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 Arccosine Distance in Adaptive K-Means Clustering Algorithm for Penguin Species Clustering

Herlina Br Nainggolan[1], Pandi Barita Nauli Simangungsong[2]
Faculty of computer sciene, University Katolik Santo Thomas, Medan, Indonesia

Article Info
Keywords:
Clustering,
Arccosine distance,
Euclidean distance

ABSTRACT
Clustering is an important method in unsupervised learning for grouping data based on similarity of characteristics. This study aims to cluster penguin species based on weight, height, and wing length attributes using the K-Means algorithm with two distance approaches: Euclidean and Arccosine. The dataset consists of 342 data points after cleaning. Evaluation results show that the Arccosine distance yields a clustering accuracy of 89.6%, higher than the Euclidean distance at 63.09%. This indicates that Arccosine is more optimal for classifying penguin species.

Corresponding Author:
Herlina Br Nainggolan
University Katolik Santo Thomas, Medan, Indonesia
herlinanainggolan32@gmail.com

INTRODUCTION

Data is a collection of information obtained through observation and research from various sources ('Monica rizkiana_tugas uas_Pbd osf', no date). In general, data is divided into two types, namely primary data and secondary data. The process of grouping a set of data into classes containing examples that are similar is called clustering (Harifi, Khalilian and Mohammadzadeh, 2023). Clustering is an analysis method in unsupervised learning that aims to group objects based on the statistical characteristics of each data (Dzakiansyah and Pramiyati, 2020). Cluster analysis is used in various fields, one of which is pattern recognition (Flasiński, 2016), data analysis (Harifi, et all, 2017). This technique is also used to organize data so that the information retrieval process becomes more efficient (Bhatia, 2004). This study uses the K-Means and Adaptive K-Means algorithms. K-Means is a distance-based clustering method that divides data into several clusters and can only be used on numerical attributes (Handayani, 2022). Meanwhile, Adaptive K-Means is a modified version of K-Means that selects K elements from the input data and calculates the distance between these elements and the formed clusters (Mughnyanti and Hafiz Nanda Ginting, 2023).

Classifying animal species is a major challenge in biology (Adrianto *et al.*, 2022), Classifying animal species is a major challenge in biology, especially when there are significant morphological similarities between species (Adji *et al.*, 2025). Clusters are the largest virial structures found in the universe (Nelson *et al.*, 2024). One clear example of this problem is the penguin, a flightless seabird from the Spheniscidae family, which is widely distributed in the southern hemisphere, including Antarctica, the southern coast of Africa, Australia, and New Zealand. (Gill, Frank, 2023). Adaptation to diverse environmental

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

conditions (Anggraeni *et al.*, 2023) has led to variations in body size, shape, and weight among different penguin species. However, convergent evolution has also resulted in physical similarities between species (Ramussen *et al.*, 1996), making identification difficult if relying solely on visual observation.

Generally, adult penguins are about 1.1 meters tall and weigh around 35 kilograms. In contrast, the smallest penguin, Eudyptula minor, is only 30–33 centimeters tall and weighs between 1.2 and 1.3 kilograms (2.6–2.9 pounds) (Grabski, Valerie (2009). These size differences are usually closely related to the habitats where they live—larger species tend to inhabit colder regions, while smaller ones are more commonly found in temperate or tropical climates. Although morphometric data such as height and weight can provide important information, the similarity in body shape among species remains a challenge in achieving accurate classification.

Previous research has generally relied on conventional methods such as visual observation and basic morphological analysis to distinguish penguin species (S. P. Collins *et al.*, 2021).These methods are subjective and ineffective, especially when applied on a large scale involving hundreds of samples from various habitats with high diversity in form and environment. Therefore, a more systematic, objective approach that can efficiently handle large amounts of data is needed.

In this context, machine learning-based methods offer a promising alternative solution. The ability of machine learning to explore and analyze data in depth enables this method to group penguin morphometric data more accurately. Unfortunately, the application of this technique for penguin species classification has not been extensively developed in previous studies.

Considering this, this study aims to utilize machine learning methods in the process of classifying penguin species based on morphometric data. Through this approach, it is hoped that a more accurate, efficient, and objective classification system will be obtained, while also considering biological diversity and the influence of the environment on the physical form of penguins.

METHODS

The model used in this study is the K-Means clustering method, because this method, whether using Euclidean distance or Arccosine, is able to group data based on similarities between data. This method is considered suitable for this study because it is effective in grouping penguin species based on their attributes.

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

1. Clustering
   Clustering is one of the unsupervised (unlabeled) techniques in data mining, where the characteristics of each cluster are not predetermined. Clusters are formed automatically based on the similarity of attributes between data in a group.
2. K-Means
   K-Means is a clustering method that aims to divide a number of N observations into K clusters. Each observation will be assigned to a cluster based on its proximity to the mean of that cluster.
3. Adaptive K-Means
   Adaptive K-Means is an extension of the K-Means algorithm with a more systematic and accurate approach to selecting centroid points. Unlike the conventional K-Means method, which sets centroids randomly, Adaptive K-Means determines the initial centroid point (c_1) randomly, but subsequent centroid points are calculated based on a specific formula to improve the accuracy of centroid position selection.

$$D(x)^2 = \| x - c \|^2 = (x_x - 1)^2 + (x_y - 2)^2$$

This observation can be described as follows:

1. Data Source
   The penguin species classification data was obtained from data.go.id (https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species?resource=download), which contains 345 pieces of penguin species classification data.
2. Data Cleaning
   The 345 data points were cleaned to create a valid dataset, where data cleaning was performed by removing NAN values.
3. Dataset
   After data cleaning, the dataset is ready for use in Euclidean distance and arcossine distance comparisons. The following dataset will be used:

Table 1. Dataset

N0	Culmen length mm	Culmen depth mm	Flipper legth mm	Body mass g
1	39.1	18.7	181	3750
2	39.5	17.4	186	3800
3	40.3	18	195	3250
4	36.7	19.3	193	3450
5	39.3	20.6	190	3650
6	38.9	17.8	181	3625
7	39.2	19.6	195	4675
8	34.1	18.1	193	3475
9	37.8	17.1	186	3300
.....	.....	.....	.....	.....

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

338	48.8	16.2	222	6000
339	47.2	13.7	214	4925
340	46.8	14.3	215	4850
341	50.4	15.7	222	5750
342	45.2	14.8	212	5200

RESULTS AND DISCUSSION

Euclidean distante calculation for k-means

In euclidean distance calculation, there are several steps that need to be taken, namely:

a. The initial step is to determine the centroid point, where C1 is located at the 20th data point.

$$C1 = 49.8, 17.3, 198, 3675$$

b. Perform calculations for the next centroid using the Adaptive K-Means method with the formula:

$$D(Xi)2 = ||Xi-Cj||2$$

Table 2. Determining te Centroid

No	$\|\|Xi-Cj\|\|^2$	$D(Xi)^2$
1	$\|\|39.1-49.8\|\|^2+\|\|18.7-17.3\|\|^2+\|\|181-198\|\|^2+\|\|3750-3675\|\|^2$	77.656
2	$\|\|41.1-49.8\|\|^2+\|\|17.6-17.3\|\|^2+\|\|182-198\|\|^2+\|\|3200-3675\|\|^2$	475.349
3	$\|\|34.3-49.8\|\|^2+\|\|18.4-17.3\|\|^2+\|\|184-198\|\|^2+\|\|3325-3675\|\|^2$	350.624
4	$\|\|36.7-49.8\|\|^2+\|\|19.3-17.3\|\|^2+\|\|193-198\|\|^2+\|\|3450-3675\|\|^2$	225.445
5	$\|\|36-49.8\|\|^2+\|\|18.5-17.3\|\|^2+\|\|186-198\|\|^2+\|\|3100-3675\|\|^2$	575.292
6	$\|\|42.3-49.8\|\|^2+\|\|21.2-17.3\|\|^2+\|\|191-198\|\|^2+\|\|4150-3675\|\|^2$	475.127
7	$\|\|39.7-49.8\|\|^2+\|\|18.4-17.3\|\|^2+\|\|190-198\|\|^2+\|\|3900-3675\|\|^2$	225.371
8	$\|\|34.1-49.8\|\|^2+\|\|18.1-17.3\|\|^2+\|\|193-198\|\|^2+\|\|3475-3675\|\|^2$	200.679
9	$\|\|37.8-49.8\|\|^2+\|\|17.1-17.3\|\|^2+\|\|186-198\|\|^2+\|\|3300-3675\|\|^2$	375.384
10	$\|\|37.8-49.8\|\|^2+\|\|20-17.3\|\|^2+\|\|203-198\|\|^2+\|\|4725-3675\|\|^2$	1050.084
11	$\|\|48.8-49.8\|\|^2+\|\|16.2-17.3\|\|^2+\|\|222-198\|\|^2+\|\|6000-3675\|\|^2$	2325.124
12	$\|\|41-49.8\|\|^2+\|\|20-17.3\|\|^2+\|\|190-198\|\|^2+\|\|4250-3675\|\|^2$	575.129
13	$\|\|40.2-49.8\|\|^2+\|\|17-17.3\|\|^2+\|\|176-198\|\|^2+\|\|3450-3675\|\|^2$	226.277
14	$\|\|50.4-49.8\|\|^2+\|\|15.7-17.3\|\|^2+\|\|222-198\|\|^2+\|\|5750-3675\|\|^2$	2075.139
15	$\|\|39-49.8\|\|^2+\|\|17.1-17.3\|\|^2+\|\|191-198\|\|^2+\|\|3050-3675\|\|^2$	625.063

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

No	$\|\|Xi-Cj\|\|^2$	$D(Xi)^2$
16	$\|\|49.9-49.8\|\|^2+\|\|16.1-17.3\|\|^2+\|\|213-198\|\|^2+\|\|5400-3675\|\|^2$	1725.066
17	$\|\|41.5-49.8\|\|^2+\|\|18.3-17.3\|\|^2+\|\|195-198\|\|^2+\|\|4300-3675\|\|^2$	625.063
18	$\|\|41.1-49.8\|\|^2+\|\|17.5-17.3\|\|^2+\|\|191-198\|\|^2+\|\|3175-198\|\|^2$	500.125
19	$\|\|40.6-49.8\|\|^2+\|\|17.2-17.3\|\|^2+\|\|187-198\|\|^2+\|\|3475-3675\|\|^2$	200.513
20	$\|\|49.8-49.8\|\|^2+\|\|17.3-17.3\|\|^2+\|\|198-198\|\|^2+\|\|3675-3675\|\|^2$	0.000
21	$\|\|45.6-49.8\|\|^2+\|\|19.4-17.3\|\|^2+\|\|194-198\|\|^2+\|\|3525-3675\|\|^2$	150.127
22	$\|\|48.7-49.8\|\|^2+\|\|15.7-17.3\|\|^2+\|\|208-198\|\|^2+\|\|5350-3675\|\|^2$	1675.031
23	$\|\|50.5-49.8\|\|^2+\|\|15.9-17.3\|\|^2+\|\|225-198\|\|^2+\|\|5400-3675\|\|^2$	1725.212
24	$\|\|55.1-49.8\|\|^2+\|\|16-17.3\|\|^2+\|\|230-198\|\|^2+\|\|5850-3675\|\|^2$	2175.242
25	$\|\|55.9-49.8\|\|^2+\|\|17-17.3\|\|^2+\|\|228-198\|\|^2+\|\|5600-3675\|\|^2$	1925.243
	Total	20559.437

Calculating Probability

$$P(Xi) = D(Xi)^2 / \text{total } D(Xi)^2$$

Table 3. Calculating Probability

No	$D(Xi)^2$	P(Xi)
1	77.656/20559.437	0.004
2	475.349/20559.437	0.023
3	350.624/20559.437	0.017
4	225.445/20559.437	0.011
5	575.292/20559.437	0.028
6	475.127/20559.437	0.023
7	225.371/20559.437	0.011
8	200.679/20559.437	0.010
9	375.384/20559.437	0.018
10	1050.084/20559.437	0.051
11	2325.124/20559.437	0.113
12	575.129/20559.437	0.028
13	226.277/20559.437	0.011
14	2075.139/20559.437	0.101
15	625.063/20559.437	0.030
16	1725.066/20559.437	0.084
17	625.063/20559.437	0.030
18	500.125/20559.437	0.024

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

No	$D(Xi)^2$	P(Xi)
19	200.513/20559.437	0.010
20	0.000/20559.437	0.000
21	150.127/20559.437	0.007
22	1675.031/20559.437	0.081
23	1725.212/20559.437	0.084
24	2175.242/20559.437	0.106
25	1925.243/20559.437	0.093

From the probability results, select the highest value to be used as the next centroid. Perform calculations to determine the centroid and probability to determine the next centroid. In this observation, there are only two centroid points, which can be described as follows:

C1 = 49.8, 17.3, 198, 3675

C2 = 48.8, 16.2, 222, 6000

Determining Clustering using Euclidean Distance

After all the sequences for finding the centroid points have been performed, the next step is to determine the clustering using Euclidean distance. The following are the calculations from data-1 to c1, c2, c3, c4, and c5.

Table 4. Determing Cluster Manually

C1	$\sqrt{(39{,}1-49{,}8)^2+(18{,}7-17{,}3)^2+(181-198)^2+(3750-3675)^2}$	77,656
C2	$\sqrt{(39{,}1-48{,}8)^2+(18{,}7-16{,}2)^2+(181-222)^2+(3750-6000)^2}$	2250,396

The calculation was performed on 25 data samples, and teh results show the distance of data 1 from C1, C2, C3, C4, and C5, as well as the cluster division of each data.

Table 5. Cluster

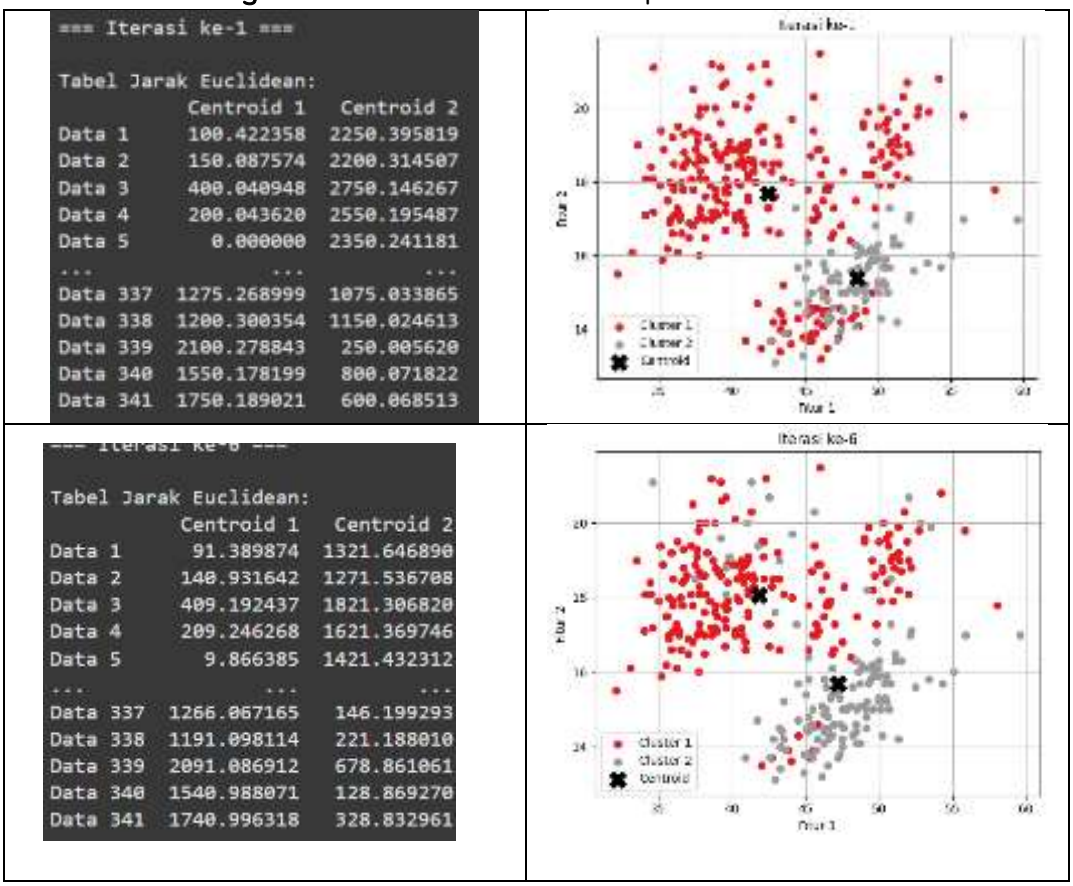
DATA	c1	c2	J.terkecil
1	77,656	2250,396	C1
2	475,349	2800,297	C1
3	350,624	2675,31	C1
4	225,445	2550,195	C1
5	575,292	2900,253	C1
6	475,127	1850,278	C1
7	225,371	2100,265	C1
8	200,679	2525,21	C1
9	375,384	2700,263	C1
10	1050,084	1275,195	C1
11	2325,124	0	C2
12	575,129	1750,314	C1
13	226,277	2550,429	C1
14	2075,139	250,006	C2
15	625,133	2950,179	C1
16	1725,066	600,069	C2
17	625,063	1700,231	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

DATA	c1	c2	J.terkecil
18	500,125	2825,181	C1
19	200,513	2525,256	C1
20	0	2325,124	C1
21	150,127	2475,163	C1
22	1675,031	650,151	C2
23	1725,212	600,01	C2
24	2175,242	150,345	C2
25	1925,243	400,109	C2

The purpose of this process is to perform data clustering using Adaptive K-Means method, by determining the best centroid based on the minimum (closest) distance value from each data to all centroids. To simplify the calculation od Euclidean distance using big data, Google Colab can be used to generate a scatter plot output in the form of data distribution options.

Figure 1. Euclidean Distance Output



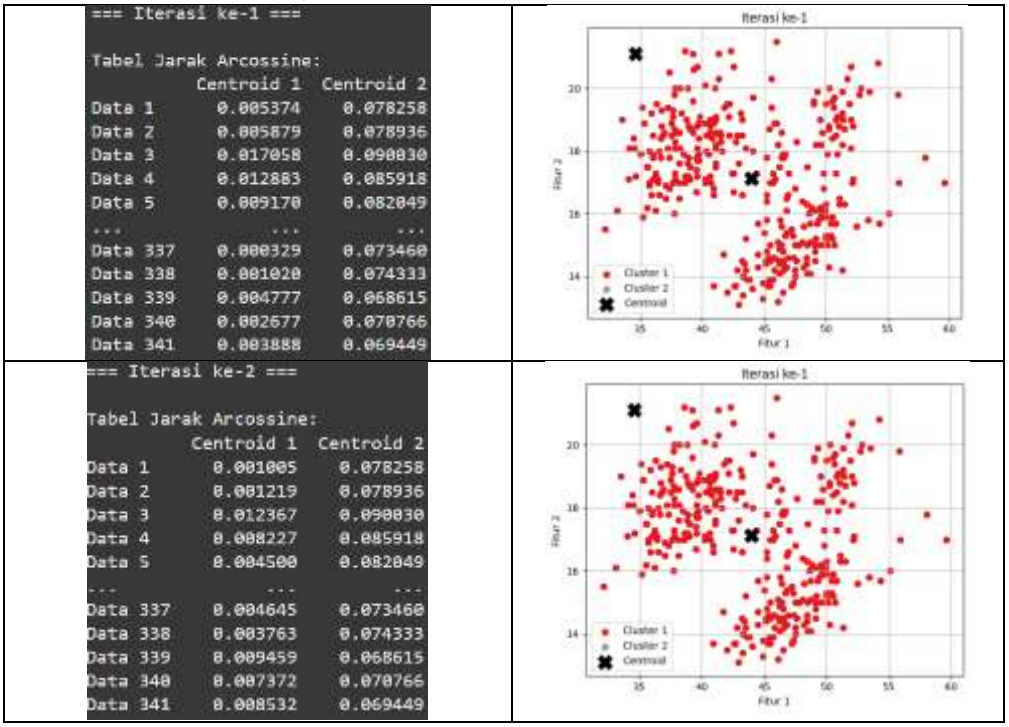
The output obtained from Google Colab is 7 iterations. The iteration stops when the cluster point no longer changes.

Testing with Arcossine Distante

To shorten the testing time, Google Colab tools are used. The following are the iteration results and scatter plot distribution points:

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

Figure 2. Arcossine Distance Output



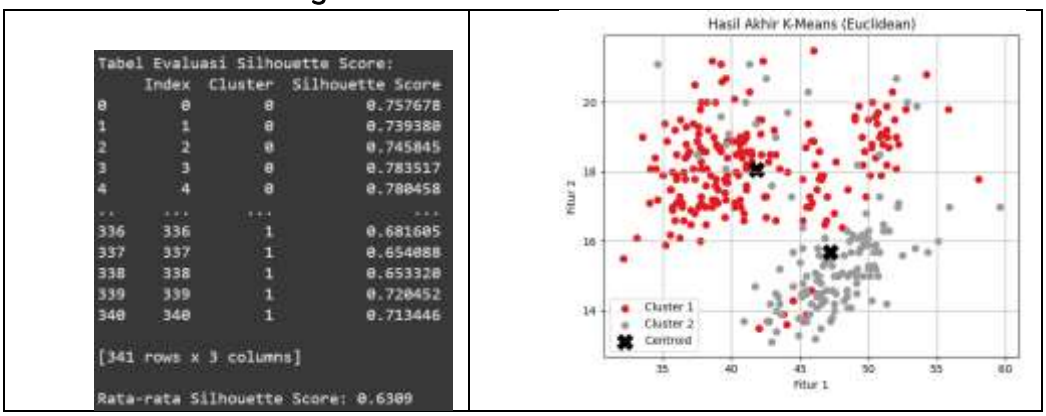
Evalution of Euclidean Distance and Arcossine Distance

After conducting a series of k-means tests on Euclidean distance and arcossine distance, an evaluation was carried out to see how accurately the data was clustered. The measurement used to assess the evaluation was -1<0<1.

a. Evaluation of Euclidean distance

The evaluation was carried out using the silhouette score. The following is the output.

Figure 3. Evaluation of Euclidean Distance



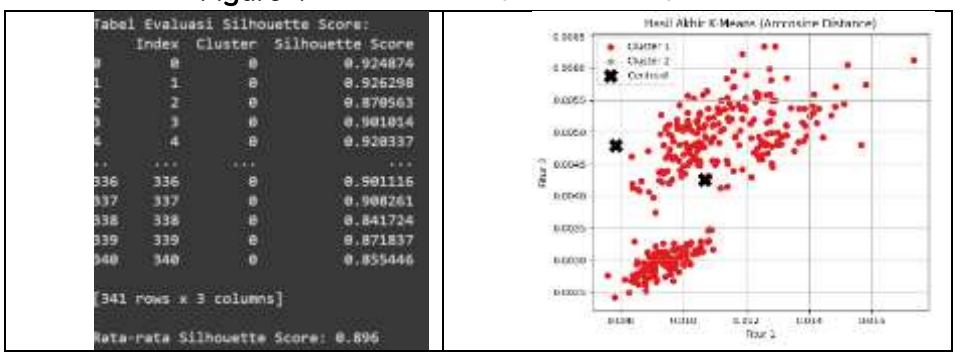
This, from the evaluation of the Euclidean distance, the average obtained was 63.9%.

b. Arcossine distance evaluation

The following are the results of the silhouette score evaluation calculation:

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

Figure 4. Arcossine distance evaluation



Thus, from the evaluation results of the arcossine distance, the average obtained was 89.6%.

CONCLUSION

The data used is the penguin species classification data, which consists of 345 initial data points reduced to 342 data points after data cleaning. After conducting Euclidean distance and arcossine distance tests, the evaluation results of both distances were obtained using the same dataset, namely the penguin species classification data consisting of 342 data points. The Euclidean distance evaluation was 0.6309 or 63.9%, while the arcossine distance was 0.896 or 89.6%. From the percentage, it can be seen that the data is in the correct cluster, as tested by the arcossine distance evaluation. As a recommendation for further research, it is suggested to examine the use of other distance measurement methods, such as Manhattan distance and Mahalanobis distance, to evaluate the potential for improving accuracy in the classification process. In addition, the use of larger and more diverse datasets, as well as the application of cross-validation techniques, is expected to produce more accurate, stable, and better generalized model evaluations.

REFERENCE

Adji, D.R. *et al.* (2025) 'Metode dan Algoritma Dalam Data Clustering: Systematic Literature Review', *Science Technology and Management Journal*, 5(1), pp. 9–15. Available at: https://doi.org/10.53416/stmj.v5i1.326.

Adrianto, H. *et al.* (2022) 'Pembekalan Klasifikasi Baru Makhluk Hidup Hewan Kepada Guru-Guru Biologi', *Sebatik*, 26(2), pp. 638–643. Available at: https://doi.org/10.46984/sebatik.v26i2.2152.

Anggraeni, R. *et al.* (2023) 'Perilaku Makan , Adaptasi Dan Menghindari Predator Pada Hewan', *Jurnal Lingkungan*, (July), pp. 1–21. Available at: https://osf.io/preprints/z572m/.

Bhatia, S.K. (2004) 'Pengelompokan K-Means Adaptif'.

Dzakiansyah, F. and Pramiyati, T. (2020) 'Perancangan Sistem Informasi E-Learning Pembelajaran Bahasa Inggris Berbasis Web', *Prosiding Seinasi-Kesi*, 14, pp. 143–148. Available at: https://conference.upnvj.ac.id/index.php/seinasikesi/article/view/811.

Flasiński M (2016) Pattern recognition and cluster analysis. Introduction to Artifcial

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

Intelligence. Springer, Cham. https://doi.org/10. 1007/978-3-319-40022-8_10

Gill, Frank; Donsker, David; Rasmussen, Pamela, ed. (2023). "Kagu, Sunbittern, tropicbirds, loons, penguins". World Bird List Version 13.1. International Ornithologists' Union. Diakses tanggal 15 juli 2025. https://id.wikipedia.org/wiki/Penguin

Grabski, Valerie (2009). "Little Penguin – Penguin Project". Penguin Sentinels/University of Washington. Diarsipkan dari asli tanggal 16 December 2011. Diakses tanggal 15 juli 2025. https://id.wikipedia.org/wiki/Penguin

Handayani, F. (2022) 'Aplikasi Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokan Mahasiswa Berdasarkan Gaya Belajar', *Jurnal Teknologi dan Informasi*, 12(1), pp. 46–63. Available at: https://doi.org/10.34010/jati.v12i1.6733.

Harifi, S., Khalilian, M. and Mohammadzadeh, J. (2023) 'Swarm based automatic clustering using nature inspired Emperor Penguins Colony algorithm', *Evolving Systems*, 14(6), pp. 1083–1099. Available at: https://doi.org/10.1007/s12530-023-09507-y.

Harif S, Byagowi E, Khalilian M (2017) Comparative study of apache spark MLlib clustering algorithms. In: Data mining and big data: second international conference, DMBD 2017, Fukuoka, Japan, July 27–August 1, 2017, Proceedings 2. Springer International Publishing, pp 61–73

'Monica rizkiana_tugas uas_Pbd osf' (no date).

Mughnyanti, M. and Hafiz Nanda Ginting, S. (2023) 'Data Mining Manhattan Distance dan Euclidean Distance Pada Algoritma X-Means Dalam Klasifikasi Minat dan Bakat Siswa', *Remik*, 7(1), pp. 835–842. Available at: https://doi.org/10.33395/remik.v7i1.12162.

Nelson, D. *et al.* (2024) 'Introducing the TNG-Cluster simulation: Overview and the physical properties of the gaseous intracluster medium', *Astronomy and Astrophysics*, 686, pp. 1–25. Available at: https://doi.org/10.1051/0004-6361/202348608.

Rasmussen, L.E.L., Lee, T.D., Roelofs, W.L., Zhang, A., Doyle Davies Jr, G. (1996). Insect pheromone in elephants. Nature. 379: 684

S. P. Collins *et al.*, "No Title 済無No Title No Title No Title," pp. 12–64, 2021.