02	0	32.26	30.79	16.82	17.17	48.68	33.65
Student 5	344.1	32.9	348.3	32.26	0	37.6	25.3	34.32	14.46	3.6
Student 6	362.05	29.62	362.2	30.79	37.6	0	36.37	14.21	48.34	35.47
Student 7	327.97	12.29	365,27	16.82	25.3	36.37	0	24.48	35.04	23.67
Student 8	348.23	16.35	362.12	17.17	34.32	14.21	24.48	0	46.11	10.66
Student 9	342.29	44.12	349.71	48.68	14.46	48.34	35.04	46.11	0	17.8
Student 10	344.38	31.04	348.05	33.65	3.6	35.47	23.67	10.66	17.8	0

TABLE V
SINGLE LINKED RESULT

Data	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6	Student 7	Student 8	Student 9	Student 10
Student 1	0	332.47	692.37	333.47	344.1	362.05	327.97	348.23	342.29	344.38
Student 2	332.47	0	362.18	11.95	32.9	29.62	12.29	26.35	44.12	31.04
Student 3	692.37	362.18	0	362.02	348.3	362.2	365.27	362.12	349.71	348.05
Student 4	333.47	11.95	362.02	0	32.26	30.79	16.82	17.17	48.68	33.65
Student 5	344.1	32.9	348.3	32.26	0	37.6	25.3	34.32	14.46	3.6
Student 6	362.05	29.62	362.2	30.79	37.6	0	36.37	14.21	48.34	35.47
Student 7	327.97	12.29	365,27	16.82	25.3	36.37	0	24.48	35.04	23.67
Student 8	348.23	16.35	362.12	17.17	34.32	14.21	24.48	0	46.11	10.66
Student 9	342.29	44.12	349.71	48.68	14.46	48.34	35.04	46.11	0	17.8
Student 10	344.38	31.04	348.05	33.65	3.6	35.47	23.67	10.66	17.8	0

TABLE VI
SINGLE LINKED RESULT

Data	Student 5 & 10	Student 1	Student 2	Student 3	Student 4	Student 6	Student 7	Student 8	Student 9
Student 5 & 10	0	344.1	31.04	348.05	32.26	35.47	35.47	10.66	14.45
Student 1	344.1	0	332.47	692.37	333.47	362.05	327.97	348.23	342.29
Student 2	31.04	332.47	0	362.18	11.95	29.62	12.29	16.35	44.12
Student 3	348.05	692.37	362.18	0	362.02	362.2	365.27	362.12	349.71
Student 4	32.26	333.47	11.95	362.02	0	30.79	26.82	17.17	48.68
Student 6	35.47	362.05	29.62	362.2	30.79	0	36.37	14.21	48.34
Student 7	23.67	327.97	12.29	365.27	26.82	36.37	0	24.48	35.04
Student 8	10.66	348.23	16.35	362.12	17.17	14.21	24.48	0	46.11
Student 9	14.45	342.29	44.12	349.71	48.68	48.34	35.04	46.11	0

TABLE VII
GROUPING USING K-MEDOID AND AHC METHODS

College class		NIM	School Name	Region I	Region II	Region III	MTK	Length of Study	GPA	TOEFL	State Madrasah (MA)	Senior Hogh School (SMA)	Vocational High School (SMK)	Cluster
0	2014	1400019002	SMK N 2 Yogyakarta	0	1	0	42.55	1748	3.42	466	0	0	1	1
1	2014	1400019006	SMK Kharya Dharma 1 Kotabumi	0	1	0	56	1749	2.91	470	0	0	1	1
2	2014	1400019012	SMA N 12 Merangin	0	1	0	57.67	1513	3.09	413	0	1	0	2
3	2014	1400019014	SMK N 3 Yogyakarta	0	1	0	80.33	1513	3.34	410	0	0	1	2
4	2014	1400019017	SMA Perintis 2 Bandar Lampung	0	1	0	86.33	1842	3.22	456	0	1	0	1
71	2015	1500019163	SMA N 5 Tebo	0	1	0	88.33	1455	3.48	406	0	1	0	2
72	2015	1500019165	SMA N 1 Bandongan	0	1	0	84.67	1455	3.66	413	0	1	0	2
73	2015	1500019166	SMA Budi Utomo, Batak	0	1	0	84	1478	3.30	463	0	1	0	2
74	2015	1500019206	SMA N 1 Temblahan Hulu	0	1	0	85	1455	3.44	406	0	1	0	2
75	2015	1500019207	SMA N 2 Cirebon	0	1	0	87	1455	3.57	436	0	1	0	2

B. Clustering Accuracy Test

The silhouette coefficient [30]–[32] is used in a test to find data groupings that resemble each other as closely as feasible. Data from 4 departments are used to conduct the test. Three trials are conducted for each department, using cluster sizes of 2, 3, and 4. The silhouette coefficient approach is used to conduct this test. Figures 2 through Figure 5 display the test findings. On the graphs of the two tests, a value that is exactly proportional to the test results using the K-Medoids approach and the AHC method can be seen. Figure 3 contrasts the accuracy of clusters 3 and 4, nevertheless. The accuracy of the AHC approach is rising, but the accuracy of the K-Medoids method is falling. The data distribution in the S Department accounts for the accuracy discrepancy. The AHC approach is superior for grouping student data, as shown by the test in Figures 2 to 5. The best outcomes for the R department are displayed in Table 8.

Fig. 1 Data Comparison between K-Medoids and AHC Methods for Department R

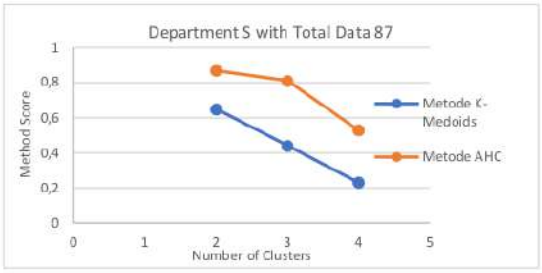

Fig. 2 Data Comparison between K Medoids and AHC Methods for Department S

Fig. 3 Data Comparison between K Medoids and AHC Methods for Department T

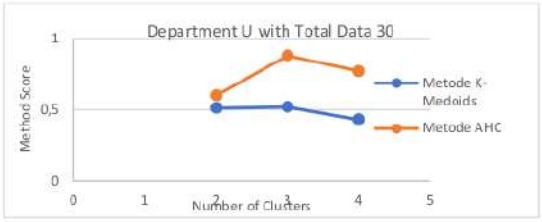

Fig. 4 Data Comparison between K Medoids and AHC Methods for Department U

TABLE VIII
THE BEST SILHOUETTE COEFFICIENT RESULT

Department	Amount of Data	Amount of Cluster	Silhouette Results AHC Method
R	90	4	0.88
S	87	2	0.87
T	76	4	0.88
U	30	3	0.88

C. *Grouping Results*

All of the data in Table 8's grouping, which employs the AHC findings, have accuracy values above 0.8, suggesting that the grouping's resulting structure is substantial [33]. In order to graduate on time and with a GPA over 3, students from region 2, namely Java and Kalimantan, who have a math score of 80 or more and the name of a high school, can be referred to the R department. The distribution of the grouped data in the R department is shown in Table 9, and the graph of the results is shown in Figure 5.

Regarding the S department, the math score has no impact on timely graduation and GPA rankings because students can still graduate on schedule and achieve GPAs greater than 3. The favored origins of the students are SENIOR HIGH SCHOOL and Java. Figure 6 depicts the graph of the outcomes of grouping the S department data, and Table 10 presents the distribution of the grouped data in the S department.

To graduate on time and with a GPA above three in the T department, a student must have a math score of at least 75. A nearby Javanese high school inspired the name of the school. The distribution of the grouped data for department T is shown in Table 11, and the dendrogram of those results is shown in Figure 7.

If a student has a math score of at least 80, is from the Java region, and graduated from high school with a GPA of at least three, they may be referred for the U department. Additionally, Sumatran students can graduate on time with a GPA in the top three using a math score of 75. The distribution of grouped data in the U department is shown in Table 12, and Figure 8 displays the graph of the data from the U department after being grouped.

TABLE IX
DISTRIBUTION OF GROUPING RESULT DATA IN THE R DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	4	Senior High School dan Vocational High School	2	Kalimantan	63	206	3.44	3 years, 10 months
Cluster 2	43	Senior High School	2	Java	80	424	3.5	3 years, 9 months
Cluster 3	40	Senior High School	2	Java	82	428	3.38	4 years, 3 months
Cluster 4	3	Senior High School	2	Java	79	426	2.96	5 years

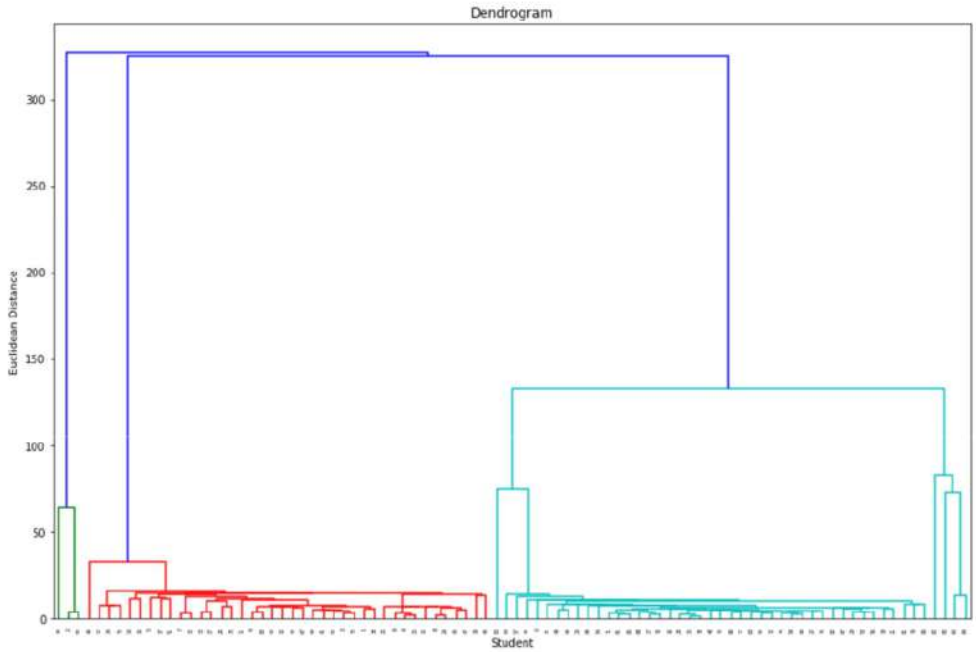

Fig. 5 Dendogram of Clustering of R Department Using AHC Method

TABLE X
DISTRIBUTION OF GROUPING RESULT DATA IN THE S DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	15	Senior High School	2	Java	60	420	3.33	4 years, 2 months
Cluster 2	72	Senior High School	2	Java	67	438	3.53	3 years, 9 months

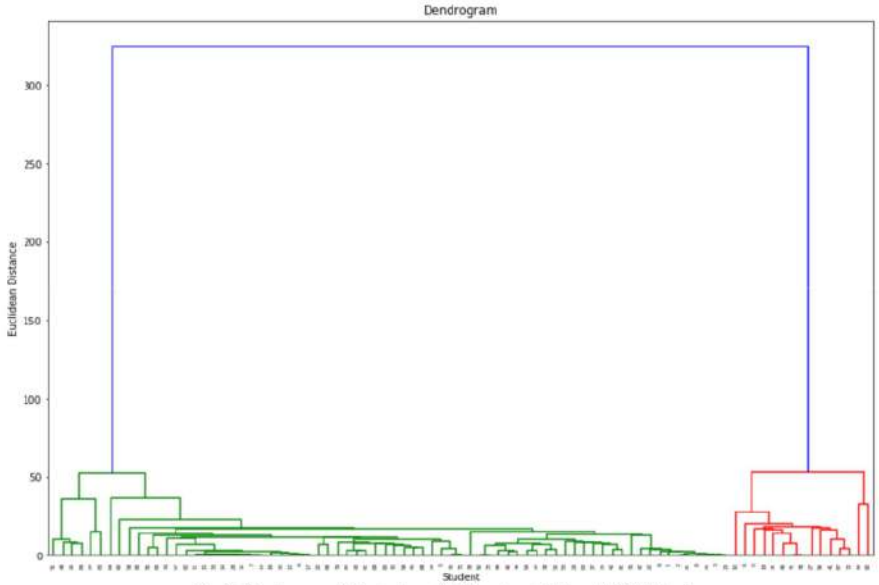

Fig. 6 Dendogram of Clustering of S Department Using AHC Method

TABLE XI
DISTRIBUTION OF GROUPING RESULT DATA IN THE T DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	49	Senior High School	2	Java	75	428	3.36	4 years, 2 months
Cluster 2	24	Senior High School	2	Java	84	423	3.56	3 years, 10 months
Cluster 3	2	Senior High School	2	Sumatera	84	455	3.22	5 years
Cluster 4	1	Senior High School	2	Sumatera	78	423	3.31	6 years

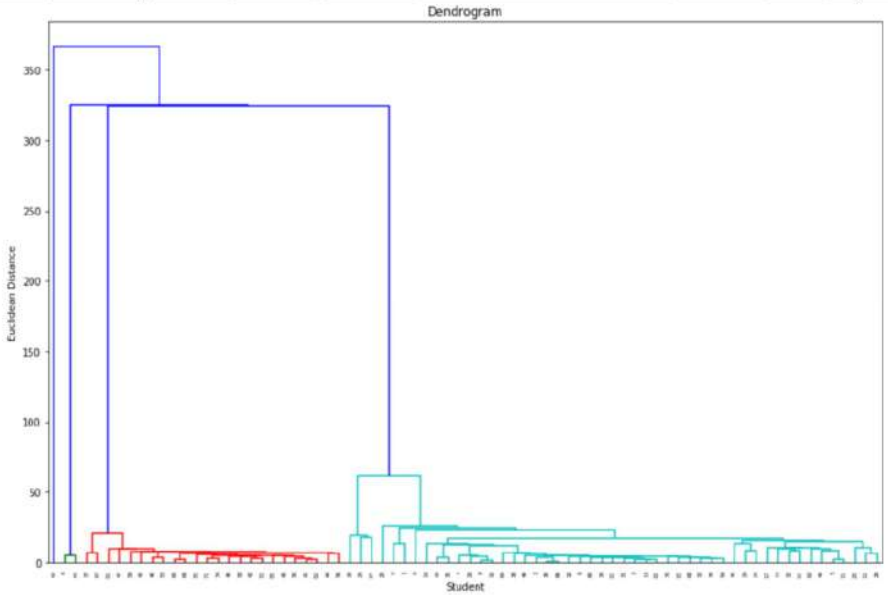

Fig. 7 Dendogram of Clustering of T Department Using AHC Method

TABLE XII

DISTRIBUTION OF GROUPING RESULT DATA IN THE U DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	10	Senior High School and Vocational High School	2	Sumatera	74	435	3.23	4 years, 4 months
Cluster 2	1	Senior High School	2	Java	80	443	3.31	5 years
Cluster 3	19	Senior High School and Vocational High School	2	Java	80	421	3.56	3 years, 10 months

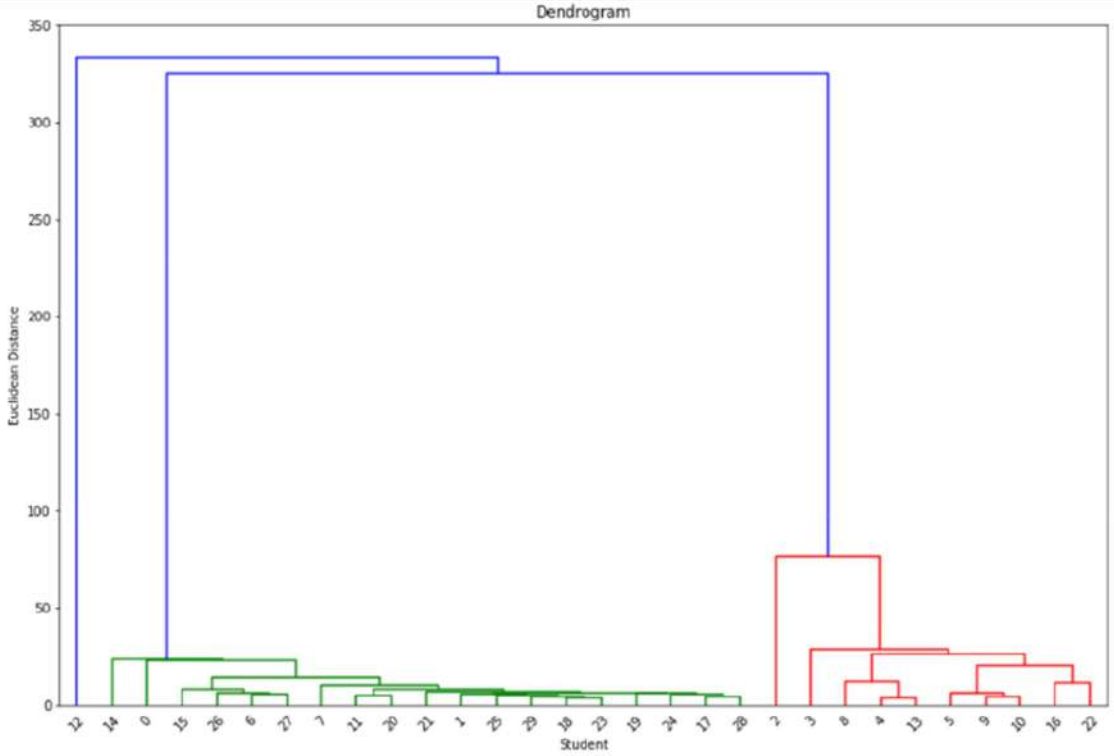

Fig. 8 Dendogram of Clustering of U Department Using AHC Method

IV. CONCLUSION

K-Medoids performed more accurately when categorizing student data using the AHC approach. However, the outcomes of the experiments conducted with both AHC and K-Medoids are directly proportionate. Only 1 of the 4 test data had not directly proportional results. According to the AHC test results, all research data has an accuracy value greater than 0.8. The grouping that results from this has a solid structure. Additionally, it can be advised that Faculty Y at University X pick new students for the R, T, and U departments with the name of their high school and origin from the Java region using a mathematics score of at least 80 based on the results of the overall grouping. A minimum math score of 76 is required for S majors to choose a new high school and Java students.

REFERENCES

[1] B. Hodge, B. Wright, and P. Bennett, "The Role of Grit in Determining Engagement and Academic Outcomes for University Students," *Res. High. Educ.*, vol. 59, no. 4, pp. 448–460, 2018.

[2] M. F. Parnes, C. Suárez-Orozco, O. Osei-Twumasi, and S. E. O. Schwartz, "Academic Outcomes Among Diverse Community College Students: What Is the Role of Instructor Relationships?," *Community Coll. Rev.*, vol. 48, no. 3, pp. 277–302, 2020, doi: 10.1177/0091552120909908.

[3] A. Alhadabi and A. C. Karpinski, "Grit, self-efficacy, achievement orientation goals, and academic performance in University students," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 519–535, 2020, doi: 10.1080/02673843.2019.1679202.

[4] M. J. Gormley, G. J. DuPaul, L. L. Weyandt, and A. D. Anastopoulos, "First-Year GPA and Academic Service Use Among College Students With and Without ADHD," *Physiol. Behav.*, pp. 1766–1779, 2019, doi: 10.1177/1087054715623046.First-Year.

[5] S. Chaturapruek, T. S. Dee, R. Johari, R. F. Kizilcec, and M. L. Stevens, "How a data-driven course planning tool affects college students' GPA: Evidence from two field experiments," *Proc. 5th Annu.*

ACM Conf. Learn. Scale, L S 2018, 2018, doi: 10.1145/3231644.3231668.

[6] D. Aggarwal and D. Sharma, "Application of clustering for student result analysis," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 50–53, 2019.

[7] A. Almasri, R. S. Alkhawaldeh, and E. Çelebi, "Clustering-Based EMT Model for Predicting Student Performance," *Arab. J. Sci. Eng.*, vol. 45, no. 12, pp. 10067–10078, 2020, doi: 10.1007/s13369-020-04578-4.

[8] D. S. Lamb, J. Downs, and S. Reader, "Space-time hierarchical clustering for identifying clusters in spatiotemporal point data," *ISPRS Int. J. Geo-Information*, vol. 9, no. 2, 2020, doi: 10.3390/ijgi9020085.

[9] L. Zappia and A. Oshlack, "Clustering trees: a visualization for evaluating clusterings at multiple resolutions," *Gigascience*, vol. 7, no. 7, pp. 1–9, 2018, doi: 10.1093/gigascience/giy083.

[10] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[11] Mardonov, "Structure and Mechanisms of Action of The Educational Cluster," *Int. J. Psychol. Rehabil.*, vol. 24, no. 07, pp. 1475–7192, 2020, [Online]. Available: https://hozir.org/pars_docs/refs/541/540182/540182.pdf.

[12] L. Zahrotun, N. hutami Putri, and A. N. Khusna, "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesisi Titles," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018.

[13] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, "A self-training approach for short text clustering," *ACL 2019 - 4th Work. Represent. Learn. NLP. RepL4NLP 2019 - Proc. Work.*, no. 2017, pp. 194–199, 2019, doi: 10.18653/v1/w19-4322.

[14] E. A. Anaam, S.-C. Haw, and P. Naveen, "Applied Fuzzy and Analytic Hierarchy Process Techniques in Hybrid Recommendation Approaches For E-CRM," *Int. J. Informatics Vis.*, vol. 6, no. 2, p. 2, 2022.

[15] H.-S. Park and C.-H. Jun, "Expert Systems with Applications An International Journal," *Expert Syst. Appl.*, vol. 145, no. 2, p. 3341, 2020.

[16] D. Sun, H. Fei, and Q. Li, "A Bisecting K-Medoids clustering Algorithm Based on Cloud Model," vol. 51, no. 11, pp. 308–315, 2018, doi: 10.1016/j.ifacol.2018.08.301.

[17] Martanto, S. Anwar, C. L. Rohmat, F. M. Basysyar, and Y. A. Wijaya, "Clustering of internet network usage using the K-Medoid method," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012036, 2021, doi: 10.1088/1757-899x/1088/1/012036.

[18] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means," *Am. J. Distance Educ.*, vol. 34, no. 2, pp. 137–156, 2020, doi: 10.1080/08923647.2020.1696140.

[19] S. Sinche *et al.*, "Analysis of Student Academic Performance Using Human-in-the-Loop Cyber-Physical Systems," *Telecom*, vol. 1, no. 1, pp. 18–31, 2020, doi: 10.3390/telecom1010003.

[20] O. Tinuke Omolewa, A. Taye Oladele, A. Adekanmi Adeyinka, and O. Roseline Oluwaseun, "Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions," *J. Eng. Appl. Sci.*, vol. 14, no. 22, pp. 8254–8260, 2019, doi: 10.36478/jeasci.2019.8254.8260.

[21] J. Oyelade *et al.*, "Data Clustering: Algorithms and Its Applications," *Proc. - 2019 19th Int. Conf. Comput. Sci. Its Appl. ICCSA 2019*, no. July, pp. 71–81, 2019, doi: 10.1109/ICCSA.2019.000-1.

[22] A. Naeem, M. Rehman, M. Anjum, and M. Asif, "Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm," *Curr. Sci.*, vol. 117, no. 6, pp. 1045–1053, 2019, doi: 10.18520/cs/v117/i6/1045-1053.

[23] S. Bipasha Biswas and M. Tariq Iqbal, "Solar Water Pumping System Control Using a Low Cost ESP32 Microcontroller," *Can. Conf. Electr. Comput. Eng.*, vol. 2018-May, pp. 1–5, 2018, doi: 10.1109/CCECE.2018.8447749.

[24] M. T. Lwin and M. M. Aye, "A Modified Hierarchical Agglomerative Approach for Efficient Document Clustering System," *Am. Sci. Res. J. Eng.*, vol. 29, no. 1, pp. 228–238, 2017, [Online]. Available: http://asrjetsjournal.org/.

[25] W. Xiaochun and W. Xia Li, " A Fast K -medoids Clustering Algorithm for Image Segmentation based Object Recognition ," *J. Robot. Auiom.*, vol. 4, no. 1, pp. 202–211, 2020, doi: 10.36959/673/371.

[26] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco, "An information-theoretic approach to hierarchical clustering of uncertain data," *Inf. Sci. (Ny).*, vol. 402, pp. 199–215, 2017, doi: 10.1016/j.ins.2017.03.030.

[27] A. Triayudi and I. Fitri, "Comparison of parameter-free agglomerative hierarchical clustering methods," *ICIC Express Lett.*, vol. 12, no. 10, pp. 973–980, 2018, doi: 10.24507/icicel.12.10.973.

[28] A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, and M. I. Awang, "Silhouette index for determining optimal k-means clustering on images in different color models," *Int. J. Eng. Technol.*, vol. 7, no. April, pp. 105–109, 2018, doi: 10.14419/ijet.v7i2.14.11464.

[29] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021, doi: 10.3390/e23060759.

[30] R. J. Roiger, *Data Mining A Tutorial-Based Primer*. Boca Raton, London, New York, 2017.

[31] N. Nidheesh, K. A. A. Nazeer, and P. M. Ameer, "A Hierarchical Clustering algorithm based on Silhouette Index for cancer subtype discovery from genomic data," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11459–11476, 2020, doi: 10.1007/s00521-019-04636-5.

[32] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, 2019, doi: 10.1088/1757-899X/569/5/052024.

[33] J. Han, J. Pei, and H. Tong, *Data Mining Concepts and Techniques*. Cambridge, MA 02139, United States: Elsevier Inc., 2023.

HASIL CEK_Comparison of K-Medoids Method and Analytical Hierarchy

ORIGINALITY REPORT

13%	13%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.joiv.org Internet Source	7%
2	isimmed.uny.ac.id Internet Source	1%
3	downloads.hindawi.com Internet Source	1%
4	trilogi.ac.id Internet Source	1%
5	Submitted to Telkom University Student Paper	1%
6	Submitted to University of Queensland Student Paper	1%
7	Ridho Ananda. "Silhouette Density Canopy K-Means for Mapping the Quality of Education Based on the Results of the 2019 National Exam in Banyumas Regency", Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika, 2019 Publication	1%

8	journals.ums.ac.id Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Comparison of K-Medoids Method and Analytical Hierarchy Clustering on Students' Data Grouping

Lisna Zahrotun [a,*], Utaminingsih Linarti [b], Banu Harli Trimulya Suandi As [a], Herri Kurnia [a], Liya Yusrina Sabila [c]

[a] *Informatics Department, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[b] *Industrial Engineering Department, Faculty of Industrial Technology, Universitas, Ahmad Dahlan, Yogyakarta, Indonesia*
[c] *Electrical Engineering Department, Faculty of Industrial Technology, Universitas, Ahmad Dahlan, Yogyakarta, Indonesia*

*Corresponding author: *lisna.zahrotun@tif.uad.ac.id*

***Abstract*— One sign of how successfully the educational process is carried out on campus in a university is the timely graduation of students. This study compares the Analytic Hierarchy Clustering (AHC) approach with the K-Medoids method, a data mining technique for categorizing student data based on school origin, region of origin, average math score, TOEFL, GPA, and length study. This study was carried out at University X, which contains a variety of architectural styles. The R department, the S department, the T department, and the U department make up one of them. K-Medoids and AHC techniques Utilize the number of clusters 2, 3, and 4 and the silhouette coefficient approach. The evaluation's findings indicate a value. Although there is a linear silhouette between the AHC and K-Medoids methods, the AHC approach (departments R: 0.88, S: 0.87, T: 0.88, and U: 0.88) has a more excellent Silhouette value than K-Medoids (department R: 0.35, department S: 0.65 number of cluster 2, department T: 0.67 number of cluster 2 and program Study U: 0,52). The results of the second approach, which includes the K-Medoids and AHC procedures, are determined by the data distribution to be clustered rather than by the quantity of data or clusters. Based on this methodology, University X can refer to the grouping outcomes for the four departments with two achievements to receive results on schedule.**

***Keywords*— Grouping; K-medoids; silhouette coefficient; analytical hierarchy clustering.**

Manuscript received 12 Sep. 2022; revised 8 Dec. 2022; accepted 20 Feb. 2023. Date of publication 30 Jun. 2023.

I. INTRODUCTION

The main metric used in universities to assess academic achievement is GPA [1]–[3]. The average grade received in all courses taken from the first semester to the last is known as the GPA. So, each semester's GPA will be determined. The GPA will be determined between the first and last semesters, though [4], [5]. Each student's GPA impacts how successful or successful they become at a university. Graduation rates or length of study are further indicators of a university's success besides GPA. University X has various faculties, and faculty Y is one of them. The R department, S department, T department, and U department are the four departments that make up Faculty Y. There are many students at Faculty Y. All departments in Faculty Y see a growth in the number of new students each year. This impacts a university's performance or success since there is an imbalance between new students and students who are graduating, which leads to poor evaluations. The ratio of instructors to students is still relevant because some factors, including the imbalance in the number of lecturers and students, can contribute to this imbalance. Another concern is the sheer volume of pupils, which will restrict how much time can be spent in places like labs. Analyzing student achievement is essential to figuring out how effective the current educational system [6]. Student graduation data, such as period of study, TOEFL, and GPA, with new student data, such as report score, school origin, and district origin, have never been combined together before now. The Faculty of Y will consider this, especially when choosing new students, by grouping the data from each department.

Clustering is a technique in data mining. Clustering is a method for organizing data that can be utilized [7]–[9]**.** Based on the values of their attributes, clustering is a technique for identifying homogeneous object groups. Clustering can assist in analyzing the variables influencing student learning results [10], [11]. Data in the form of text or numeric data can be grouped using clustering. Clustering has been employed in texts before [12], [13]. The Analytical Hierarchy Clustering

(AHC) approach and the K-Medoids method are two examples of the numerous techniques that can be applied. The accuracy of the AHP approach in this study's relationship management system research on electronic clients is 66.6%, and it performs better in terms of time complexity [14]. There has been an investigation into the center point using the K-Medoids technique [15]–[17]. In order to group data, this study contrasts the K-Medoids method with the Analytical Hierarchy Clustering (AHC) method. The data used in this study are the school's name, region of origin, math score, and GPA. The Silhouette Coefficient approach was used to conduct the test.

II. Material and Method

This study aims to examine the classification of student data using the K-Medoids method and the analytical Hierarchy Clustering (AHC) approach. The use of clustering approaches to analyze student academic performance has been studied in a number of publications over the past few years [6]. K-Means clustering is used to analyze student learning outcomes and performance [18]–[20] . K-Medoids and Analytical Hierarchy Clustering (AHC) are the methods applied in this study. Partition grouping is done using the K-Medoids approach, which is popular due to its effectiveness, simplicity, and convenience of usage [21]–[24]. The AHC approach, in contrast, uses hierarchical clustering to generate grouping of the individual data points inside a cluster in the shape of a tree [6]. These two grouping techniques, however, are appropriate for data having categorical data types [21]. As a result, the K-Medoids approach and the Analytical Hierarchy Clustering (AHC) method were used in this study.

A. K-Medoids

Data will be taken randomly to be used as central data in the cluster, each data has the opportunity to become central data, but most middle data is used as central data in a cluster based on the conditions of the K-Medoids Algorithm [25]. The steps of the K-Medoids Algorithm are as follows:

- Initialization of cluster centers as much as k (number of clusters).
- Group each data into the closest cluster using the Euclidean Distance approach to calculate the distance between data with the equation (1):

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

Explanation:
$x(i)$ = the first i data.
$y(i)$ = the second i data.
n = amount of data

- Then select the data randomly in each cluster used as a candidate for a new medoid.
- After that, calculate the distance of each data in each cluster with the new medoid candidate.
- Then calculate the total deviation (S) by calculating the new total distance value - the old total distance. If S < 0, replace objects with cluster data to form a new set of k objects as medoids.
- Repeat steps 3 to 5 until there is no medoid change so that clusters and their respective cluster members are obtained.

B. Analytical Hierarchy Clustering (AHC)

This grouping is a Hierarchical Grouping which allows having two main approaches, namely the hierarchical approach and the split approach [21], [26], [27]. AHC Algorithm Steps:

- The distance between data is calculated using the Euclidean formula at this stage. Euclidean Distance Formula (6):

$$\|U - V\| = \sqrt{\sum i}\,(Ui - Vi)^2 \quad (6)$$

where:
Ui = U value on training data
Vi = value of V on test data

- Based on the distance matrix, then the data is grouped using Agglomerative Hierarchical Clustering (AHC) using the single linkage method in equation (7).

$$d_{data} = min\{d_{data}\}, d_{data} \in \mathrm{D} \quad (7)$$

where:
d_{data} = the distance between the nearest/smallest neighbor of the data group
D= Euclidean distance proximity matrix

C. Silhouette Coefficient

This method will calculate the level of proximity between data or objects in a cluster. For example, the steps in the silhouette coefficient process [16], [28], [29] are as follows :

- Calculate the average distance from a document, i with all other documents in one cluster [17].

$$a(i) = \frac{1}{[A]-1} \sum j \epsilon \mathrm{C}\; d(i,j) \quad (8)$$

where
$a(i)$ = The mean difference of object (i) to all other objects on A
$d(i,j)$ = Distance between data i to j
A = cluster

- Calculate the average distance from document i to all documents in other clusters, and take the smallest value [17].

$$d(i,C) = \frac{1}{[A]} \sum j \epsilon\; C\; d(i,j) \quad (9)$$

$$b(i) = min C \neq \mathrm{A}\; d(i,C) \quad (10)$$

where
$d(i,C)$ = the average distance of document, i with all objects in another cluster C
C = Other clusters other than cluster A or C are not the same as cluster A.
$b(i)$ = The average distance of the object with all other objects that are different in the other clusters.

- Calculating the value of the Silhouette Coefficient

$$s(i) = \frac{\mathrm{b(i)-a(i)}}{\min(\mathrm{a(i),b(i)})} \quad (11)$$

Explain:
$s(i)$) = Silhouette Coefficient Value.
$a(i)$ = The average distance i to all objects in cluster
$b(i)$ = Average distance i to all other cluster objects

III. Result and Discussion

Data on the number of incoming students and graduate students in a Y faculty at an X university are listed in Table 1. This investigation used the Python programming language to carry out the grouping process.

A. Calculation Process

This study uses student data from faculty Y's R, S, T, and U departments, with a total of 90, 87, 76, and 30 students. The information is comprised of student records from the classes of 2014 and 2015, and it includes a number of characteristics, including college class, ID, name, department, entrance path, school, name of the school, district of origin, mathematics score, length of study, GPA, and TOEFL. The data load stage precedes the data processing stage. As illustrated in Table 2, the S department dataset, for instance, is loaded first before processing.

TABLE I
DATA ON THE NUMBER OF STUDENTS AND GRADUATES OF Y FACULTY

Year	Number of Students	Number of Graduates
2012	360	14 or 4%
2013	340	22 or 6%
2014	591	55 or 9%
2015	837	181 or 22%

TABLE II
DATA OF STUDENTS

College class		ID	Name	Department	Entrance	School Type	Name of School	County Origin	Mathematics Score	Length of Study	GPA	TOEFL
0	2014	1400019002	Muhammad Andrean Pratama	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMK N 2 Yogyakarta	Yogyakarta	42.55	4 years, 9 months	3.42	466
1	2014	1400019006	Dedi Mustaal	Industry Engineering	Achivement path –Raport	Vocational High School (SMK)	SMK Kharya Dharma 1 Kotabumi	Lampung	56	4 years, 9 months	2.91	470
2	2014	1400019012	Novrawan	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMA N 12 Merangin	Jambi	57.67	4 years, 1 months	3.09	413
3	2014	1400019014	Bangun Sajiwo Prihatmoko	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMK N 3 Yogyakarta	Yogyakarta	80.33	4 years, 1 months	3.34	410
4	2014	1400019017	Muhammad Khrisna Putra	T	Achivement path – Mathematics Score	Vocational High School (SMK)	SMA Perintis 2 Bandar Lampung	Lampung	86.33	5 years, 0 months	3.22	456
					Achivement path – Mathematics Score							
71	2015	1500019163	Wahdi Luthfi Ramadhan	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 5 Tebo	Jambi	88.33	3 years, 11 months	3.48	406
72	2015	1500019165	Intan Pratiwi	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 1 Bandongan	Central Java	84.67	3 years, 11 months	3.66	413
73	2015	1500019166	Rama Yudhi Fernando	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA Budi Utomo, Perak	East Java	84	4 years, 0 months	3.30	463
74	2015	1500019206	Sava Luna Wahyu Ellenora	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 1 Temblahan Hulu	Riau	85	3 years, 11 months	3.44	406
75	2015	1500019207	Dea Arivah Avelia	T	Achivement path – Mathematics Score	Senior High School (SMA)	SMA N 2 Cirebon	West Java	87	3 years, 11 months	3.57	436

The next step is to tidy up the data, although the ID and School Name characteristics have already been saved. The One Hot Encoding method is then applied to the School attribute during data transformation. The data is translated into three categories for the district origin attribute, with origin one, grey, comprising North Maluku and Central Kalimantan. According to Figure 1, origin two, which is brown, is made up of Java Island, Sumatra Island, Sulawesi Island, West Kalimantan, South Kalimantan, Bali, NTT, and NTB; origin three, which is green, is made up of Papua, East Kalimantan, and Maluku. Based on a set of criteria, this classification is based on the caliber of the education received.

In Table 3, the categories are listed. As a consequence of the transformation findings based on mapping the quality of education in Indonesia, the school origin attribute is then changed into three new attributes, namely region I, region II, and region III, and placed into the dataset along with the values. Enter the K-Medoids Algorithm step after receiving the processing results, where each data is measured in relation to other data using the Euclidean Distance method, as indicated in Table 4.

The K-Medoids technique is used to process the data from the Euclidean Distance computation, and the result is some clusters with members that are similar to the medoid or the central data. The single linkage approach is used in the equation to perform calculations for the AHC Process from the Euclidean Table (3). Table 5 displays the results of the single link calculation from the first to the ninth student with the deletion of the rows and columns of the matrix in groups of students five and student ten and the addition of rows and columns for the group (school student 5, student school 10). The next step is to choose the smallest distance from the group to calculate the distance between the fifth and tenth students and the remaining groups. To achieve the results of clusters or grouping utilizing the K-Medoids and AHC methods, the single linkage step is carried out until there is only one cluster or grouping, as shown in Table 7.

TABLE III
DATA OF TRANSFORMATION RESULT

	Region I	Region II	Region III	Math Score	Length of Study	GPA	TOEFL	State Madrasah (MA)	Senior High School (SMA)	Vocational High School (SMK)
0	0	1	0	42.55	1748	3.42	466	0	0	1
1	0	1	0	56	1749	2.91	470	0	0	1
2	0	1	0	87.67	1513	3.09	413	0	1	0
3	0	1	0	60.33	1513	3.34	410	0	0	1
4	0	1	0	86.33	1842	3.22	456	0	1	0
71	0	1	0	88.33	1455	3.48	406	0	1	0
72	0	1	0	84.67	1455	3.66	413	0	1	0
73	0	1	0	84	1478	3.30	463	0	1	0
74	0	1	0	85	1455	3.44	406	0	1	0
75	0	1	0	87	1455	3.57	436	0	1	0

TABLE IV
EUCLIDEAN DISTANCE RESULT

Euclidean Distance										
Data	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 5**	**Student 6**	**Student 7**	**Student 8**	**Student 9**	**Student 10**
Student 1	0	332.47	692.37	333.47	344.1	362.05	327.97	348.23	342.29	344.38
Student 2	332.47	0	362.18	11.95	32.9	29.62	12.29	26.35	44.12	31.04
Student 3	692.37	362.18	0	362.02	348.3	362.2	365.27	362.12	349.71	348.05
Student 4	333.47	11.95	362.02	0	32.26	30.79	16.82	17.17	48.68	33.65
Student 5	344.1	32.9	348.3	32.26	0	37.6	25.3	34.32	14.46	3.6
Student 6	362.05	29.62	362.2	30.79	37.6	0	36.37	14.21	48.34	35.47
Student 7	327.97	12.29	365,27	16.82	25.3	36.37	0	24.48	35.04	23.67
Student 8	348.23	16.35	362.12	17.17	34.32	14.21	24.48	0	46.11	10.66
Student 9	342.29	44.12	349.71	48.68	14.46	48.34	35.04	46.11	0	17.8
Student 10	344.38	31.04	348.05	33.65	3.6	35.47	23.67	10.66	17.8	0

TABLE V
SINGLE LINKED RESULT

Data	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 5**	**Student 6**	**Student 7**	**Student 8**	**Student 9**	**Student 10**
Student 1	0	332.47	692.37	333.47	344.1	362.05	327.97	348.23	342.29	344.38
Student 2	332.47	0	362.18	11.95	32.9	29.62	12.29	26.35	44.12	31.04
Student 3	692.37	362.18	0	362.02	348.3	362.2	365.27	362.12	349.71	348.05
Student 4	333.47	11.95	362.02	0	32.26	30.79	16.82	17.17	48.68	33.65
Student 5	344.1	32.9	348.3	32.26	0	37.6	25.3	34.32	14.46	3.6
Student 6	362.05	29.62	362.2	30.79	37.6	0	36.37	14.21	48.34	35.47
Student 7	327.97	12.29	365,27	16.82	25.3	36.37	0	24.48	35.04	23.67
Student 8	348.23	16.35	362.12	17.17	34.32	14.21	24.48	0	46.11	10.66
Student 9	342.29	44.12	349.71	48.68	14.46	48.34	35.04	46.11	0	17.8
Student 10	344.38	31.04	348.05	33.65	3.6	35.47	23.67	10.66	17.8	0

TABLE VI
SINGLE LINKED RESULT

Data	**Student 5 & 10**	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 6**	**Student 7**	**Student 8**	**Student 9**
Student 5 & 10	0	344.1	31.04	348.05	32.26	35.47	35.47	10.66	14.45
Student 1	344.1	0	332.47	692.37	333.47	362.05	327.97	348.23	342.29
Student 2	31.04	332.47	0	362.18	11.95	29.62	12.29	16.35	44.12
Student 3	348.05	692.37	362.18	0	362.02	362.2	365.27	362.12	349.71
Student 4	32.26	333.47	11.95	362.02	0	30.79	26.82	17.17	48.68
Student 6	35.47	362.05	29.62	362.2	30.79	0	36.37	14.21	48.34
Student 7	23.67	327.97	12.29	365.27	26.82	36.37	0	24.48	35.04
Student 8	10.66	348.23	16.35	362.12	17.17	14.21	24.48	0	46.11
Student 9	14.45	342.29	44.12	349.71	48.68	48.34	35.04	46.11	0

TABLE VII
GROUPING USING K-MEDOID AND AHC METHODS

College class		NIM	School Name	Region I	Region II	Region III	MTK	Length of Study	GPA	TOEFL	State Madrasah (MA)	Senior Hogh School (SMA)	Vocational High School (SMK)	Cluster
0	2014	1400019002	SMK N 2 Yogyakarta	0	1	0	42.55	1748	3.42	466	0	0	1	1
1	2014	1400019006	SMK Kharya Dharma 1 Kotabumi	0	1	0	56	1749	2.91	470	0	0	1	1
2	2014	1400019012	SMA N 12 Merangin	0	1	0	57.67	1513	3.09	413	0	1	0	2
3	2014	1400019014	SMK N 3 Yogyakarta	0	1	0	80.33	1513	3.34	410	0	0	1	2
4	2014	1400019017	SMA Perintis 2 Bandar Lampung	0	1	0	86.33	1842	3.22	456	0	1	0	1
71	2015	1500019163	SMA N 5 Tebo	0	1	0	88.33	1455	3.48	406	0	1	0	2
72	2015	1500019165	SMA N 1 Bandongan	0	1	0	84.67	1455	3.66	413	0	1	0	2
73	2015	1500019166	SMA Budi Utomo, Perak	0	1	0	84	1478	3.30	463	0	1	0	2
74	2015	1500019206	SMA N 1 Temblahan Hulu	0	1	0	85	1455	3.44	406	0	1	0	2
75	2015	1500019207	SMA N 2 Cirebon	0	1	0	87	1455	3.57	436	0	1	0	2

B. *Clustering Accuracy Test*

The silhouette coefficient [30]–[32] is used in a test to find data groupings that resemble each other as closely as feasible. Data from 4 departments are used to conduct the test. Three trials are conducted for each department, using cluster sizes of 2, 3, and 4. The silhouette coefficient approach is used to conduct this test. Figures 2 through Figure 5 display the test findings. On the graphs of the two tests, a value that is exactly proportional to the test results using the K-Medoids approach and the AHC method can be seen. Figure 3 contrasts the accuracy of clusters 3 and 4, nevertheless. The accuracy of the AHC approach is rising, but the accuracy of the K-Medoids method is falling. The data distribution in the S Department accounts for the accuracy discrepancy. The AHC approach is superior for grouping student data, as shown by the test in Figures 2 to 5. The best outcomes for the R department are displayed in Table 8.

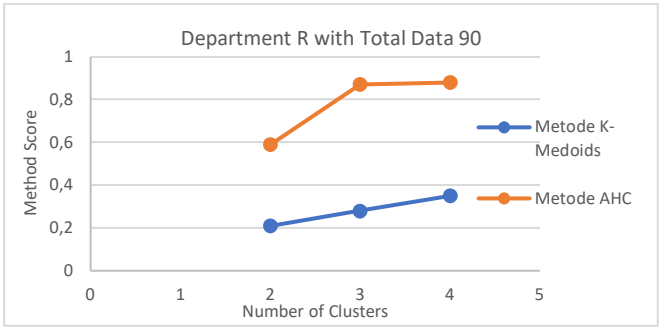

Fig. 1 Data Comparison between K-Medoids and AHC Methods for Department R

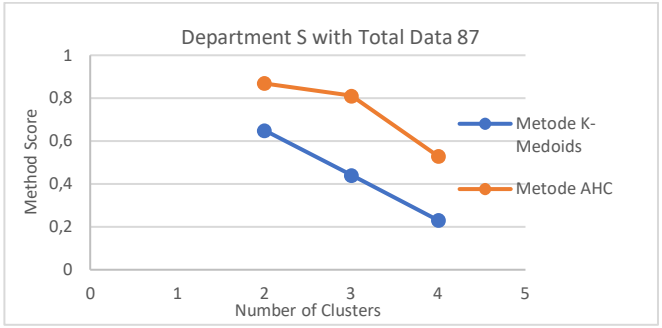

Fig. 2 Data Comparison between K Medoids and AHC Methods for Department S

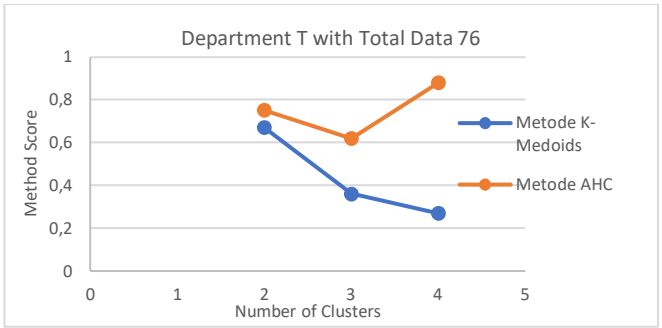

Fig. 3 Data Comparison between K Medoids and AHC Methods for Department T

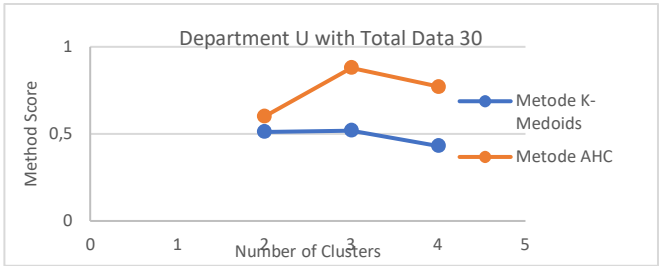

Fig. 4 Data Comparison between K Medoids and AHC Methods for Department U

TABLE VIII
THE BEST SILHOUETTE COEFFICIENT RESULT

Department	Amount of Data	Amount of Cluster	Silhouette Results AHC Method
R	90	4	0.88
S	87	2	0.87
T	76	4	0.88
U	30	3	0.88

C. Grouping Results

All of the data in Table 8's grouping, which employs the AHC findings, have accuracy values above 0.8, suggesting that the grouping's resulting structure is substantial [33]. In order to graduate on time and with a GPA over 3, students from region 2, namely Java and Kalimantan, who have a math score of 80 or more and the name of a high school, can be referred to the R department. The distribution of the grouped data in the R department is shown in Table 9, and the graph of the results is shown in Figure 5.

Regarding the S department, the math score has no impact on timely graduation and GPA rankings because students can still graduate on schedule and achieve GPAs greater than 3. The favored origins of the students are SENIOR HIGH SCHOOL and Java. Figure 6 depicts the graph of the outcomes of grouping the S department data, and Table 10 presents the distribution of the grouped data in the S department.

To graduate on time and with a GPA above three in the T department, a student must have a math score of at least 75. A nearby Javanese high school inspired the name of the school. The distribution of the grouped data for department T is shown in Table 11, and the dendrogram of those results is shown in Figure 7.

If a student has a math score of at least 80, is from the Java region, and graduated from high school with a GPA of at least three, they may be referred for the U department. Additionally, Sumatran students can graduate on time with a GPA in the top three using a math score of 75. The distribution of grouped data in the U department is shown in Table 12, and Figure 8 displays the graph of the data from the U department after being grouped.

TABLE IX
DISTRIBUTION OF GROUPING RESULT DATA IN THE R DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	4	Senior High School dan Vocational High School	2	Kalimantan	63	206	3.44	3 years, 10 months
Cluster 2	43	Senior High School	2	Java	80	424	3.5	3 years, 9 months
Cluster 3	40	Senior High School	2	Java	82	428	3.38	4 years, 3 months
Cluster 4	3	Senior High School	2	Java	79	426	2.96	5 years

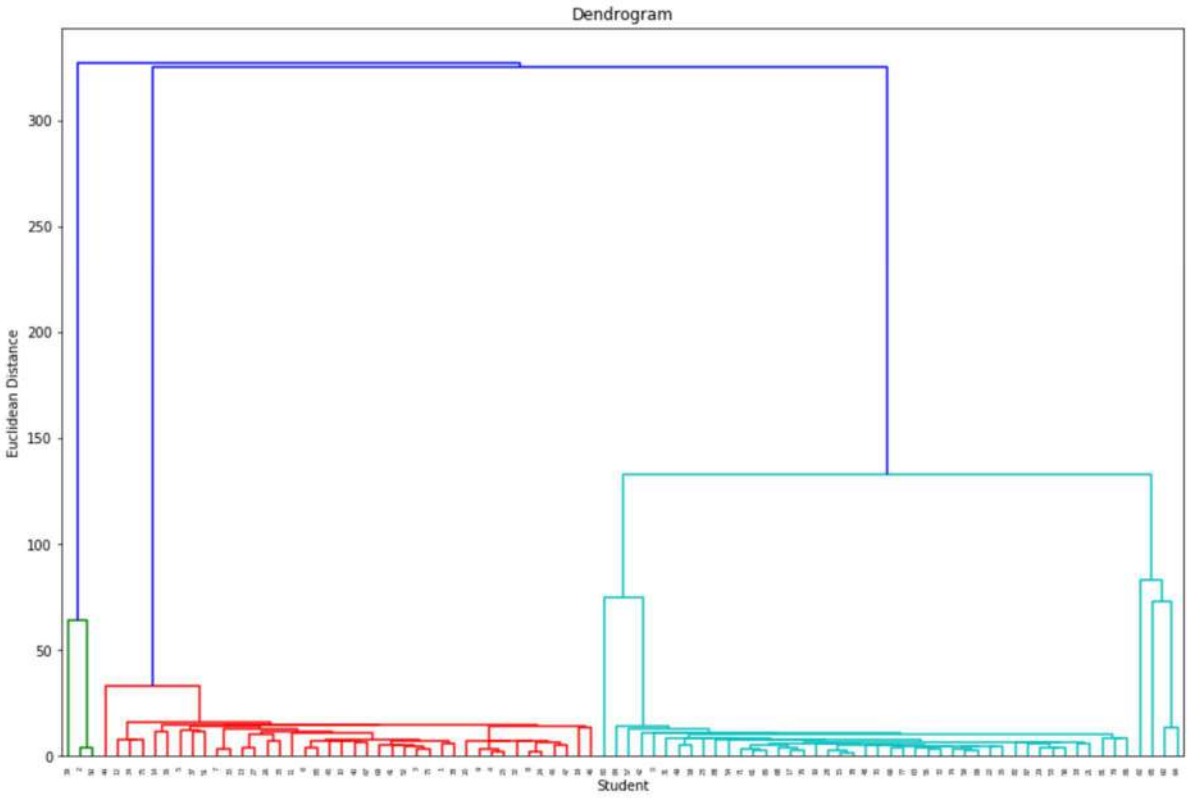

Fig. 5 Dendogram of Clustering of R Department Using AHC Method

TABLE X
DISTRIBUTION OF GROUPING RESULT DATA IN THE S DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	15	Senior High School	2	Java	60	420	3.33	4 years, 2 months
Cluster 2	72	Senior High School	2	Java	67	438	3.53	3 years, 9 months

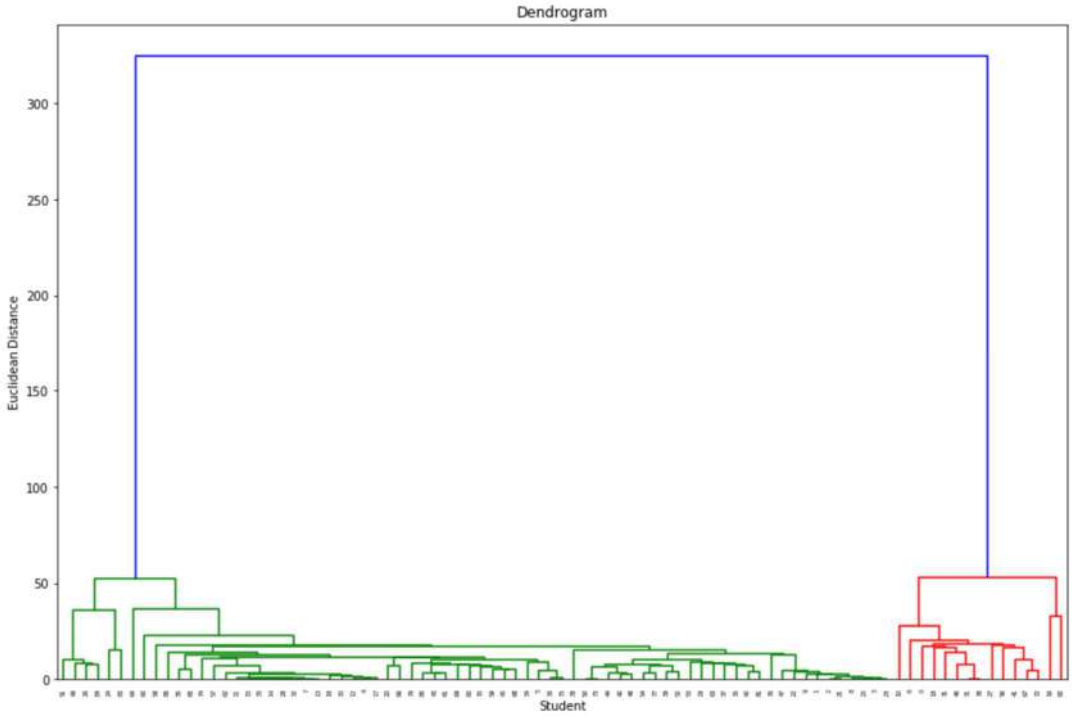

Fig. 6 Dendogram of Clustering of S Department Using AHC Method

TABLE XI
DISTRIBUTION OF GROUPING RESULT DATA IN THE T DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	49	Senior High School	2	Java	75	428	3.36	4 years, 2 months
Cluster 2	24	Senior High School	2	Java	84	423	3.56	3 years, 10 months
Cluster 3	2	Senior High School	2	Sumatera	84	455	3.22	5 years
Cluster 4	1	Senior High School	2	Sumatera	78	423	3.31	6 years

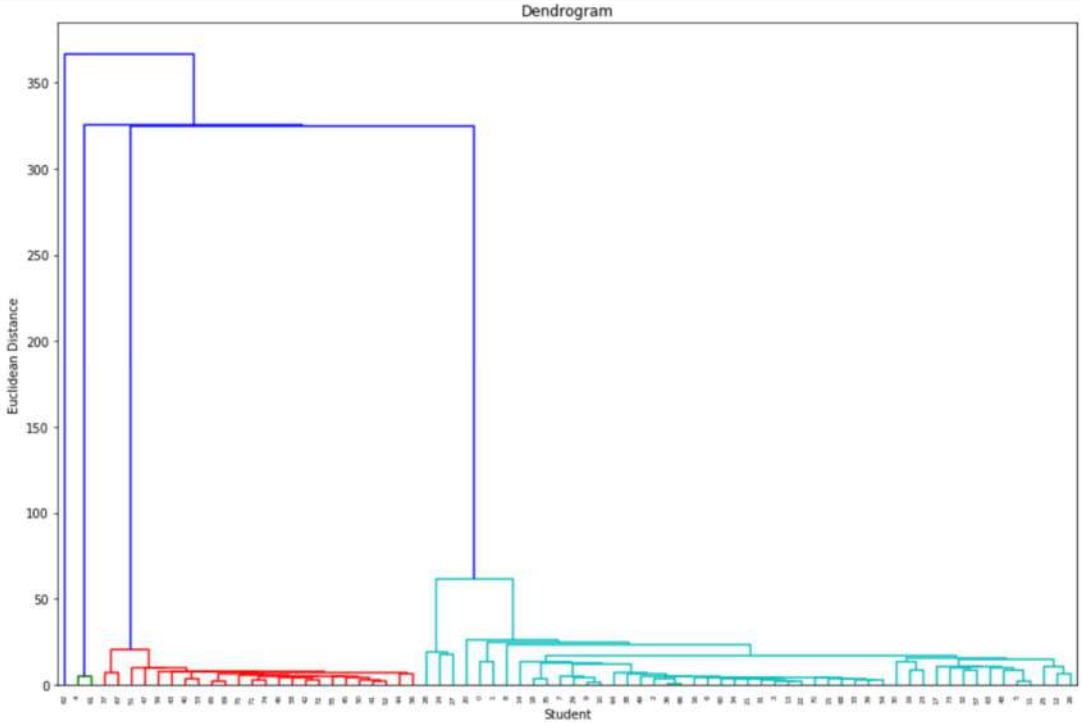

Fig. 7 Dendogram of Clustering of T Department Using AHC Method

TABLE XII

DISTRIBUTION OF GROUPING RESULT DATA IN THE U DEPARTMENT USING AHC METHOD

K Value	Data	School Type	Region	Island	Average of Mathematics	TOEFL	GPA	Length of Study
Cluster 1	10	Senior High School and Vocational High School	2	Sumatera	74	435	3.23	4 years, 4 months
Cluster 2	1	Senior High School	2	Java	80	443	3.31	5 years
Cluster 3	19	Senior High School and Vocational High School	2	Java	80	421	3.56	3 years, 10 months

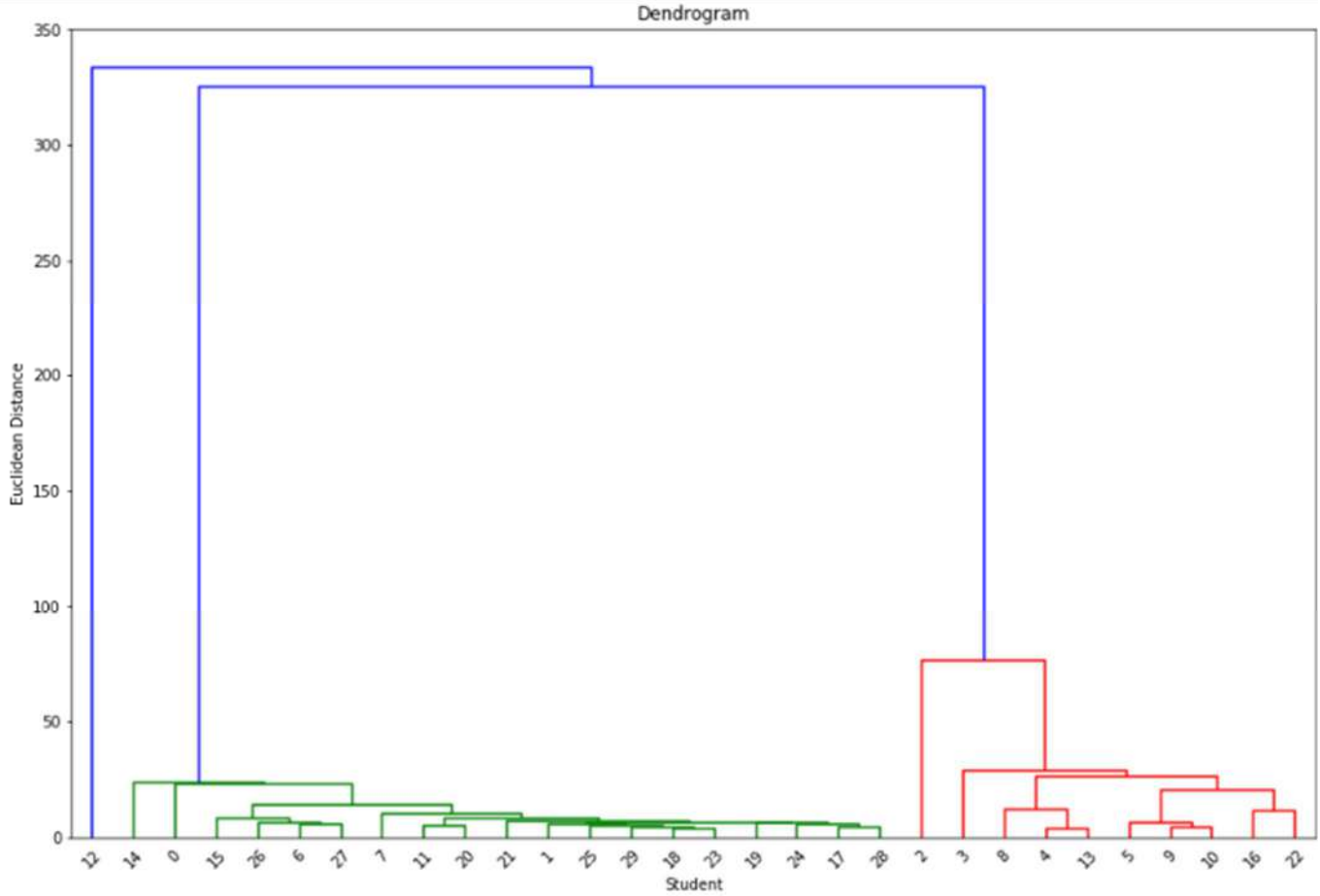

Fig. 8 Dendogram of Clustering of U Department Using AHC Method

IV. CONCLUSION

K-Medoids performed more accurately when categorizing student data using the AHC approach. However, the outcomes of the experiments conducted with both AHC and K-Medoids are directly proportionate. Only 1 of the 4 test data had not directly proportional results. According to the AHC test results, all research data has an accuracy value greater than 0.8. The grouping that results from this has a solid structure. Additionally, it can be advised that Faculty Y at University X pick new students for the R, T, and U departments with the name of their high school and origin from the Java region using a mathematics score of at least 80 based on the results of the overall grouping. A minimum math score of 76 is required for S majors to choose a new high school and Java students.

REFERENCES

[1] B. Hodge, B. Wright, and P. Bennett, "The Role of Grit in Determining Engagement and Academic Outcomes for University Students," *Res. High. Educ.*, vol. 59, no. 4, pp. 448–460, 2018.

[2] M. F. Parnes, C. Suárez-Orozco, O. Osei-Twumasi, and S. E. O. Schwartz, "Academic Outcomes Among Diverse Community College Students: What Is the Role of Instructor Relationships?," *Community Coll. Rev.*, vol. 48, no. 3, pp. 277–302, 2020, doi: 10.1177/0091552120909908.

[3] A. Alhadabi and A. C. Karpinski, "Grit, self-efficacy, achievement orientation goals, and academic performance in University students," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 519–535, 2020, doi: 10.1080/02673843.2019.1679202.

[4] M. J. Gormley, G. J. DuPaul, L. L. Weyandt, and A. D. Anastopoulos, "First-Year GPA and Academic Service Use Among College Students With and Without ADHD," *Physiol. Behav.*, pp. 1766–1779, 2019, doi: 10.1177/1087054715623046.First-Year.

[5] S. Chaturapruek, T. S. Dee, R. Johari, R. F. Kizilcec, and M. L. Stevens, "How a data-driven course planning tool affects college students' GPA: Evidence from two field experiments," *Proc. 5th Annu.*

ACM Conf. Learn. Scale, L S 2018, 2018, doi: 10.1145/3231644.3231668.

[6] D. Aggarwal and D. Sharma, "Application of clustering for student result analysis," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 50–53, 2019.

[7] A. Almasri, R. S. Alkhawaldeh, and E. Çelebi, "Clustering-Based EMT Model for Predicting Student Performance," *Arab. J. Sci. Eng.*, vol. 45, no. 12, pp. 10067–10078, 2020, doi: 10.1007/s13369-020-04578-4.

[8] D. S. Lamb, J. Downs, and S. Reader, "Space-time hierarchical clustering for identifying clusters in spatiotemporal point data," *ISPRS Int. J. Geo-Information*, vol. 9, no. 2, 2020, doi: 10.3390/ijgi9020085.

[9] L. Zappia and A. Oshlack, "Clustering trees: a visualization for evaluating clusterings at multiple resolutions," *Gigascience*, vol. 7, no. 7, pp. 1–9, 2018, doi: 10.1093/gigascience/giy083.

[10] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[11] Mardonov, "Structure and Mechanisms of Action of The Educational Cluster," *Int. J. Psychol. Rehabil.*, vol. 24, no. 07, pp. 1475–7192, 2020, [Online]. Available: https://hozir.org/pars_docs/refs/541/540182/540182.pdf.

[12] L. Zahrotun, N. hutami Putri, and A. N. Khusna, "The Implementation of K-Means Clustering Method in Classifying Undergraduate Thesisi Titles," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018.

[13] A. Hadifar, L. Sterckx, T. Demeester, and C. Develder, "A self-training approach for short text clustering," *ACL 2019 - 4th Work. Represent. Learn. NLP, RepL4NLP 2019 - Proc. Work.*, no. 2017, pp. 194–199, 2019, doi: 10.18653/v1/w19-4322.

[14] E. A. Anaam, S.-C. Haw, and P. Naveen, "Applied Fuzzy and Analytic Hierarchy Process Techniques in Hybrid Recommendation Approaches For E-CRM," *Int. J. Informatics Vis.*, vol. 6, no. 2, p. 2, 2022.

[15] H.-S. Park and C.-H. Jun, "Expert Systems with Applications An International Journal," *Expert Syst. Appl.*, vol. 145, no. 2, p. 3341, 2020.

[16] D. Sun, H. Fei, and Q. Li, "A Bisecting K-Medoids clustering Algorithm Based on Cloud Model," vol. 51, no. 11, pp. 308–315, 2018, doi: 10.1016/j.ifacol.2018.08.301.

[17] Martanto, S. Anwar, C. L. Rohmat, F. M. Basysyar, and Y. A. Wijaya, "Clustering of internet network usage using the K-Medoid method," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012036, 2021, doi: 10.1088/1757-899x/1088/1/012036.

[18] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means," *Am. J. Distance Educ.*, vol. 34, no. 2, pp. 137–156, 2020, doi: 10.1080/08923647.2020.1696140.

[19] S. Sinche *et al.*, "Analysis of Student Academic Performance Using Human-in-the-Loop Cyber-Physical Systems," *Telecom*, vol. 1, no. 1, pp. 18–31, 2020, doi: 10.3390/telecom1010003.

[20] O. Tinuke Omolewa, A. Taye Oladele, A. Adekanmi Adeyinka, and O. Roseline Oluwaseun, "Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions," *J. Eng. Appl. Sci.*, vol. 14, no. 22, pp. 8254–8260, 2019, doi: 10.36478/jeasci.2019.8254.8260.

[21] J. Oyelade *et al.*, "Data Clustering: Algorithms and Its Applications," *Proc. - 2019 19th Int. Conf. Comput. Sci. Its Appl. ICCSA 2019*, no. July, pp. 71–81, 2019, doi: 10.1109/ICCSA.2019.000-1.

[22] A. Naeem, M. Rehman, M. Anjum, and M. Asif, "Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm," *Curr. Sci.*, vol. 117, no. 6, pp. 1045–1053, 2019, doi: 10.18520/cs/v117/i6/1045-1053.

[23] S. Bipasha Biswas and M. Tariq Iqbal, "Solar Water Pumping System Control Using a Low Cost ESP32 Microcontroller," *Can. Conf. Electr. Comput. Eng.*, vol. 2018-May, pp. 1–5, 2018, doi: 10.1109/CCECE.2018.8447749.

[24] M. T. Lwin and M. M. Aye, "A Modified Hierarchical Agglomerative Approach for Efficient Document Clustering System," *Am. Sci. Res. J. Eng.*, vol. 29, no. 1, pp. 228–238, 2017, [Online]. Available: http://asrjetsjournal.org/.

[25] W. Xiaochun and W. Xia Li, " A Fast K -medoids Clustering Algorithm for Image Segmentation based Object Recognition ," *J. Robot. Autom.*, vol. 4, no. 1, pp. 202–211, 2020, doi: 10.36959/673/371.

[26] F. Gullo, G. Ponti, A. Tagarelli, and S. Greco, "An information-theoretic approach to hierarchical clustering of uncertain data," *Inf. Sci. (Ny).*, vol. 402, pp. 199–215, 2017, doi: 10.1016/j.ins.2017.03.030.

[27] A. Triayudi and I. Fitri, "Comparison of parameter-free agglomerative hierarchical clustering methods," *ICIC Express Lett.*, vol. 12, no. 10, pp. 973–980, 2018, doi: 10.24507/icicel.12.10.973.

[28] A. R. Mamat, F. S. Mohamed, M. A. Mohamed, N. M. Rawi, and M. I. Awang, "Silhouette index for determining optimal k-means clustering on images in different color models," *Int. J. Eng. Technol.*, vol. 7, no. April, pp. 105–109, 2018, doi: 10.14419/ijet.v7i2.14.11464.

[29] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, pp. 1–17, 2021, doi: 10.3390/e23060759.

[30] R. J. Roiger, *Data Mining A Tutorial-Based Primer*. Boca Raton, London, New York, 2017.

[31] N. Nidheesh, K. A. A. Nazeer, and P. M. Ameer, "A Hierarchical Clustering algorithm based on Silhouette Index for cancer subtype discovery from genomic data," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11459–11476, 2020, doi: 10.1007/s00521-019-04636-5.

[32] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 569, no. 5, 2019, doi: 10.1088/1757-899X/569/5/052024.

[33] J. Han, J. Pei, and H. Tong, *Data Mining Concepts and Techniques*. Cambridge, MA 02139, United States: Elsevier Inc., 2023.