



K-Means for Majoring Informatics Students' Interests Based on Brainwave Signals

¹Qori Aulia Robin, ^{2,*}Ahmad Azhari

^{1,2} Universitas Ahmad Dahlan, Yogyakarta, Indonesia

¹qoriauliaronbin@gmail.com, ^{2,*}ahmad.azhari@tif.uad.ac.id

*correspondence email

Abstract

This study investigates the potential of utilizing EEG (electroencephalogram) as a determinant for the specialization choices of Informatics students. EEG, measuring brain activity patterns, is employed to discern majors of interest among students. A questionnaire revealed that some students opt for specializations due to class availability and peer influence, leading to potential mismatches between their abilities and interests, consequently affecting their final project or thesis. EEG data from 30 respondents, recorded using NeuroSky Mindwave and MyndPlayer Pro software, were subjected to K-Means Clustering after feature extraction through PCA. However, the evaluation using Silhouette indicated a low score of 0.453, possibly due to significant distance between cluster data and centroids, minimal dataset size, and random respondent selection without considering their specific areas of interest. This suggests limitations in using EEG alone for determining specialization choices, necessitating further refinement and integration with additional factors for more accurate predictions.

Keywords: Brain Wave, K-Means, Clustering, Principal Component Analysis, Silhouette

INTRODUCTION

The brain is the control center of all human activities which is the center of communication and body decisions [1]. Brain activity in the form of communication between neurons generates flow and creates brain wave signals that can only be known by recording using an Electroencephalogram (EEG) [2].

The brain waves recorded by the EEG are in the frequency between 0.5 Hz to 100 Hz. Brain wave signals from EEG recordings can identify all conditions of a person, including in a state of not doing anything, full of concentration and thinking until a person's condition is in a state of high mental activity such as panic and fear [3]. The recording results can be used to evaluate a person based on the activities carried out during the process of recording brain wave signals. The recording results can be used to evaluate a person based on the activities carried out during the process of recording brain wave signals in answering or deciding on a certain condition [4].

The activity is used as a recorded brain wave stimulus. One of the activities that have been used in previous research is completing a learning ability test or can be called an achievement test in the form of a Basic Mathematical Test in Yumiko, Triroasmoro, and Fauzi's research in 2021 [5]. In this study, the achievement test was used as a stimulus to obtain an EEG signal in determining the majors in the field of interest in the Informatics study program at Ahmad Dahlan University. Areas of interest that will be offered include "sistem cerdas" and "relata". Students begin to determine their area of interest in the sixth semester by taking several relevant specialization courses. The reality is that not all students can confidently choose an area of interest based on their abilities even though there is data on learning outcomes as material for decision considerations.

Based on the results of a questionnaire to 30 Informatics students who are in the 6th semester of college to students who have just finished college, there are 12 students taking an area of interest based on their ability in the chosen field of interest compared to other fields of interest, 14 students taking an interest field because it is based on their interests and interests. interest in their specialization courses as well as the development of knowledge in the field of interest, 2 students confidently determined the field of interest because of their abilities and interests, but the other 2 students chose the field of interest because the class was full and unsure of their abilities and interests so they followed their friends when choosing courses interest. This reason can be a factor in changes in the field of interest and unpreparedness of students in the process of working on their final project or thesis.

These factors are the reason for conducting research that can determine the majors in the field of interest based on EEG data, achievement test data and learning outcomes data. The research was conducted by recording brain wave signals using EEG and accompanied by a stimulus as a brain wave signal stimulant in the form of working on questions from achievement tests. The achievement test is a collection of basic course questions that represent each area of interest.

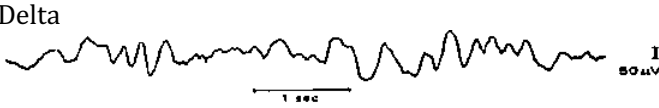
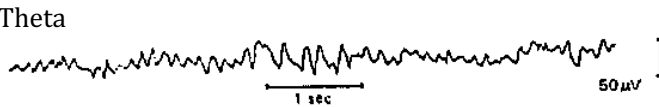
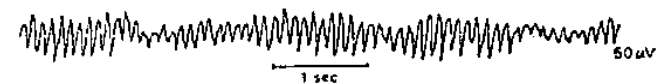
EEG data, stimulus result data and value transcript data will be processed using the k-means clustering method. The k-means method is a clustering method that is widely used because it includes a simple unsupervised clustering technique and can be used for large datasets [6]. This method groups data into a cluster that has similar characteristics with one another, so that data with different characteristics will be included in other clusters [2], [6], [7].

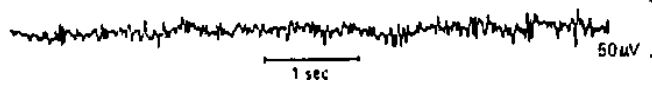

LITERATURE REVIEW

Brain Wave Signal

Brain wave signals are divided into 5 categories based on frequency, including delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), gamma (>30 Hz) . Each category can identify a person's condition [7], [8]. These categories can be seen in table 1 categories of brain wave signals.

Table 1. Categories of Brain Wave Signal

Signal	Frequency (Hz)	Condition
Delta 	0.5 – 4	A sleep without dreams or coma. It can also be an indication if there is a brain defect.
Theta 	4 – 8	Light sleep, daydreaming or praying solemnly.
Alpha 	8 – 13	In a state between conscious and unconscious as well as a state of relaxation.
Beta	13 – 30	Thinking, full concentration, doing

		daily activities or a state of relaxation.
<p>Gamma</p> 	>30	In high mental activity, for example panic, fear, and others.

Field of Interest

The Informatics Engineering Study Program at Ahmad Dahlan University concentrates on 2 areas of interest, namely “sistem cerdas” and “relata”. The field of interest determines the topic of the student's thesis. Students choose the field of interest in the 6th semester.

METHODS

Data Collection

This study acquired data from 30 respondents. Respondents are students of Informatics Engineering from Ahmad Dahlan University at least semester 5, because respondents are students who have or will choose a major in the field of interest. For each respondent 3 data were taken, namely, brain wave signal data, research stimulus data, and value transcript data. The brain wave signal data collection was done by recording the respondent's brain wave signal which was taken using an Electroencephalogram (EEG) and MyndPlayerPro software. During the recording process the respondents were given a test as a research stimulus and the time given to complete the entire test was 30 minutes. The recording result is still a file with *.log format and must be reprocessed with the same software to convert it into *.csv format data that is ready to proceed to the next stage.

Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is the initial stage of processing EEG data. At this stage, the EEG data will be analyzed and examined to find patterns, find anomalies, test hypotheses and check assumptions using statistical calculations [8], [9]. EDA is used to analyze and select the best data variables to be processed at the extraction stage to data grouping.

Feature Extraction

Feature extraction is the process of separating EEG signals based on their feature categories with the aim of producing distinguishing characteristics between one object and another. In this study, the feature extraction process goes through two stages of extraction, namely Order I and the Principal Component Analysis (PCA) method.

Orde I

Orde I is a method with a statistical approach with several parameters including the mean, median, standard deviation, skewness, and kurtosis.

- **Mean (\bar{x})**

Used to determine the mean of the data distribution. The means equation can be seen in equation 1.

$$\bar{x} = \frac{x_1+x_2+x_3+\dots+x_n}{n} \quad (1)$$

Where n is the number of data.

- **Median (*Med*)**

Used to calculate the mean value of the data distribution. To find the middle value, there is a difference in determining the midpoint when the data is even and odd. For data with an odd number, use the equation in equation 2 below.

$$Med = X \frac{(n+1)}{2} \quad (2)$$

Data with even numbers count using the equation in equation 3 below.

$$Med = \frac{(X(\frac{n}{2}) + X(\frac{n}{2} + 1))}{2} \quad (3)$$

Where Med is the Median, X is the data value, and n is the number of data.

- **Standard Deviation (s)**

Used to measure the value of the standard deviation of the data distribution by finding the value of the variance first. The standard deviation equation can be seen in equation 4.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}} \quad (4)$$

Where $(x - \bar{x})$ is the difference between x and \bar{x} .

- **Skewness (α_3)**

Skewness is used to measure the level of slope of the data distribution. The skewness equation can be seen in equation 5.

$$\alpha_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^3} \quad (5)$$

- **Kurtosis (α_4)**

Kurtosis is used to measure the height distribution of the data. The equation of kurtosis can be seen in equation 6.

$$\alpha_4 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} \quad (6)$$

Principal Component Analysis (PCA)

The PCA method is used to reduce dimensions that have many variables by selecting the most important components with the aim of making calculations more optimal when processing signals [10]–[12]. Data reduction is needed to reduce the complexity of the data, most of which have correlations between other data, and convert the data into small and uncorrelated datasets so that the data is easier to interpret [11].

The PCA algorithm begins with looking for data $X_{i,j}^*$ which has dimensions $m \times n$, where m is the number of samples and n is the number of attributes. And using the zero-mean technique, namely by subtracting all the values of $X_{i,j}$ in the X matrix, with the average value being the matrix value of X. Zero-mean is a process-to-process data into a standard normal distribution. According to the central limit theorem, this is done if the data taken is close to the population, the data is closer to the normal distribution. So the results of these calculations can represent a number of population data. Find the matrix value $X_{i,j}^*$ by using the equation contained in equation 7 below.

$$X_{i,j}^* = X_{i,j} - \bar{X} \quad (7)$$

Then do the calculations to find the covariance value of the $X_{i,j}$ matrix, as shown in equation 8.

$$C_x = \frac{1}{m-1} \times X_{i,j}^{*T} \times X_{i,j}^* \quad (8)$$

Where, C_x is the covariance matrix of $j \times j$, and m is the number of samples. The next step is to find the eigen values, as seen in equation 9.

$$|C_x - \lambda I| = 0 \text{ and } (C_x - \lambda I) \times v = 0 \quad (9)$$

Where, I is the identity matrix, λ is the eigenvalue and v is the eigenvector. Eigenvector is the main component to determine the new variable. To determine the number of new variables used depending on the perception of the cumulative contribution of V_r variation, the calculation of V_r can be seen in equation 10.

$$V_r = \frac{\sum_{j=1}^r \lambda_j}{\sum_{j=1}^D \lambda_j} \times 100\% \quad (10)$$

Where, D is the number of initial attributes and r is the number of selected components. [10]

Clustering

This stage is the stage of grouping data using K-Means Clustering. K-Means Clustering is a non-hierarchical data clustering method that attempts to group data into one or more clusters. The method works by grouping data based on similar characteristics so that data with the same characteristics will be grouped into one cluster and data with different characteristics will be included in other clusters [6], [7], [13]. In processing data, generally k-means clustering uses the following algorithm:

- Determine the number of clusters to be used.
- Determine the initial center point (centroid) at random as many as k . Random centroid determination only applies to the first iteration.
- Calculate the distance between the centroids with all data using the Euclidean distance with the equation contained in equation 11 below:

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} : i = 1, 2, 3 \dots n \quad (11)$$

- Group each data into the nearest cluster.
- Re-determine the centroid value to start the next iteration using the equation contained in equation 12 below.

$$v = \frac{\sum_{i=1}^n x_i}{n} : i = 1, 2, 3 \dots n \quad (12)$$

Repeat Steps 3 and 4, if there is still a change in the position of the centroid. If there is no change in the position of the centroid, the clustering process is complete.

Evaluation

In this study, the evaluation of the system was carried out by analyzing the validity of clustering using the silhouette technique. The Silhouette technique is a comparison of tightness to object separation. Silhouettes can reflect grouped data, so that objects are grouped into clusters that have a match [14]–[17]. Silhouette can be defined by the equation contained in equation 13 below.

$$\bar{S} = \frac{1}{n} \sum_{i=1}^n \left(\frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \right) \quad (13)$$

With, $a(i)$ is the average distance from data i to other data in the same cluster, $b(i)$ is the minimum distance from data i to other clusters [14].

RESULT AND DISCUSSIONS

Data Collection

EEG Data

EEG data was obtained from the results of recording brain wave signals in 30 respondents using EEG tools and MyndPlayer Pro software. The recording process accompanied by giving stimulus to the respondents was carried out within ± 30 minutes. The recorded EEG data will be saved in a file with the format (*.log).

The recording results with the format (*.log) are then imported back into the software and the software will automatically translate the data into waveforms which are divided by category. After getting the waveform based on the data category, it can be exported to get the data in a file with the format (*.csv). This data will be used for further processing. An example of EEG data from one respondent that is ready to be processed can be seen in table 2, an example of EEG data from Respondent 1 below.

Table 2. EEG Data of Respondent 1

'Time'	'Vid Time'	'Signal Level'	'Blink'	'Attention'	'Meditation'	'Zone'	'Delta'	'Theta'	'Low Alpha'	'High Alpha'	'Low Beta'	'High Beta'	'Low Gamma'	'Mid Gamma'
36:27.0	00:00.0	100	1	77	56	66	1274215	152889	5894	44034	18478	16507	12426	7644
36:28.0	00:00.0	100	1	53	50	51	850913	172396	6594	90694	85145	17434	10597	1585
36:29.0	00:00.0	100	1	54	51	52	206847	48925	8582	6521	6177	6386	2348	634
36:30.0	00:00.0	100	1	60	54	57	158397	4104	2961	7816	4414	3006	1373	858
36:31.0	00:00.0	100	1	47	69	58	66988	61996	27002	40018	7960	6621	3321	978
36:32.0	00:00.0	100	1	41	84	62	57888	46257	20982	47780	7628	5669	3227	1785
36:33.0	00:00.0	100	3	56	84	70	491009	154795	30202	17164	23502	43085	9135	4708
36:34.0	00:00.0	100	2	56	78	67	74559	22637	16968	2401	2670	7818	3466	1442
36:35.0	00:00.0	100	2	80	63	71	2059364	277570	40665	76501	92020	61151	18191	10164
36:36.0	00:00.0	100	2	87	38	62	188712	239352	19895	17924	18298	20160	13491	1281

Stimulus Data

The stimulus carried out will be the data needed in this study. The stimulus is a collection of several questions representing each subject that forms the basis for specialization courses from semester one to semester 5. The courses include data structures, artificial intelligence, algorithmic strategies, and automata language theory for areas of interest in "sistem cerdas". And courses on web programming, data communication and computer networks, databases and courses on human and computer interaction for fields of interest. The results of the stimulus are scored for each question, the average value is calculated and labeled based on the average value of the area of interest which is greater than other areas of interest. The stimulus data can be seen in table 3 of the stimulus data.

Table 3. The Stimulus Data

Respondent	Stimulus Score		Label
	<i>Sistem Cerdas</i>	<i>Relata</i>	
1	0.63	0.7	Relata
2	0.85	0.45	Sistem Cerdas
3	0.3	0.7	Relata
4	0.3	0.78	Relata
5	0.5	0.58	Relata
6	0.7	0.45	Sistem Cerdas
7	0.73	0.53	Sistem Cerdas
8	0.5	0.8	Relata
9	0.58	0.7	Relata

10	0.65	0.58	Sistem Cerdas
11	0.65	0.4	Sistem Cerdas
12	0.7	0.63	Sistem Cerdas
13	0.85	0.65	Sistem Cerdas
14	0.55	0.63	Relata
15	0.45	0.7	Relata
16	0.5	0.65	Relata
17	0.23	0.63	Relata
18	0.75	0.7	Sistem Cerdas
19	0.75	0.7	Sistem Cerdas
20	0.28	0.58	Relata
21	0.23	0.63	Relata
22	0.23	0.38	Relata
23	0.7	0.75	Relata
24	0.35	0.35	Relata
25	0.15	0.15	Relata
26	0.48	0.5	Relata
27	0.75	0.75	Relata
28	0.63	0.53	Sistem Cerdas
29	0.3	0.65	Relata
30	0.23	0.45	Relata

Transcript Data

The transcript data was obtained through an online questionnaire by the respondents. The data is then searched for the average value for the basic courses from each area of interest. The average value obtained from each area of interest can be seen in table 4 of the Transcript Data.

Table 4. The Transcript Data

Respondent	Transcript	
	<i>Sistem Cerdas</i>	<i>Relata</i>
1	0.89	0.86
2	0.9	0.91
3	0.77	0.77
4	0.9	0.93
5	0.91	0.9
6	0.9	0.9
7	0.9	0.89
8	0.94	0.96
9	0.73	0.76
10	0.83	0.86
11	0.84	0.89
12	0.84	0.89
13	0.85	0.84
14	0.84	0.85
15	0.8	0.83
16	0.81	0.82
17	0.93	0.95
18	0.88	0.91
19	0.92	0.92
20	0.76	0.81

21	0.93	0.95
22	0.67	0.79
23	0.91	0.93
24	0.68	0.75
25	0.94	0.91
26	0.8	0.85
27	0.88	0.91
28	0.8	0.83
29	0.89	0.86
30	0.77	0.84

Extraction Data Analysis (EDA)

Data analysis is the stage to find the best data variables using Exploratory Data Analysis (EDA) techniques. The data analyzed is only EEG data from one of the respondents by displaying data variables that have the possibility to be used for processing and have a good data distribution. The data variables analyzed were “Low Beta” and “High Beta” because these waves were waves obtained under conditions of concentration; "Attention", "Meditation" and "Zone" variables because the data is the result of brain wave processing by MyndPlayer Pro software which becomes data based on the conditions of each variable. The analysis shows that the data with the best distribution is in the variables "Zone" to "Meditation" and "Zone" to "Attention" as shown in Figure 1 the results of exploratory data analysis.

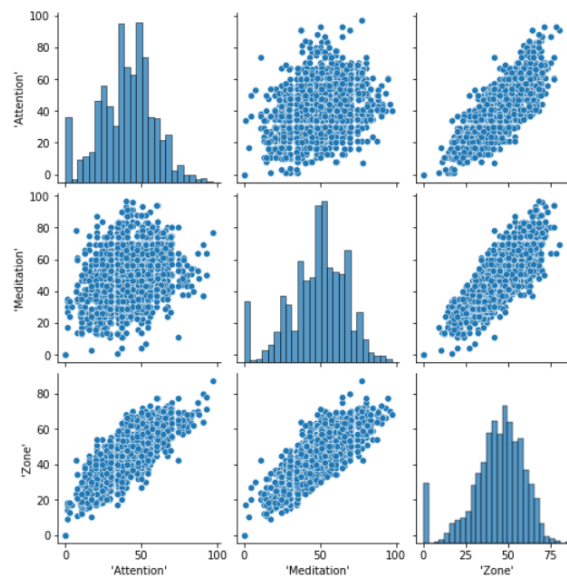


Fig 1. The Results of Exploratory Data Analysis

The distribution pattern of the same variable only shows an increase in the middle and continues to decrease until the end. The variable "Meditation" to "Attention" shows the pattern of data distribution but the distribution is too wide and irregular. The variables “Zone” against “Meditation” and “Zone” against “Attention” show regular graphic patterns. The increase in the graph in the "Zone" variable is directly proportional to the increase in the "Meditation" and "Attention" variables, and vice versa.

So it is determined that the variable used in this system is a variable "Zone" to "Attention". The variable “Zone” against “Meditation” was not selected because this variable is derived and generated from brain waves when a person is in a state of meditation or relaxation. Whereas in

this study when the process of recording brain wave signals, the respondents were in a state of full thinking and concentration [18].

Feature Extraction

First Orde

The data that has been selected from the results of the EDA analysis will be extracted to obtain its characteristic features. Feature extraction will be carried out using two methods, namely Order I and PCA. The first-order method was performed on EEG data from EDA to obtain the mean, median, standard deviation, skewness, and kurtosis. The results of the first-order feature extraction can be seen in table 5 of the first-order feature extraction.

Table 5. The First-order Feature Extraction

Responden t	Attention					Zone				
	mean	med	std	skew	kurt	mean	med	std	skew	kurt
1	47.51	48	21.45	-0.18	0.00	49.06	51	17.23	-0.93	1.37
2	40.44	41	17.61	-0.19	0.08	44.01	45.5	15.55	-0.84	0.95
3	52.89	51	17.94	0.17	-0.17	53.24	54	12.37	-0.35	-0.02
4	57.80	57	18.44	-0.15	-0.22	55.32	56	11.18	-0.41	-0.10
5	45.42	47	16.55	-0.01	-0.16	51.87	53	12.68	-0.23	-0.40
6	44.54	44	18.32	0.13	-0.25	51.47	51.5	13.28	-0.01	-0.26
7	49.20	51	19.46	-0.62	0.34	49.45	52	16.19	-1.13	1.79
8	53.30	53	18.76	-0.08	-0.27	53.96	54	13.75	-0.10	-0.24
9	53.50	53	17.92	0.10	-0.27	58.31	59	14.03	-0.17	-0.30
10	55.92	57	20.56	-0.63	0.44	51.55	54	15.03	-1.21	2.21
11	54.49	54	16.75	-0.13	-0.18	53.11	54	12.89	-0.37	-0.10
12	57.96	57	18.64	-0.16	-0.05	58.85	60	13.57	-0.43	0.43
13	43.45	44	22.31	-0.10	-0.07	48.20	51	19.59	-0.98	0.74
14	57.92	57	15.84	-0.09	-0.07	56.48	57	12.27	-0.35	-0.05
15	49.20	50	16.67	-0.30	-0.04	55.17	56	13.14	-0.44	0.09
16	41.16	43	28.00	-0.07	-1.02	44.73	50	25.12	-0.63	-0.64
17	37.44	38	27.07	0.25	-0.72	38.19	42	22.97	-0.36	-0.75
18	42.01	48	26.32	-0.38	-1.03	38.77	45	21.97	-0.66	-0.78
19	46.89	47	17.07	0.20	-0.15	55.90	56	12.55	-0.13	-0.11
20	46.83	48	16.63	-0.07	0.23	51.50	52	12.95	-0.32	0.88
21	56.67	57	15.73	-0.25	0.47	52.49	53	11.61	-0.28	0.98
22	45.12	44	17.46	0.27	-0.03	52.84	53	12.73	-0.13	-0.23
23	53.25	53	20.36	-0.08	-0.66	50.35	51	13.96	-0.27	-0.63
24	46.85	48	16.87	-0.34	0.36	50.60	51	14.52	-0.48	0.91
25	38.65	38	22.25	0.30	-0.28	43.47	44	16.85	-0.11	0.71
26	53.57	54	18.19	-0.12	-0.21	55.80	57	13.69	-0.29	-0.33
27	46.86	48	21.27	-0.34	-0.13	48.19	51	17.95	-1.02	1.11
28	45.42	48	21.65	-0.21	-0.18	49.30	51	18.73	-0.70	0.61
29	61.86	63	20.49	-0.36	-0.06	60.18	61	13.88	-0.46	0.14
30	59.41	61	21.14	-0.92	1.11	52.87	55	15.96	-1.38	2.65

Principal Component Analysis (PCA)

Before grouping the data, the data to be used must be reduced to minimize the data by extracting the most important information. This reduction process needs to be done because the python 3 programming language can only perform k-means clustering with 2 dimensions. The results of the data reduction process using PCA can be seen in table 6 of PCA feature extraction.

Table 6. PCA Feature Extraction

PCA Feature Extraction	
1.92966934e-01	4.26309857e-01
4.85886236e-01	2.53497434e-01
-2.05273057e-01	-3.25560304e-01
-5.13876122e-01	-1.96529205e-01
8.00904129e-04	-4.12205682e-01
1.84755928e-01	-5.13568290e-01
-1.00222287e-01	7.34535901e-01
-2.10696064e-01	-3.48869548e-01
-3.83526164e-01	-4.51107931e-01
-4.21211098e-01	8.62985810e-01
-3.31232108e-01	-2.17333844e-01
-6.30746815e-01	-8.97962931e-02
4.75357534e-01	3.50955039e-01
-6.12913010e-01	-2.61047525e-01
-2.85947729e-01	-1.30908226e-01
1.07549944e+00	9.84667662e-02
1.42571529e+00	-1.26945948e-01
1.02535618e+00	2.12039407e-01
-1.35705796e-01	-5.16803238e-01
-1.03948683e-01	-1.04952008e-01
-5.76262732e-01	7.73144204e-03
7.51490261e-02	-5.24989729e-01
2.24761611e-02	-3.01910942e-01
-6.58828248e-02	1.24083428e-01
8.72370658e-01	-2.26599089e-01
-3.47717526e-01	-2.83923294e-01
2.38735642e-01	4.89841787e-01
3.14145591e-01	2.32872921e-01
-7.91806725e-01	2.15095640e-02

Clustering

The reduced data using PCA was then grouped using the k-means clustering method and the Euclidean distance formula to measure the closest distance of the data to the centroid. The visualized grouping results can be seen in Figure 3 visualization of k-means clustering.

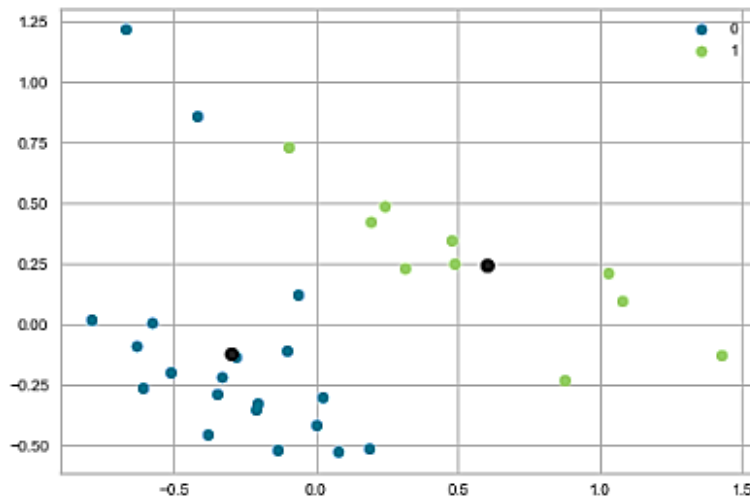


Fig 2. Visualization of K-Means Clustering

Based on the clustering process, the percentage of interest areas based on clusters is obtained, namely cluster 0 has 75% with a temporary label “relata” and 25% with a temporary label “sistem cerdas”. In contrast to cluster 1, the percentage is 50% with the temporary label “relata” and 50% with the temporary label “sistem cerdas”. Respondents and their temporary labels that fall into each cluster can be seen in table 7 field of interest based on clusters.

Table 7. Field of Interest Based on Clusters

Cluster 0		Cluster 1	
Respondent	Temporary Label	Respondent	Temporary Label
2	Relata	0	Relata
3	Relata	1	Sistem Cerdas
4	Relata	6	Sistem Cerdas
5	Sistem Cerdas	12	Sistem Cerdas
7	Relata	15	Relata
8	Relata	16	Relata
9	Sistem Cerdas	17	Sistem Cerdas
10	Sistem Cerdas	24	Relata
11	Sistem Cerdas	26	Relata
13	Relata	27	Sistem Cerdas
14	Relata		
18	Sistem Cerdas		
19	Relata		
20	Relata		
21	Relata		
22	Relata		
23	Relata		
25	Relata		
28	Relata		
29	Relata		

Evaluation

At the evaluation stage, the silhouette score was 0.453. The small score obtained can be caused by the distance between the data in the cluster and the far centroid as can be seen in Figure 3

Visualization of Kmeans Clustering. Silhouette visualization results can be seen in Figure 4 Silhouette Visualization.

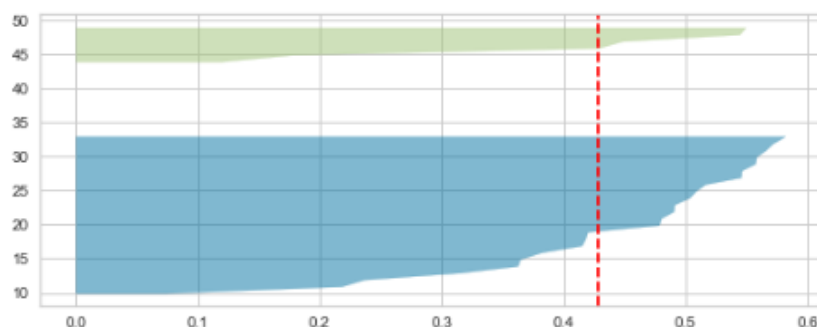


Fig 3. Silhouette Visualization

CONCLUSION

The results of the grouping show that cluster 0 or relata cluster contains 20 objects with 15 objects of interest in “relata” and 5 objects of interest in “sistem cerdas”. Comparison of the percentage of areas of interest in cluster 0 is 75% “relata” and 25% “sistem cerdas”. In cluster 1 or a “sistem cerdas” cluster that has 5 objects and 5 objects of “relata” interest. The comparison of the percentage of interest fields in cluster 1 is 50% of the field of interest in “relata” and 50% of the field of interest in “sistem cerdas”.

Based on the evaluation results, the silhouette results show a small score of 0.453. This can be caused because the data in the cluster and the centroid has a long distance. Another factor that can be the cause of the small silhouette score is because the data used is too minimal and the data collection is done by random respondents without considering the comparison of the areas of interest that the respondents have chosen.

REFERENCE

- [1] A. Azhari, “Analisis Pengaruh Cognitive Task Berdasarkan Hasil Ekstraksi Ciri Gelombang Otak Menggunakan Jarak Euclidean,” p. 6, 2017.
- [2] A. Azhari and L. Hernandez, “Brainwaves feature classification by applying K-Means clustering using single-sensor EEG,” *Int. J. Adv. Intell. Informatics*, vol. 2, no. 3, p. 167, Nov. 2016, doi: 10.26555/ijain.v2i3.86.
- [3] Y. Akbar, “Pola Gelombang Otak Abnormal Pada Elektroencephalograph,” p. 6.
- [4] M. Teplan, “Fundamentals Of eeg Measurement,” *Measurement Science Review*, vol. 2, p. 11, 2002.
- [5] F. Yumiko, I. I. Tritasmoro, and H. Fauzi, “Klasifikasi Sinyal Eeg Terhadap Konsentrasi Individu Menggunakan Metode K-Nearest Neighbor,” p. 16.
- [6] Md. Z. Hossain, Md. N. Akhtar, R. B. Ahmad, and M. Rahman, “A dynamic K-means clustering for data mining,” *IJEECS*, vol. 13, no. 2, p. 521, Feb. 2019, doi: 10.11591/ijeecs.v13.i2.pp521-526.
- [7] K. P. Sinaga and M.-S. Yang, “Unsupervised K-Means Clustering Algorithm,” *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
- [8] S. K. Mukhiya and U. Ahmed, *Hands-On Exploratory Data Analysis with Python*. Packt Publishing Ltd., 2020.
- [9] S. Morgenthaler, “Exploratory data analysis,” *WIREs Comp Stat*, vol. 1, no. 1, pp. 33–44, Jul. 2009, doi: 10.1002/wics.2.

-
- [10] R. Pujiyanto and A. A. Rahmawati, "Analisis Ekstraksi Fitur Principle Component Analysis pada Klasifikasi Microarray Data Menggunakan Classification And Regression Trees," p. 12.
- [11] O. D. Nurhayati, D. Eridani, and A. Ulinuha, "Ekstraksi Ciri Orde Pertama dan Metode Principal Component Analysis untuk Mengidentifikasi Jenis Telur Ayam Kampung dan Ayam Arab," *J. Sistem Info. Bisnis*, vol. 9, no. 2, p. 133, Nov. 2019, doi: 10.21456/vol9iss2pp133-140.
- [12] S. K. Prabhakar and H. Rajaguru, "PCA and K-means clustering for classification of epilepsy risk levels from EEG signals — A comparative study between them," in *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Okinawa, Japan, Nov. 2015, pp. 83–86. doi: 10.1109/ICIIBMS.2015.7439467.
- [13] M. H. Dunham, "Data Mining Introductory and Advanced Topics," pp. 140–141.
- [14] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient Untuk Evaluasi Clustering Obat," vol. 6, no. 2, p. 7, 2021.
- [15] S. Aranganayagi and K. Thangavel, "Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure," in *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*, Sivakasi, Tamil Nadu, India, Dec. 2007, pp. 13–17. doi: 10.1109/ICCIMA.2007.328.
- [16] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, "The Clustering Validity with Silhouette and Sum of Squared Errors," in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 2015, pp. 44–51. doi: 10.12792/iciae2015.012.
- [17] 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, p. 052024, Jul. 2019, doi: 10.1088/1757-899X/569/5/052024.
- [18] S. Morshad, Md. R. Mazumder, and F. Ahmed, "Analysis of Brain Wave Data Using Neurosky Mindwave Mobile II," in *Proceedings of the International Conference on Computing Advancements*, Dhaka Bangladesh, Jan. 2020, pp. 1–4. doi: 10.1145/3377049.3377053.