Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 1, November 2023, pp. 103∼112
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i1.3269

Application of Numerical Measure Variations in K-Means Clustering for Grouping Data

Relita Buaton[1], **Solikhun**[2]
[1]Sekolah Tinggi Manajemen Informatika dan Komputer Kaputama, Binjai, Indonesia
[2]Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematang Siantar, Indonesia

Article Info

Article history:

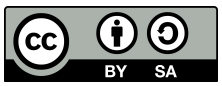
Received August 08, 2023
Revised October 08, 2023
Accepted November 05, 2023

Keywords:

Data Grouping
Distance Calculation
K-Means Clustering
Numerical Measures

ABSTRACT

The K-Means Clustering algorithm is commonly used by researchers in grouping data. The main problem in this study was that it has yet to be discovered how optimal the grouping with variations in distance calculations is in K-Means Clustering. The purpose of this research was to compare distance calculation methods with K-Means such as Euclidean Distance, Canberra Distance, Chebychev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity, and Manhattan Distance to find out how optimal the distance calculation is in the K-Means method. The best distance calculation was determined from the smallest Davies Bouldin Index value. This research aimed to find optimal clusters using the K-Means Clustering algorithm with seven distance calculations based on types of numerical measures. This research method compared distance calculation methods in the K-Means algorithm, such as Euclidean Distance, Canberra Distance, Chebychev Distance, Cosine Smilirity, Dynamic Time Warping Distance, Jaccard Smilirity and Manhattan Distance to find out how optimal the distance calculation is in the K-Means method. Determining the best distance calculation can be seen from the smallest Davies Bouldin Index value. The data used in this study was on cosmetic sales at Devi Cosmetics, consisting of cosmetics sales from January to April 2022 with 56 product items. The result of this study was a comparison of numerical measures in the K-Means Clustering algorithm. The optimal cluster was calculating the Euclidean distance with a total of 9 clusters with a DBI value of 0.224. In comparison, the best average DBI value was the calculation of the Euclidean Distance with an average DBI value of 0.265.



Corresponding Author:

Relita Button,
STMIK Kaputama, Binjai, Sumatera Utara,
Email: bbcbuaton@gmail.com

How to Cite:
R. Buaton and S. Solikhun, "Application of Numerical Measure Variations in K-Means Clustering for Grouping Data", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 1, pp. 103-112, Nov, 2023.

1. INTRODUCTION

Data Mining is solving problems by analyzing the data contained in the database. The data is stored electronically, and the data search is done automatically, like on a computer [1–3]. The K-Means method is a data mining method widely used in grouping data. K-Means has drawbacks, namely needing to determine the number of clusters [4–6]. Research [7] uses the K-Means algorithm, which is equipped with a web-based application program. The research results show that the k-means algorithm can produce the selection and distribution of superior classes for prospective new students according to the student's ability scores. The application of excellent classes has a positive impact on improving education.

Research [8] uses the K-Means algorithm method, processed with the rapid miner application, to group the best universities from several existing universities. The test results show that the best cluster in cluster 2 is Harvard University, with 667 universities. In cluster 1, the University of Haifa has 667 universities; in cluster 0, National Chung Cheng University has 666 universities. Research [9] aims to determine three cluster groups with product similarities to be used as a recommendation for company management in planning inventory. In this study, 3 clusters were chosen, with Cluster 1 being the best-selling product, Cluster 2 being the best, and Cluster 3 being the least-selling product. Cluster evaluation with DBI obtained quite good results with a value of 0.431. Measurements of accuracy, recall, and precision from Microsoft Excel calculations received discounts of 62%, 67%, and 59%, respectively. For analyses using Rapidminer, the accuracy value was 64%, recall was 81%, and precision was 88%. The clustering comparison results prove that Rapidminer calculations get higher accuracy, recall, and precision values.

The dataset used in the research [10] is the basic data of primary and secondary education in Tegal City. The research results show that the three methods compared have a good level of accuracy, namely 84.47% for Euclidean distance, 83.85% for Manhattan distance, and 83.85% for Minkowski distance. This research informs us that there are still six schools where the availability of teachers is still deficient (in the HIGH disparity label category) and need to receive more attention, namely SMP Atmaja Wacana, SMKN 3 Tegal, SMAS Muhammadiyah, SMAS Pancasakti Tegal, SMKS Muhammadiyah 1 Kota Tegal, and IC Bias Assalam Middle School. Research [11] uses data mining techniques with the K-Means clustering method to cluster patient medical record data. This research produced a 4-cluster column of sub-district, disease diagnosis, age, and gender. This grouping of patient medical record data creates new information regarding the grouping pattern of disease spread in each sub-district based on 534 patient medical record data from Anwar Medika Hospital with a completion time of 0.06 seconds. In Research [12], the analysis used is data on social assistance recipients that have yet to be grouped. Based on the results of grouping social assistance recipients using the K-Means method, of the 257 data, 196 data are included in cluster 1 with the status of receiving social assistance on target and 61 data in cluster 2 with the group of aid recipients. Social media is not on target. From the results of data analysis, a conclusion can be drawn that the people who receive assistance are right on target because the majority of aid recipients are received by people who need help from the government, where the recipients of assistance work as laborers, have no assets and have an income below Rp. 500,000.

The primary reference in this research is Research [13]. This research optimized the number of clusters needed to ensure policies that could be taken regarding grouping results, including ensuring regional groups with ODP, PDP, and Positive COVID-19 status in Riau province. This research compares two distance measurements, Euclidean and Manhattan, to find the best grouping by looking for DBI values for the two distance measurements by examining COVID-19 distribution data for the Riau region. The research results show that the lowest DBI values are at k=8 for Euclidean and k=7 for Manhattan, with values of 0.394 and 0.434, respectively. In addition, DBI works better on Euclidean than Manhattan because it has lower DBI values on all k tests. The main problem in this study is that it is not yet known which distance calculation is the most optimal in the K-Means method from the existing distance calculations, namely Euclidean Distance, Canberra Distance, Chebychev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity and Manhattan Distance. The purpose of this research is to compare distance calculation methods with K-Means such as Euclidean Distance, Canberra Distance, Chebychev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity, and Manhattan Distance to find out how optimal the distance calculation is in the K-Means method. The determination of the best distance calculation is seen from the smallest Davies Bouldin Index value.

The novelty of this research is to carry out clustering optimization on the K-means algorithm by comparing nine distance calculation methods in the K-means algorithm with several k tests, determining the optimal cluster by looking for the lowest DBI value. In previous research, we optimized the number of clusters using only two distance calculations, namely Euclidean and Manhattan distances. Hence, it needs to be improved again by adding several existing distance calculations to optimize the grouping results. The purpose of this research is to compare distance calculation methods with K-Means such as Euclidean Distance, Canberra Distance, Chebychev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity, and Manhattan Distance to find out how optimal the distance calculation is in the K-Means method. This research implies that the results can be applied in the optimal data grouping process to support appropriate decision-making based on the data resulting from the grouping.

2. RESEARCH METHOD

2.1. Research Framework

Arrange a research framework consisting of steps or stages to achieve the research objectives. The following is the research framework that the authors compiled. The design of the research stages for each step is explained in Figure 1. The first step involves collecting data for the research. Data relevant to the research topic should be collected according to the research objectives. Data can be in text, images, numbers, or other data types. Data normalization is a step that involves adjusting the data to be similar or conform to a specific scale. This is especially important if distance metrics such as the Euclidean Distance, Canberra Distance, Chebyshev Distance, and Manhattan Distance are used, as distances can be significantly affected by the scale of the data. Distance Calculation with Various Metrics: In this step, the distance between data pairs is calculated using various distance metrics mentioned, such as Euclidean Distance, Canberra Distance, Chebyshev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity, and Manhattan Distance. This provides insight into how similar or different the data pairs are in various aspects.

The clustering result step involves grouping or clustering data based on previously calculated distance metrics. Similar or more closely spaced data is combined into clusters or groups. Evaluation Dunn's Index or Davies-Bouldin Index (DBI) is an evaluation metric used to measure the quality of the resulting clusters. This helps assess how well clustering works and whether the resulting clusters are mutually exclusive and compact. The last step is the conclusion, where findings are summarized, the clustering results are interpreted, the DBI is evaluated, and whether the research objectives were achieved is concluded. The implications of the findings in the context of the problem can also be discussed.

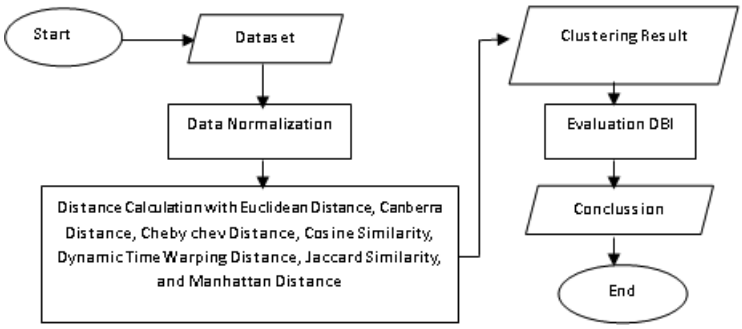


Figure 1. Research stages

2.2. Data Normalization

The research data utilizes a comprehensive dataset detailing the sales of 55 distinct cosmetic products at Devi Cosmetics from January to April 2022. The normalization process for this research data is conducted meticulously, employing the specific formula presented in Equation (1). Where x' = normalization result, x = data to be normalized, a = smallest data from the dataset, and b = largest data from the dataset.

$$x' = \frac{(x-a)}{(b-a)} \tag{1}$$

2.3. Distance Calculation Type

The formula for the seven distance calculations is as follows:

1. Euclidean Distance

Euclidean Distance [14] is a metric or distance measure used in Euclidean geometry to measure the distance between two points in dimensional space. This metric is the length of the straight line connecting the two points. In two-dimensional space, the Euclidean Distance between two points $(x1, y1)$ and $(x2, y2)$ can be calculated using the following formula as seen in Equation (2). D_{ij} = object distance between data values and cluster centre values, m = several data dimensions, X_{ij} = data values from the k-th dimension, and X_{jk} = cluster centre values from the k-th dimension [15].

$$d_{ij} = \sqrt{\Sigma_{k=i}^{m} x_{ij} - c_{ij}^{\ 2}} \tag{2}$$

2. Canberra Distance

Canberra Distance [16] is a metric that measures the difference between two vectors or points in a multidimensional space. This metric is often used in data analysis, especially in cases where the data has high dimensions and contains various attributes. The Canberra Distance emerges as an alternative to other distance metrics, such as the Euclidean Distance or the Manhattan Distance, as it accommodates differences in the scale and magnitude of attribute values. The formula for Canberra Distance is as follows in Equation (3). D_{ij} = difference level, n = the number of vectors, X_{ik} = input image vector, and X_{jk} = comparison image vector [17].

$$d_i j = \sqrt{\Sigma_{k=1}^{n} \frac{|X_{ik} - X_{jk}|}{|X_{ik}| - |X_{jk}|}} \tag{3}$$

3. Chebychev Distance

Chebyshev Distance [18], also known as Supremum Distance or Infinity Norm, is a metric used to measure the maximum distance between two points in a multidimensional space. This metric counts the most significant difference between the corresponding components of two vectors or points. The formula for Chebyshev Distance is as follows in Equation (4). This formula is used to calculate the distance or difference between elements in two rows (indices i and k) of a matrix X. This distance is calculated by finding the maximum value (largest value) of the difference in values between elements in the same column (index j) in both rows [19].

$$d_{ij} = max_k \, |X_{ij} - X_i k| \tag{4}$$

4. Cosine Similarity

Cosine Similarity [20] is a metric used to measure the degree of similarity between two vectors in a multidimensional space, especially in data analysis and text processing. This metric counts the angle between two vectors, not their geometric distance. Cosine Similarity ranges from -1 to 1, with a higher value indicating a more significant similarity between the vectors. The formula for Cosine Similarity is as follows, as seen in Equation (5). A = is the weight of each feature in vector A, and B = is the weight of each feature in vector B [21].

$$Cosine\ Similarity = \frac{A.B}{||A||, ||B||} \tag{5}$$

5. Dynamic Time Warping Distance

Dynamic Time Warping (DTW) Distance [22] is a method used to measure the similarity between two temporal data sequences or time series that have different time lengths or time distortions. DTW is an algorithm used to compare two timelines, which may have different rates or patterns of change, so they sometimes differ from the linear comparison method. The formula for Dynamic Time Warping is as follows in Equation (6). m is the number of variables A and B, A1 is the ith data matrix A, and B1 is the ith data matrix B [23].

$$D_{DTW} = (A, B) = \Sigma_{i=m}^{m} D_{DTW}(A_i - B_i) \tag{6}$$

6. Jaccard Similarity

Jaccard Similarity [24] is a metric used to measure the degree of similarity between two sets. This metric measures how many elements are the same or similar between two groups in proportion to the total components in those sets. Jaccard Similarity is particularly useful in data analysis involving groups, such as text processing, cluster analysis, and content-based recommendations.

The Jaccard Similarity formula between two groups, A and B, is defined in Equation (7). x = the value of the key and y = the value of the document [25].

$$J(x,y) = \frac{\Sigma_i^p x_i y_i}{\Sigma_{j=i}^p x_i^2 + \Sigma_{j=i}^p y_i^2 - \Sigma_{j=i}^p x_i y_i} \tag{7}$$

7. Manhattan Distance

Manhattan Distance [26], or City Block Distance or L1 Distance, is a metric that measures the distance between two points in a multidimensional space. The name "Manhattan" refers to the layout of the streets in the city of Manhattan, New York, where the distance traveled to move from one point to another must follow a path parallel to the coordinate axes. The Manhattan Distance formula is as seen in Equation (8). This formula calculates the distance or difference between two vectors, namely vector i and vector j. This distance is calculated by adding up the difference (absolute value) between the corresponding components of the two vectors [27].

$$d(i,j) = |X_{i1} - X_{j1}| + |X_{i2} - X_{j2}| + \cdots + |X_{jp} - X_{jp}| \tag{8}$$

3. RESULT AND ANALYSIS

This research is a novelty in the form of an optimal cluster from comparing nine distance calculations in the K-means algorithm using DBI values. Each distance calculation is tested with the k test value, $k = 2$ to $k = 9$. The results of this optimal cluster can be used to group data to make an optimal decision in grouping cosmetic sales using seven distance calculations: Euclidean Distance, Canberra Distance, Chebychev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity, and Manhattan Distance. The author makes a comparison of the seven existing distance calculations. Evaluation of 7 distance calculations using DBI evaluation. The optimal cluster is the one with the smallest DBI value. Before the data is processed, normalization is carried out firstthe results of the normalization of Devi cosmetics sales data from January to April 2022. The results of the data normalization are in Table 1.

Table 1. Normalization of Cosmetic Sales Data

No	January 2022	February 2022	March 2022	April 2022	No	January 2022	February 2022	March 2022	April 2022
1	0.1111	0.1944	0.3056	0.0278	29	0.3333	0.0833	0.1667	0.1111
2	0.3611	0.3333	0.4167	0.2222	30	0.2222	0.3056	0.4444	0.2222
3	0.0556	0.1944	0.2778	0.0833	31	0.2778	0.3611	0.2500	0.1389
4	0.2778	0.3889	0.1667	0.0556	32	0.0556	0.2222	0.2778	0.1111
5	0.9444	0.1667	0.1667	0.3056	33	0.1389	0.1111	0.1111	0.0556
6	0.7222	0.1944	0.2222	0.0556	34	0.0556	0.0278	0.0278	0.0278
7	0.2222	0.0833	0.3333	0.3333	35	0.5556	0.5000	0.6944	0.2500
8	0.0556	0.2500	0.0833	0.2778	36	0.2500	0.1111	0.1111	0.0833
9	1.0000	0.3056	0.2500	0.3611	37	0.1389	0.3056	0.2222	0.1111
10	0.0000	0.1111	0.0833	0.0833	38	0.2500	0.1111	0.1389	0.2222
11	0.1944	0.1111	0.3333	0.0000	39	0.1111	0.1944	0.1389	0.1111
12	0.1667	0.1389	0.3333	0.0278	40	0.1111	0.1389	0.2778	0.0833
13	0.1944	0.0556	0.0000	0.0833	41	0.1389	0.2222	0.0278	0.0833
14	0.0556	0.2500	0.2222	0.1944	42	0.3611	0.4167	0.2778	0.1111
15	0.1944	0.0278	0.1667	0.1667	43	0.0556	0.0000	0.0556	0.0556
16	0.0000	0.1389	0.3056	0.0556	44	0.4444	0.2222	0.4167	0.1667
17	0.1111	0.1944	0.0000	0.0278	45	0.1667	0.0833	0.2500	0.0556
18	0.0278	0.0278	0.0556	0.0278	46	0.4167	0.3333	0.1111	0.3333
19	0.0278	0.1389	0.0556	0.0278	47	0.0833	0.0833	0.2778	0.0556
20	0.0833	0.2222	0.1111	0.0000	48	0.3056	0.5000	0.1667	0.3333
21	0.5000	0.4167	0.5556	0.1944	49	0.3333	0.4444	0.4167	0.1667
22	0.5833	0.6389	0.6389	0.2222	50	0.2778	0.2778	0.1111	0.1389
23	0.0000	0.1389	0.3333	0.5833	51	0.1667	0.3333	0.4444	0.1667
24	0.1111	0.0000	0.1944	0.0278	52	0.0278	0.1667	0.0000	0.0000
25	0.0278	0.2500	0.0556	0.0833	53	0.3056	0.3056	0.0556	0.3889

No	January 2022	February 2022	March 2022	April 2022	No	January 2022	February 2022	March 2022	April 2022
26	0.6389	0.4444	0.4444	0.2778	54	0.2500	0.0833	0.1389	0.0556
27	0.3056	0.1111	0.0556	0.0278	55	0.3056	0.6111	0.0556	0.3056
28	0.1111	0.1944	0.1111	0.0556	56	0.1944	0.1667	0.0833	0.0833

The DBI value for grouping with the K-Means Clustering algorithm uses k=2 to k=9 with the Euclidean Distance calculation in Table 2. The DBI value for grouping with the K-Means Clustering algorithm uses k=2 to k=9 with the calculation of the Canberra Distance in Table 3. The DBI values for grouping with the K-Means Clustering algorithm use k=2 to k=9 with the calculation of the Chebychev Distance in Table 4.

Table 2. DBI Value from Euclidean Distance Calculation

Number of Clusters(k)	DBI Value
k=2	0.263
k=3	0.304
k=4	0.274
k=5	0.283
k=6	0.280
k=7	0.248
k=8	0.245
k=9	0.224
k Best	0.224
Average	0.265

Table 3. DBI Value from Canberra Distance Calculation

Number of Clusters(k)	DBI Value
k=2	0.312
k=3	0.409
k=4	0.399
k=5	0.478
k=6	0.422
k=7	0.416
k=8	0.401
k=9	0.343
k Best	0.312
Average	0.398

Table 4. DBI Value from Chebychev Distance Calculation

Number of Clusters(k)	DBI Value
k=2	0.254
k=3	0.301
k=4	0.374
k=5	0.325
k=6	0.303
k=7	0.262
k=8	0.245
k=9	0.256
k Best	0.245
Average	0.290

The DBI value of grouping with the K-Means Clustering algorithm using k = 2 to k = 9 with the calculation of the Cosine Similarity distance is in Table 5. The DBI values for grouping with the K-Means Clustering algorithm use k=2 to k=9 with the Dynamic Time Warping Distance calculation shown in Table 6. The DBI value of grouping with the K-Means Clustering algorithm using k = 2 to k = 9 with the Jaccard Similarity distance calculation is in Table 7. DBI values for grouping with the K-Means Clustering algorithm using k=2 to k=9 with the calculation of the Manhattan Distance are in Table 8.

Table 5. DBI Value from Cosine Similarity Distance Calculation

Number of Clusters(k)	DBI Value
k=2	0.608
k=3	0.506
k=4	0.480
k=5	0.452
k=6	0.465
k=7	0.427
k=8	0.354
k=9	0.423
k Best	0.354
Average	0.464

Table 6. DBI Value of Dynamic Time Warping Distance Calculation

Number of Clusters(k)	DBI Value
k=2	0.277
k=3	0.347
k=4	0.439
k=5	0.359
k=6	0.346
k=7	0.406
k=8	0.364
k=9	0.366
k Best	0.277
Average	0.346

Table 7. DBI Value from Jaccard Similarity Calculation

Number of Clusters(k)	DBI Value
k=2	$-\infty$
k=3	$-\infty$
k=4	$-\infty$
k=5	$-\infty$
k=6	$-\infty$
k=7	$-\infty$
k=8	$-\infty$
k=9	$-\infty$
k Best	$-\infty$
Average	$-\infty$

Table 8. DBI Value from Manhattan Distance Calculation

Number of Clusters(k)	DBI Value
k=2	0.263
k=3	0.332
k=4	0.239
k=5	0.268
k=6	0.265
k=7	0.270
k=8	0.364
k=9	0.234
k Best	0.234
Average	0.279

The results of applying a numerical measure variation in the K-Means Clustering algorithm with case studies selling cosmetic products using several cluster numbers from the number of clusters 2 to 9 with an evaluation of the DBI value is the calculation of the

Euclidean Distance, which is the optimal/best cluster with the number of clusters 9 with a DBI value of 0.224. The average value of the test is 0.265. The following is a diagram of the average DBI value from each distance calculation, the best DBI value from each distance calculation, and the number of existing clusters, as seen in Figure 2 and Figure 3.

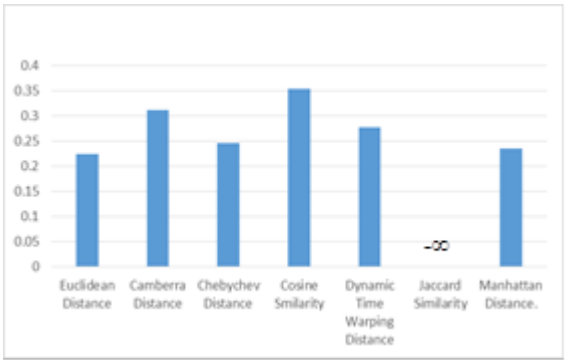


Figure 2. The Best Number of Clusters (k)

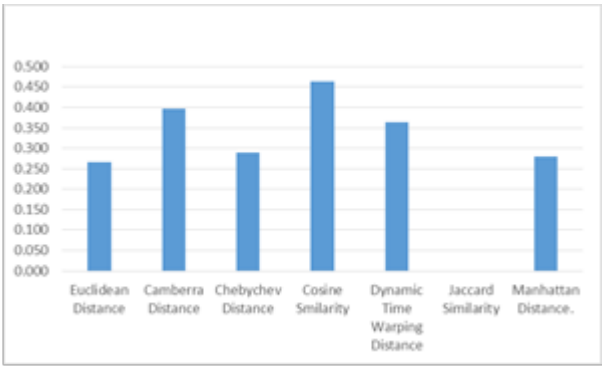


Figure 3. Average DBI value

The findings of this research reveal the identification of optimal clusters through the meticulous comparison of seven different distance calculations within the framework of the K-means algorithm, predicated on the analysis of DaviesBouldin index (DBI) values. Each distance calculation is thoroughly tested across the range of values for k, from k=2 to k=9. This study's discerned optimal cluster results can be effectively utilized for data grouping, facilitating informed and reasonable decisions. The results of this research are in line with or supported by research [13], where this research optimizes the number of clusters needed to ensure that policies can be taken regarding the grouping results, including providing that regional groups have ODP, PDP and Positive COVID-19 status in Riau Province. This research compares two distance measurements, namely Euclidean and Manhattan, to find the best grouping by looking for DBI values for the two distance measurements by examining data on the distribution of COVID-19 in the Riau region. The research results show that the lowest DBI values are at k=8 for Euclidean and k=7 for Manhattan, with values of 0.394 and 0.434. Additionally, DBI performs better on Euclidean than Manhattan as it has lower DBI values in all k tests.

This study showed better results, as evidenced by the remarkable achievement of the lowest recorded DaviesBouldin index (DBI) value of 0.224, obtained specifically when k was set to 7 and Euclidean Distance calculation was used. In contrast to previous research, which only examined two distance calculations using the K-means algorithm, this study expands the analysis to include a comprehensive evaluation of seven different distance calculations.

4. CONCLUSION

The research data uses a dataset of cosmetic product sales at Devi Cosmetics from January to April 2022, comprising 55 items. Before this data is used in research, the data is normalized first. The study results compare using seven distance calculations: Eu-

clidean Distance, Canberra Distance, Chebychev Distance, Cosine Similarity, Dynamic Time Warping Distance, Jaccard Similarity, and Manhattan Distance. The weakness of this research is that it needs to use larger data so that the results are more optimal. The optimal cluster is the Euclidean Distance calculation distance with a total of k = 9 with a DBI value = 0.224, and the smallest average DBI value is 0.265. Future research can compare with other methods to get optimal clusters in grouping data to increase accuracy in the form of grouping data.

5. DECLARATIONS

AUTHOR CONTIBUTION
All authors contributed to this article.
FUNDING STATEMENT
-
COMPETING INTEREST
There is no conflict of interest.

REFERENCES

[1] D. Priyanto, B. K. Triwijoyo, D. Jollyta, H. Hairani, and N. G. A. Dasriani, "Data Mining Earthquake Prediction with Multivariate Adaptive Regression Splines and Peak Ground Acceleration," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 3, pp. 583–592, jul 2023.

[2] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM (Jurnal Riset Komputer)*, vol. 8, no. 6, pp. 219–225, dec 2021.

[3] P. Subarkah, E. P. Pambudi, and S. O. N. Hidayah, "Perbandingan Metode Klasifikasi Data Mining untuk Nasabah Bank Telemarketing," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 139–148, sep 2020.

[4] L. Fimawahib and E. Rouza, "Penerapan K-Means Clustering pada Penentuan Jenis Pembelajaran di Universitas Pasir Pengaraian," *INOVTEK Polbeng - Seri Informatika*, vol. 6, no. 2, pp. 234–247, nov 2021.

[5] N. Dwitri, J. A. Tampubolon, S. Prayoga, F. I. R.H Zer, and D. Hartama, "Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 Di Indonesia," *Jurnal Teknologi Informasi*, vol. 4, no. 1, pp. 128–132, 2020.

[6] C. S. D. B. Sembiring, L. Hanum, and S. P. Tamba, "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Menentukan Judul Skripsi Dan Jurnal Penelitian (Studi Kasus Ftik Unpri)," *Jurnal Sistem Informasi dan Ilmu Komputer Prima(JUSIKOM PRIMA)*, vol. 5, no. 2, pp. 80–85, 2022.

[7] C. Satria and A. Anggrawan, "Aplikasi K-Means berbasis Web untuk Klasifikasi Kelas Unggulan," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 1, pp. 111 – 124, nov 2021.

[8] F. Dikarya and S. Muharni, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Universitas Terbaik Di Dunia," *Jurnal Informatika*, vol. 22, no. 2, pp. 124–131, 2022.

[9] F. Amin, D. S. Anggraeni, and Q. Aini, "Penerapan Metode K-Means dalam Penjualan Produk Souq.Com," *Applied Information System and Management (AISM)*, vol. 5, no. 1, pp. 7–14, 2022.

[10] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 4, no. 1, pp. 20–24, 2019.

[11] A. Ali, "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 19, no. 1, pp. 186–195, 2019.

[12] L. G. Rady Putra and A. Anggrawan, "Pengelompokan Penerima Bantuan Sosial Masyarakat dengan Metode K-Means," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 1, pp. 205–214, nov 2021.

[13] W. Gie and D. Jollyta, "Perbandingan Euclidean dan Manhattan Untuk Optimasi Cluster Menggunakan Davies Bouldin Index : Status Covid-19 Wilayah Riau," in *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS) 2020*, 2020, pp. 187–191.

[14] L. Han, F. Gao, B. Zhou, and S. Shen, "FIESTA: Fast Incremental Euclidean Distance Fields for Online Motion Planning of Aerial Robots," *IEEE International Conference on Intelligent Robots and Systems*, pp. 4423–4430, 2019.

[15] Y. Zhao, R. Dai, Y. Yang, F. Li, Y. Zhang, and X. Wang, "Integrated evaluation of resource and environmental carrying capacity during the transformation of resource-exhausted cities based on Euclidean distance and a Gray-TOPSIS model: A case study of Jiaozuo City, China," *Ecological Indicators*, vol. 142, no. July, p. 109282, 2022.

[16] M. Santosh and A. Sharma, "Proposed framework for emotion recognition using canberra distance classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 9, pp. 3778–3782, 2019.

[17] H. Ren, Y. Gao, and T. Yang, "A Novel Regret Theory-Based Decision-Making Method Combined with the Intuitionistic Fuzzy Canberra Distance," *Discrete Dynamics in Nature and Society*, vol. 2020, no. October, pp. 1–9, oct 2020.

[18] X. Gao and G. Li, "A KNN Model Based on Manhattan Distance to Identify the SNARE Proteins," *IEEE Access*, vol. 8, pp. 112 922–112 931, 2020.

[19] G. T. Pranoto, W. Hadikristanto, and Y. Religia, "Grouping of Village Status in West Java Province Using the Manhattan, Euclidean and Chebyshev Methods on the K-Mean Algorithm," *JISA(Jurnal Informatika dan Sains)*, vol. 5, no. 1, pp. 28–34, 2022.

[20] R. H. Singh, S. Maurya, T. Tripathi, T. Narula, and G. Srivastav, "Movie Recommendation System using Cosine Similarity and KNN," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 5, pp. 556–559, 2020.

[21] K. Park, J. S. Hong, and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Applied Artificial Intelligence*, vol. 34, no. 5, pp. 396–411, 2020.

[22] G. Ilharco, V. Jain, A. Ku, E. Ie, and J. Baldridge, "General Evaluation for Instruction Conditioned Navigation using Dynamic Time Warping," no. NeurIPS, jul 2019.

[23] W. S. Moola, W. Bijker, M. Belgiu, and M. Li, "Vegetable mapping using fuzzy classification of Dynamic Time Warping distances from time series of Sentinel-1A images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, no. June, pp. 1–16, oct 2021.

[24] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Information Sciences*, vol. 483, no. May, pp. 53–64, 2019.

[25] M. Tang, Y. Kaymaz, B. L. Logeman, S. Eichhorn, Z. S. Liang, C. Dulac, and T. B. Sackton, "Evaluating single-cell cluster stability using the Jaccard similarity index," *Bioinformatics*, vol. 37, no. 15, pp. 2212–2214, 2021.

[26] S. S. Khan, Q. Ran, M. Khan, and M. Zhang, "Hyperspectral image classification using nearest regularized subspace with Manhattan distance," *Journal of Applied Remote Sensing*, vol. 14, no. 03, p. 1, 2019.

[27] N. Li and S. Wan, "Research on Fast Compensation Algorithm for Interframe Motion of Multimedia Video Based on Manhattan Distance," *Journal of Mathematics*, vol. 2022, no. June, pp. 1–10, jan 2022.