Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Comparison and Evaluation of Euclidean Distance and Dice Distance in the K-Means Adaptive Algorithm for Clustering Composite Indexes of Food Security and Vulnerability Maps

Emma Romasta Naulina Nainggolan[1], Paska Marto Hasugian[2]
[1,2]Faculty of computer sciene, University Katolik Santo Thomas, Medan, Indonesia

Article Info
Keywords:
Adaptive K-means,
Euclidean Distance,
Dice Distance,
Silhoutte Score,
Clustering,
Food Resilience and
Vulnerability index

ABSTRACT
This study aims to compare and evaluate the effectiveness of two distance measurement methods, namely Euclidean Distance and Dice Distance, in the K-Means Adaptive algorithm for clustering Food Security and Vulnerability Composite Index data. The dataset used includes index data from 2022 to 2024, comprising 305 entries, which were then cleaned to 298 entries. The evaluation was conducted manually using a sample dataset and automatically using the entire dataset via Google Colab with Python. The algorithm's performance was assessed using the Silhouette Score metric to measure the quality of the resulting clusters. The evaluation results showed that the Euclidean method produced an average Silhouette Score of 0.3082, indicating an suboptimal cluster structure. This study concludes that the choice of distance method significantly influences clustering results, and selection should be tailored to the characteristics of the data.

Corresponding Author:
Emma Romasta Naulina Nainggolan
University Katolik Santo Thomas, Medan, Indonesia
emmaromasta@gmail.com

INTRODUCTION

Mapping food security and vulnerability is an important step in making sustainable village development policies. Composite Index data that reflects the condition of village food security and vulnerability from year to year can be utilized to conduct further analysis, such as classification and clustering of regions. One method that can be used for this purpose is Adaptive K-Means, a variant of the K-Means algorithm that adjusts the centroid determination dynamically.

In its implementation, the selection of distance method plays an important role in determining the quality of clustering results. Therefore, this study compares two different distance measurement methods, namely Euclidean Distance, which is commonly used for continuous data, and Dice Distance, which is more suitable for binary data or categorical transformation. Utilizing index data from 2022 to 2024, this study evaluates the performance of each method in the clustering process and measures their effectiveness using the Silhouette Score. The evaluation is done both manually (with samples) and automatically (with the entire dataset), in order to obtain more objective and representative results.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

METHODS

This research uses a quantitative approach with an experimental method to compare two types of distance measurements in the Adaptive K-Means algorithm, namely Euclidean Distance and Dice Distance. The object of research is the data of the Composite Index of Food Security and Vulnerability of 305 villages for three years (2022-2024), obtained from the official source data.go.id. After going through the data cleaning process, the amount of data used was reduced to 298 entries.

Data Preparation

The data used is the Composite Index of food security and vulnerability maps for three consecutive years: 2022, 2023, 2024, which includes 305 entries. Each entry represents the index value of one village area in each of these years. Purpose of the data. This data can be used to analyze changes or trends in village development from year to year, classify village development status, predict the next year's index, and cluster villages based on development characteristics.

Initial Data

Based on data sourced from data.go.id (https://data.go.id/dataset/dataset/jumlah-penduduk-yang-mengalami-ketidakcukupan-konsumsi-pangan-provinsi-update-tahun-2024), contains data on the village development index in 2024 as follows:

Table 1. Original Data

No	Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024
1	59,32	53,13	56,51
2	60,98	57,51	59,18
3	69,83	45,37	53,50
4	60,17	51,80	56,03
5	62,31	54,12	58,05
..	.....	.....	....
303	55,28	55,28	50,28
304	48,68	45,42	51,62
305	49,27	51,39	55,44
306	56,44	52,34	59,47

After Data Cleaning

Before the data is used in the modeling process, an important stage is carried out, namely data cleaning or data cleaning. The purpose of data cleaning is to ensure that the data used is quality, valid, and does not contain errors that can affect the results of the analysis. After the cleaning process was carried out, the amount of data was reduced from 305 rows to 298 rows. Thus, 7 rows of data have been deleted because they do not meet the data eligibility criteria.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Tabel 2. Cleaning Data

No	Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024
1	59,32	53,13	56,51
2	60,98	57,51	59,18
3	69,83	45,37	53,50
4	60,17	51,80	56,03
5	62,31	54,12	58,05
...	.....	.....	....
294	51,50	38,62	39,11
295	59,78	52,66	58,87
296	65,16	47,86	54,86
297	52,46	47,38	56,65
298	54,30	43,34	49,39

Euclidean Distance against K-Means

In the initial stage of applying the K-Means clustering algorithm with the euclidean distance method, the Composite Index data of the 2022-2024 village food security and vulnerability map was used. The initial dataset consists of 305 data, then the data cleaning process is carried out to ensure that there are no missing values, duplication, and extreme values (outliers) that can affect the accuracy of the clustering results.

However, due to limitations in performing manual calculations on all data, a random sampling of 25 data from a total of 305 data was carried out. This sampling aims to simplify the process of calculating the Euclidean distance between data and visualizing the clustering process manually.

Using these 25 sample data, the process of calculating the Euclidean distance between data points, determining the cluster center (centroid), and grouping data into clusters based on proximity.

Sample Data

The sample data used in manual testing is as follows:

Table 2 sample data.

No	Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024
1	59,32	53,13	56,51
2	60,98	57,51	59,18
3	69,83	45,37	53,50
4	60,17	51,80	56,03
5	62,31	54,12	58,05
6	67,03	62,15	62,78
7	55,71	50,99	56,51
8	56,75	46,56	51,52
9	51,12	44,40	55,74
10	58,55	51,79	55,41

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

11	41,18	48,34	42,29
12	46,24	41,34	48,94
13	58,43	57,60	58,47
14	43,68	49,19	42,21
15	40,11	48,69	44,58
16	37,66	42,39	37,32
17	36,07	45,45	37,20
18	61,32	54,78	55,49
19	32,41	63,46	34,82
20	55,68	71,12	53,00
21	42,92	52,82	36,26
22	62,60	46,20	47,82
23	62,01	51,76	52,09
24	63,74	46,11	52,24
25	66,81	52,82	60,60

Determining the Initial Centroid

The initial centroid selection for the first cluster (C1) is done randomly based on the data available after the cleaning process, which is 305 data. From these data, 25 random data samples were taken to facilitate the manual calculation process in testing the K-Means algorithm.

In this test, the initial centroid is extracted from the 13th data in the sample, which has the following values:

C1 = (600.1342, 55.2287, 58.9931)

(where the numbers represent: Composite Index of 2022, 2023, and 2024)

a. Calculating the Distance to Determine the Next Centroid

To determine the next centroid, the Euclidean distance formula is used as follows:

D(Xi)2 = ||Xi-Cj||2 = (x1-c1)2 + (x2-c2)2+ (x3-c3)2

Where:

Xi is the i-th data vector

Cj is the centroid of cluster j

This calculation is performed on all data in the sample to determine the data that is furthest from the initial centroid, which is then used as the second centroid (C2). The following are the final results of 25 sample data in calculating the distance from data to the initial centroid C1).

Table 3. Initial Centroid Results (C1)

Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024	Perhitungan
59,32	53,13	56,51	D(Xi)^2 untuk C2
60,98	57,51	59,18	13,410
69,83	45,37	53,50	16,065

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024	Perhitungan
60,17	51,80	56,03	0,000
62,31	54,12	58,05	11,879
67,03	62,15	62,78	12,405
55,71	50,99	56,51	19,379
56,75	46,56	51,52	15,495
51,12	44,40	55,74	13,282
58,55	51,79	55,41	18,862
41,18	48,34	42,29	13,119
46,24	41,34	48,94	30,907
58,43	57,60	58,47	24,360
43,68	49,19	42,21	17,447
40,11	48,69	44,58	28,736
37,66	42,39	37,32	31,207
36,07	45,45	37,20	36,133
61,32	54,78	55,49	37,491
32,41	63,46	34,82	12,841
55,68	71,12	53,00	45,571
42,92	52,82	36,26	29,389
62,60	46,20	47,82	32,820
62,01	51,76	52,09	9,232
63,74	46,11	52,24	10,197
66,81	52,82	60,60	6,264
	Total		10,723

Determining the Proportionality of the Next Centroid

$$P(x_i) = \frac{D(X_i)^2}{\sum_{i=1}^{n} D(X_i)^2}$$

Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024	D(Xi)^2	Probalitas C2	Probalitas (%)
59,32	53,13	56,51			
60,98	57,51	59,18	13,41	0,027564744	2,756474426
69,83	45,37	53,50	16,065	0,033022194	3,302219363
60,17	51,80	56,03	0	0	0
62,31	54,12	58,05	11,879	0,024417718	2,44177179
67,03	62,15	62,78	12,405	0,02549893	2,549893009

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024	D(Xi)^2	Probalitas C2	Probalitas (%)
59,32	53,13	56,51			
55,71	50,99	56,51	19,379	0,039834242	3,983424154
56,75	46,56	51,52	15,495	0,031850538	3,185053783
51,12	44,40	55,74	13,282	0,027301636	2,730163559
58,55	51,79	55,41	18,862	0,038771529	3,877152918
41,18	48,34	42,29	13,119	0,026966583	2,696658314
46,24	41,34	48,94	30,907	0,063530466	6,353046613
58,43	57,60	58,47	24,36	0,050072869	5,007286877
43,68	49,19	42,21	17,447	0,035862945	3,586294505
40,11	48,69	44,58	28,736	0,059067896	5,906789643
37,66	42,39	37,32	31,207	0,064147127	6,414712708
36,07	45,45	37,20	36,133	0,0742727	7,42726998
61,32	54,78	55,49	37,491	0,077064118	7,706411835
32,41	63,46	34,82	12,841	0,026395144	2,6395144
55,68	71,12	53,00	45,571	0,093672853	9,367285315
42,92	52,82	36,26	29,389	0,060410162	6,041016175
62,60	46,20	47,82	32,82	0,067462707	6,746270743
62,01	51,76	52,09	9,232	0,018976713	1,897671283
63,74	46,11	52,24	10,197	0,020960306	2,096030553
66,81	52,82	60,60	6,264	0,012875881	1,287588054

To Calculate Proportionality using the formula above, the value that has become C1 does not need to be calculated. Then, the largest probability value is taken. In the calculation label above, C1 is in row 4. Then the probability table of the other 24 data is as follows.

Based on the table above, the largest probability value is 0.0982656 (16th row), then the centroid value C2 is taken from that row, namely (C2 = 32.41, 63.46, 34.82).
the centroid value C2 is taken from that row, namely (C2 = 32.41, 63.46, 34.82).

1. Calculate the value of D (Xi)2
2. The calculation is performed on all rows except those that have been selected as centroids before (C1 and C2).
3. Compare the D(Xi)2 values of all centroids (C1 and C2) against each data.
4. Take the smallest value of the squared distance for each row of data. Then, sum up all the smallest values as the total divisor.
5. Find the probability value P(Xi) except for the data that has become the centroid before, namely row C1 and C2).

Based on these steps, the centroid value is obtained as follows:

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

	Indeks Komposit 2022	Indeks Komposit 2023	Indeks Komposit 2024
	59,32	53,13	56,51
	60,98	57,51	59,18
C1	69,83	45,37	53,50
	60,17	51,80	56,03
	62,31	54,12	58,05
C6	67,03	62,15	62,78
C5	55,71	50,99	56,51
	56,75	46,56	51,52
	51,12	44,40	55,74
	58,55	51,79	55,41
	41,18	48,34	42,29
C4	46,24	41,34	48,94
	58,43	57,60	58,47
	43,68	49,19	42,21
	40,11	48,69	44,58
	37,66	42,39	37,32
	36,07	45,45	37,20
	61,32	54,78	55,49
C2	32,41	63,46	34,82
C3	55,68	71,12	53,00
	42,92	52,82	36,26
	62,60	46,20	47,82
	62,01	51,76	52,09
	63,74	46,11	52,24
	66,81	52,82	60,60

Calculating Euclidean Distance for Adaptive K-Means

At this stage, the process of calculating the distance between each data and the cluster center (centroid) of each cluster that has been previously determined. This process uses the Euclidean Distance formula, which is a formula for measuring the straight distance between two points in a dimensionless space.

The Euclidean Distance formula between data points X=(x1,x2,...,xn)X = (x_1, x_2,...,x_n)X=(x1,x2,...,xn) and centroid C=(c1,c2,...,cn)C = (c_1, c_2,...,c_n)C=(c1,c2,...,cn) is:

$$d(X, C) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_n)^2}$$

This step is an important part of the Adaptive K-Means algorithm, as it determines the placement of each data into the closest cluster. By calculating the Euclidean distance, we can find out how close the data is to the cluster center. The smaller the distance, the more likely the data belongs to the cluster.

After all data has been calculated the distance to each centroid, then each data will be grouped into the cluster with the shortest distance. Next, the centroid of each cluster is updated based on the average of all data in the cluster. This process is done iteratively. The

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

distance calculation, clustering, and centroid update will continue to repeat until the centroid position no longer changes significantly (convergent). Using the Euclidean Distance formula above, the following results are obtained:

C1	C2	C3	C4	C5	C6	jarak
173,805	36,07008	18,68425	19,16762	4,19745	13,42373	C5
231,5091	38,01548	15,85748	24,1627	8,803203	8,43117	C6
0	45,57097	29,3887	24,35958	15,49468	19,37927	C1
137,2428	36,83152	20,0651	18,80936	4,559262	14,13302	C5
137,7152	39,00376	18,93215	22,47041	7,473747	10,44115	C5
298,7101	44,52414	17,4653	32,51249	17,0921	0	C6
234,0646	34,18478	20,43286	15,49466	0	17,0921	C5
170,4764	34,01694	24,63387	12,01302	6,75649	21,81184	C5
352,999	33,93455	27,24922	8,916261	8,066913	24,85473	C5
170,3775	35,26348	19,69063	17,39694	3,150207	15,28406	C5
818,4117	19,01216	29,05122	10,90193	20,49614	35,76137	C4
567,9937	29,66962	31,50868	0	15,49466	32,51249	C4
284,6445	35,64809	14,84264	22,44856	7,412876	10,64459	C5
686,8992	19,63336	27,23151	10,65342	18,76867	33,71187	C4
885,4426	19,30889	28,57555	10,52038	19,76786	35,17469	C4
1027,632	21,86013	37,3648	14,48157	27,70801	43,60534	C4
1123,443	18,53464	35,96491	16,07065	28,09256	43,49945	C4
162,9526	36,58241	17,46684	21,23294	6,845946	11,8417	C5
1709,044	0	30,50914	29,66962	34,18478	44,52414	C2
862,945	30,50914	0	31,50868	20,43286	17,4653	C3
762,4337	15,02749	27,89854	17,42821	24,02015	37,04347	C2
47,25182	37,12816	26,37671	17,10362	12,08013	22,31186	C5
100,5888	36,21389	20,39209	19,15975	7,731034	15,73235	C5
36,36828	39,82494	26,28895	18,43362	10,32392	19,47597	C5
71,66866	44,29359	22,73401	26,28968	11,98041	9,585569	C6

After determining the cluster for each data in iteration 1 based on the results of the smallest Euclidean calculation, the next step is to calculate the new centroid for each cluster. This process is done by:

1. Grouping the data into their respective clusters according to the results of the previous iteration.
2. Calculating the average value of each attribute (feature) of different data in one cluster.
3. Making the average value as the new centroid for the cluster.

Formula for calculating the new centroid for each cluster

$$C_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i$$

Where:

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Ck = new centroid for the kth cluster
Nk = number of data in the kth cluster
Xi = Ith data in the kth cluster

After obtaining the new centroid, the euclidean distance between each data and the new centroid is calculated again. The results of this calculation will form the 2nd iteration of the Adaptive K-Means process.

Calculating Dice Distance for adaptive K means

At this stage, the distance between the data and the centroid that has been determined previously using the Dice Distance method in the Adaptive K-Means algorithm is calculated. Dice Distance is a similarity measure that is often used for binary or categorical data. Unlike the Euclidean or Canberra Distance which focuses on numerical differences between features, Dice Distance emphasizes on attribute similarity between two entities.
The Dice Distance formula is expressed as:

$$\text{Dice Distance} = 1 - \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

a. A and BBB are two sets of features (usually in binary form),
b. |A∩B| is the number of attributes in common.

Based on the dice distance formula above, the first iteration results are obtained as follows:

c1	c2	c3	c4	c5	c6	jarak terpendek	cluster
170,8009	830,808	336,8171	310,1158	17,61859	140,8119	17,61859	C5
225,8308	851,7517	213,2663	478,9032	70,33711	58,12086	58,12086	C6
0	1727,747	863,4454	572,5591	231,061	289,4335	0	C1
134,7126	906,6495	393,4221	303,4599	20,56268	154,1838	20,56268	C5
133,1644	981,5515	332,9026	421,8233	53,4615	86,65894	53,4615	C5
289,4335	1200,588	209,3762	865,3997	252,7492	0	0	C6
231,061	698,378	405,221	182,8102	0	252,7492	0	C5
172,461	878,3888	604,6254	137,6581	20,76638	348,8655	20,76638	C5
350,7595	713,8981	735,0119	33,20994	64,49093	568,1889	33,20994	C4
168,4701	819,6451	381,9234	260,7741	8,721375	179,2448	8,721375	C4
829,6289	305,6345	729,3036	74,70017	218,0904	859,077	74,70017	C4
572,5591	681,0108	976,2838	0	182,8102	865,3997	0	C4
279,6733	711,3261	190,3536	412,9974	51,07659	94,74904	51,07659	C5
698,2013	330,903	625,0591	68,20742	147,84	713,204	68,20742	C4
894,3662	277,5304	745,7019	91,71961	248,6302	906,0763	91,71961	C4
1043,829	471,6056	1150,319	74,81899	399,7626	1253,271	74,81899	C4
1139,766	337,8845	1043,723	120,4437	416,4001	1237,654	120,4437	C4
160,9692	911,2678	298,9208	407,9684	45,8255	86,98351	45,8255	C5
1727,747	0	600,2844	681,0108	698,378	1200,588	0	C2
863,4454	600,2844	0	976,2838	405,221	209,3762	0	C3
779,6992	223,7673	497,9112	142,9124	166,9071	668,5763	142,9124	C3
52,93618	1209,547	668,8775	291,283	70,47598	273,9382	52,93618	C1

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

c1	c2	c3	c4	c5	c6	jarak terpendek	cluster
101,9986	1013,159	415,0106	357,1459	40,28864	133,2335	40,28864	C5
37,63443	1282,758	690,5229	328,9217	88,35021	268,0899	37,63443	C1
64,57089	1297,28	459,069	555,0982	126,7529	87,12991	64,57089	C1

The purpose of this process is to perform data clustering using the Adaptive K-Means method, by determining the best centroid based on the minimum (closest) distance value from each data to all centroids.

Application of Euclidean and Dice Distance to Adaptive Kmeans with whole data

After initial manual testing using sample data, the next step is to test the entire dataset using an automated approach. In this case, the test is carried out by utilizing the Python programming language which is run on the Google Colab platform. This platform was chosen because it supports large-scale data processing efficiently and allows direct integration with files from the user's local computer storage.

The developed program has several main features. First, the user can upload the dataset file directly from the computer's internal storage through the Google Colab interface. Second, parameters such as the number of clusters (k), the maximum number of iterations and the initial value of the first centroid (C1) are determined manually by the user. The next three centroid values (C2 and C3) are not randomly determined, but systematically calculated based on a probability formula that considers the distance between the data and the initial centroid. This approach is used to improve the initialization quality and stability of the clustering results. Furthermore, users can choose the distance method that will be used in the Adaptive K-Means algorithm, namely Euclidean Distance or Dice Distance. Each method has different characteristics. Euclidean Distance measures the absolute distance between data and centroid geometrically, while Dice Distance is more suitable for categorical or binary data, focusing on attribute similarity between data pairs. The choice of distance method allows for evaluation and comparison of the clustering results produced by each approach.

This approach allows for a more thorough and efficient examination of the entire Composite Index and Food Vulnerability data, and provides a realistic overview of the performance of the Adaptive K-Means algorithm with the Euclidean and Dice distance methods automatically. By comparing these two distance methods. A deeper understanding of the effect of selecting a distance function on clustering results and accuracy is gained in the context of clustering regions based on food security and vulnerability levels.

Euclidean Distance Calculation of Adaptive K-Means with Overall Data

Iteration Results and Visualization

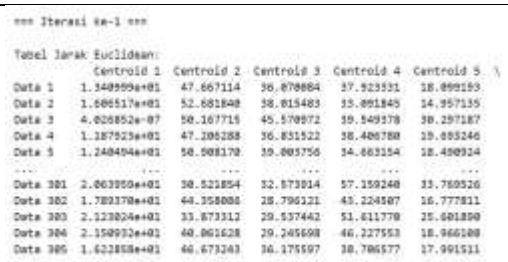
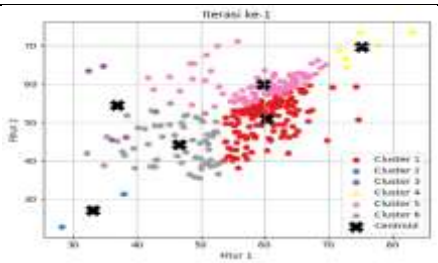
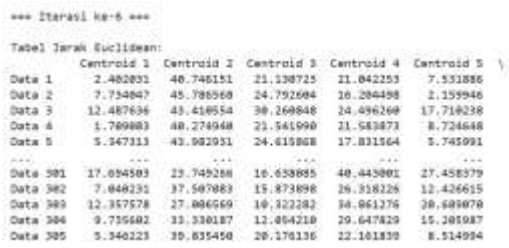
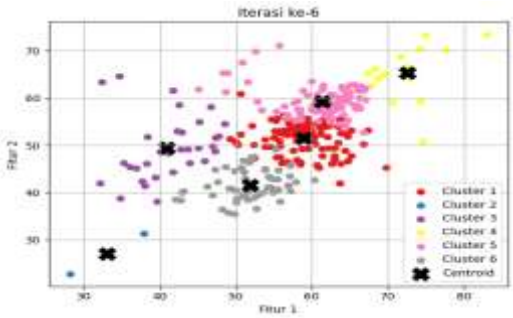
In the program, the number of clusters is 6 which represents the Composite Index of Food Security and Vulnerability Map. And the C1 value is determined randomly by the user. In this case the C1 value is taken from row 3 (69.83 45.37 53.50) and will give the amount of data included in each cluster at the last iteration. To get centroids 2-6 using the probability formula as follows.

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Perhitungan $D(X_i)^2 = ||X_i - C_j||^2$

Menghitung probabilitas $P(X_i) = \frac{D(X_i)^2}{\sum D(X_i)^2}$

After the centroi values 1-6 are obtained, the program will calculate with the Euclidean formula and provide visualization of each iteration and calculate the amount of data included in each cluster.

Iterasi	Tabel Iterasi	Visualisasi
1		
6		

Based on the number of members of each cluster (final result), the cluster results are obtained, namely:

Cluster 1: 101 data
Cluster 2: 2 data
Cluster 3: 31 data
Cluster 4: 15 data
Cluster 5: 98 data
Cluster 6: 58 data

Evaluation of the Application of Euclidean and Dice Distance to Kmeans Adaptive with overall data using Silhouette score.

In this subsection, we evaluate the effectiveness of the application of two types of distance methods, namely Euclidean Distance and Dice Distance, on the K-Means Adaptive algorithm using the whole data as the object of analysis. Adaptive K-Means is a variant of the conventional K-Means algorithm that adapts the centroid selection and cluster formation process based on the dynamics of data distribution.

The application of these two distance functions aims to assess the effect of different mathematical formulas on the quality of clustering results produced by the algorithm. Euclidean Distance is an absolute value-based distance measurement method between

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

dimensions that is commonly used in continuous data. In contrast, Dice Distance is more suitable for binary or transformed data, and emphasizes the similarity between features through the ratio of the intersection to the combined measure.

To evaluate the performance of both methods, Silhouette Score is used as a comparison metric. Silhouette Score measures how well data is clustered, by comparing the average distance between data in the same cluster to the distance to the nearest other cluster. The value of this score is in the range of -1 to 1, where a value close to 1 indicates that the data is in the correct cluster, while a value close to -1 indicates that the data fits better in another cluster.

In this experiment, the centroid starting points for each method were kept the same to make the comparison of clustering results fairer and more consistent. It is important to ensure that the difference in Silhouette scores truly reflects the effectiveness of the distance function used, rather than the variation in the initial centroid selection. The evaluation results show a significant difference between the two methods. Euclidean Distance, with its characteristic of being sensitive to absolute value differences, produces clusters that tend to be compact and radially arranged. This makes it suitable for data that is continuous and has a homogeneous distribution. Meanwhile, Dice Distance, which emphasizes more on the similarity of features in binary form, shows higher flexibility in handling data with uneven or sparse distribution patterns. The clusters produced by Dice Distance tend to be more varied but still reflect the latent structure of the binary data.

Overall, this evaluation confirms that the choice of distance function has a substantial impact on clustering performance, especially in the context of adaptation to data characteristics. Therefore, the selection of the distance method should consider the type of

Evaluation Results of Euclidean Distance Against K-Means Lagorithm

This sub-chapter presents the performance evaluation results of Adaptive K-Means algorithm using Euclidean Distance through Silhouette Score approach. Silhouette Score is an evaluation metric used to assess the quality of clustering by measuring how close a data point is to the cluster it is in compared to other nearby clusters. Mathematically, Silhouette Score for each data point is calculated based on the formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where:

1. a(i)a(i)a(i) is the average distance between point iii and all other points in the same cluster
2. b(i)b(i)b(i) is the average distance from point iii to points in other nearby clusters
3. The Silhouette Score value is in the range of -1 to 1, with the following interpretation:
4. A value close to 1 indicates that the data fits perfectly into its cluster.
5. A value close to 0 indicates that the data is at the boundary between two clusters
6. A value close to -1 indicates that the data may fit better in another cluster

Journal Of Data Science
Volume 3, Number 02, 2025, DOI 10.58471/Science.v3i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

```
Tabel Evaluasi Silhouette Score:
     Index  Cluster  Silhouette Score
0        0        0          0.401395
1        1        4          0.374875
2        2        0          0.261524
3        3        0          0.463069
4        4        0          0.059774
..     ...      ...               ...
300    300        5          0.552645
301    301        0          0.359863
302    302        5          0.287341
303    303        0          0.225492
304    304        0          0.274658

[305 rows x 3 columns]

Rata-rata Silhouette Score: 0.3082
Hasil evaluasi disimpan ke file: evaluasi_silhouette_euclidean.xlsx
```

Based on the figure above, the K-Means evaluation results using Euclidean distance show that the average Silhouette Score obtained is 0.3082. This value is in the low to medium range, which indicates that the cluster structure in the data is not fully optimized. Most of the Silhouette Score values are in the range of 0.2 to 0.4, indicating that some data has been clustered fairly well, but there is still overlap between clusters. In addition, there are very low Silhouette Score values (e.g. 0.059774), indicating that the points may not fit the cluster they are in, or are outliers. Overall, the model can still be improved by re-evaluating the number of clusters, feature selection, or distance measurement method used.

CONCLUSION

Based on the evaluation results of euclidean distance and dice distance. It is obtained that the euclidean distance evaluation result is 33.08% and the dice distance evaluation result is 30.82%. So it can be concluded that the euclidean distance is better used in calculating clustering in the data.

REFERENSI

[1.] K. P. Badan Ketahanan Pangan, “Indeks Ketahanan Pangan Indonesia 2018.” Badan Ketahanan Pangan, 2018.

[2.] Rousseeuw, P. J. (1987). Siluet: Bantuan grafis untuk interpretasi dan validasi analisis klaster. Matematika Komputasi dan Terapan, 20: 53-65.

[3.] Kepler, G., & Palomino, M. (2023). Gurun makanan dan pengelompokan k-means. *SIAM Undergraduate Research Online*.https://www.siam.org/media/bl2p3oyy/s150444.pdf

[4.] Zhou, Y. (2023). Perbandingan k-means clustering dengan hierarchical agglomerative clustering dalam analisis ketahanan pangan. JIDSS. https://www.idss.iocspublisher.org/index.php/jidss/article/download/290/161

[5.] Bora, D. J., & Gupta, A. K. (2014). Pengaruh ukuran jarak yang berbeda terhadap kinerja algoritma K-Means: Sebuah studi eksperimental di Matlab. *arXiv preprint*. https://arxiv.org/abs/1405.7471