VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.5545

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

COMPARATIVE EVALUATING NUMERICAL MEASURE VARIATIONS IN K-MEDOIDS CLUSTERING FOR EFFECTIVE DATA GROUPING

Relita Buaton[1*]; Solikhun[2]

Information Systems study program[1]
Sekolah Tinggi Manajemen Informatika dan Komputer Kaputama, Binjai, Indonesia[1]
https://header.kaputama.ac.id/index.html[1]
bbcbuaton@gmail.com[1*]

Informatics Engineering Study Program[2]
Sekolah Tinggi Ilmu Komputer Tunas Bangsa, Pematang Siantar, Indonesia[2]
https://stikomtunasbangsa.ac.id/[2]
Solikhun@amiktunasbangsa.com[2]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract *— The K-Medoids Clustering algorithm is a frequently employed technique among researchers for data categorization. The primary difficulty addressed in this investigation pertains to the extent of optimality achieved when varying distance computation methodologies are applied within the framework of K-Medoids Clustering. This study is primarily concerned with the application of K-Medoids Clustering, employing a multitude of distance calculation methods, specifically those involving numerical metrics. The aim is to undertake a comparative analysis of Davies-Bouldin Index (DBI) values in order to ascertain the most productive distance calculation technique. In this research, the distance calculation methodologies include Manhattan Distance, Jaccard Similarity, Dynamic Time Warping Distance, Cosine Similarity, Chebyshev Distance, Canberra Distance and Euclidean Distance. The dataset consists of sales data from Devi Cosmetics, covering the period between January and April 2022 and comprising 56 distinct sales items. The research provides an exhaustive evaluation of numerical metrics concerning the K-Medoids Clustering algorithm. The findings indicate that the optimal clustering is achieved using the Chebyshev distance, resulting in 9 clusters with a DBI value of 166.632. The study's contribution is that it can improve more optimal data grouping to help make decisions correctly.*

Keywords*: clustering, comparison, distance metrics, k-medoids, numerical measure*.

Intisari— *Algoritma K-Medoids Clustering adalah teknik yang sering digunakan di kalangan peneliti untuk kategorisasi data. Kesulitan utama yang diatasi dalam penyelidikan ini adalah mempertahankan tingkat optimalitas yang dicapai ketika berbagai metodologi komputasi jarak diterapkan dalam kerangka K-Medoids Clustering. Studi ini terutama berkaitan dengan penerapan K-Medoids Clustering, yang menggunakan banyak metode penghitungan jarak, khususnya yang melibatkan metrik numerik. Tujuannya adalah untuk melakukan analisis komparatif nilai Indeks Davies-Bouldin (DBI) untuk memastikan teknik penghitungan jarak yang paling produktif. Dalam penelitian ini metodologi perhitungan jarak meliputi Manhattan Distance, Jaccard Kemiripan, Dynamic Time Warping Distance, Cosine Kemiripan, Chebyshev Distance, Canberra Distance dan Euclidean Distance. Kumpulan data tersebut terdiri dari data penjualan Devi Cosmetics, yang mencakup periode antara Januari dan April 2022 dan terdiri dari 56 item penjualan yang berbeda. Penelitian ini memberikan evaluasi menyeluruh terhadap metrik numerik mengenai algoritma K-Medoids Clustering. Hasil temuan menunjukkan bahwa clustering optimal dicapai dengan menggunakan jarak Chebyshev, sehingga menghasilkan 9 cluster dengan nilai DBI sebesar 166.632. Kontribusi penelitian ini dapat meningkatkan pengelompokan data yang lebih optimal untuk membantu pengambilan keputusan secara tepat.*

Kata Kunci*: pengelompokan, perbandingan, metrik jarak, k-medoid, ukuran numerik.*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5545

INTRODUCTION

Data mining refers to the systematic procedure aimed at extracting valuable patterns, insights, and knowledge from extensive datasets. This multifaceted process incorporates a spectrum of techniques and algorithms, effectively unveiling concealed information, identifying trends, correlations, and patterns embedded within the data. It serves as a fundamental tool utilized across diverse domains, including machine learning, statistics, and artificial intelligence, enabling the revelation of substantial and actionable information pivotal for decision-making, prediction, and knowledge acquisition. The comprehensive scope of data mining encompasses various tasks, spanning from data pre-processing, exploration, pattern recognition to model building, and finds application in multifarious fields such as marketing, finance, healthcare, among others [1].

Data mining involves the systematic extraction of valuable and significant information from extensive datasets. There are several established methodologies within data mining, including the K-Means algorithm [1], K-Medoids [2], Decision Tree [3], Naive Bayes [4], Apriori [5], and Neural Network [6]. Every one of these approaches has unique advantages and disadvantages. The best strategy to use depends on the particular goals of the data analysis as well as the type of information that has to be extracted. A combination of techniques is frequently used to obtain a deeper comprehension and insights from the data.

While K-Medoids and K-Means are comparable clustering algorithms in data mining, they differ significantly in how they identify the cluster center. The average (mean) of all the data points in the group is what the K-Means method uses to calculate the group center. Meanwhile, in K-Medoids, the centre of the group is represented by one of the actual data points in the group. The data point used as the centre of this group is called a "medoid" [6].

This study [7], evaluates the impact of commercial centers in downtown Guangzhou using kernel density analysis on data from five types of commercial POIs. It correlates the integration of commercial sectors with urban vitality using population big data. Key findings indicate that living, business, financial, and leisure sectors significantly influence daytime pedestrian flow, with leisure sectors enhancing vitality both day and night. Mixed commercial sectors boost urban vitality more on weekdays, and diversifying commercial forms benefits 24-hour activity. Limitations include reliance on Tencent travel data and pre-COVID-19 data, suggesting a need for updated information and more detailed future analyses.

The research [8] aims to assess the most efficient grouping approach between two distinct methodologies, specifically, the K-Means and K-Medoids algorithms, for classifying fresh milk production. The author conducted a comparative analysis of the grouping outcomes, utilizing the lowest Davies Bouldin Index (DBI) value as the criterion for determining optimality. The dataset employed for this assessment comprises fresh milk production statistics from Indonesia spanning the years 2018 to 2020, taken from the Central Statistics Office of Indonesia. The evaluation findings show that the K-Means Clustering algorithm has a DBI value of 0.094 and the K-Medoids Clustering algorithm has a DBI value of 0.072. This difference shows that, when compared to the K-Means Clustering method, the use of the K-Medoids Clustering algorithm results in a lower DBI value, exactly 0.072. As a result, the K-Medoids algorithm, when it comes to grouping fresh milk production, performs better than the K-Means algorithm.

With a total of 1,061 data points, the research [9] found that student and academic data might be used for clustering incoming undergraduate students at the Faculty of Information Technology, Universitas Budi Luhur. Using the K-Means technique in RapidMiner Studio v.9, optimal clustering was accomplished with six clusters (k=6), producing a Davies Bouldin Index (DBI) score of 1.597. Cluster 4 had the most members (395), followed by clusters 6 (331), 5 (116), 3 (114), 2 (40), and 1 (35). The most selected study program was Informatics Engineering, followed by Information Systems, with Computer Systems being the least chosen. The highest number of students came from the SAINTEK major in cluster 4. Future research should focus on system development and incorporate additional selection data for improved clustering.

In the study documented under reference [10], protein conformational landscapes are essential for understanding biological processes and therapeutic design. Traditional structural biology has limitations, but molecular dynamics (MD) simulations offer detailed insights. Despite advances, understanding long-distance allosteric communication in proteins remains difficult. Allostery, crucial for cellular signaling, involves interactions between distant protein regions. This thesis presents the CARDS method to map allosteric networks by analyzing structural and dynamic changes. CARDS was applied to study G protein

activation linked to cancer and a druggable pocket in ebolavirus VP35. Integrating MD with experiments, the research identified potential drug targets, including for SARS-CoV-2.

Research [11] Researchers conducted a study to group temperature data originating from Riau Province, obtained from the Riau Central Statistics Agency (BPS) during the period 2019 to 2021. This research aims to compare and determine the most appropriate algorithm for grouping temperatures by testing its validity using the Davies Bouldin Index (DBI), where the smaller the DBI value indicates, the better the cluster grouping results. Based on experimental results, the K-Means algorithm shows the best DBI value of 0.2 at K=6 after 10 iterations, while the K-Medoids algorithm obtains the best value of 0.279 at K=8 after 100 iterations. These results show that the K-Means algorithm is superior in grouping temperatures in Riau Province compared to K-Medoids.

Research [12] This study examined how educational data mining can enhance teaching and learning for 100 master's students at Politecnico di Torino. Using Excel, PowerBI, and RapidMiner, student data was analyzed to assess performance and define profiles. An intelligent decision-support system was proposed to suggest improvements like flipped classes and extra resources. Future research should expand data collection and improve quality, with similar analyses planned in the Data2Learn@Edu project schools. This aligns with Sustainable Development Goal 4, promoting innovative teaching and lifelong learning.

Research [13] In this research, 1,529 items of stock data from the Ben Waras Clinic were used and evaluated through a manual calculation process and also through the use of Rstudio software to carry out computational calculations. In manual calculations, the results indicated the existence of 3 clusters, while using the Rstudio application also produced 3 clusters. In the K-Means algorithm using Rstudio, the results of the average outgoing goods show that in cluster 1, the highest average outgoing goods occurred in June, namely 87; in cluster 2, the highest average outgoing goods were recorded in January, with a total of 227; and in cluster 3, the highest average of outgoing goods was in August, namely around 14.9. Meanwhile, the results of the average outgoing goods using the K-Medoids algorithm in Rstudio show that in cluster 1, the highest average outgoing goods occurred in July, around 11.9; in cluster 2, the highest average outgoing goods was recorded in February, around 24.5; and in cluster 3, the highest average outgoing goods were found in January, around 227.

Investigate [14] In order to prevent and control the spread of diarrheal infections among children under five in Kuningan Regency, a priority regional mapping plan was developed. Data mining clustering is the methodology used, and two algorithms—the K-Means and K-Medoids algorithms—are compared. The Elbow and Silhouette Coefficient approaches were used to determine the ideal number of clusters. The K-Means approach is found to be optimal with three clusters, whereas the K-Medoids strategy yields two clusters. The DBI value of the K-Means algorithm is lower than that of K-Medoids, both in 2 clusters and 3 clusters, according to evaluation using the Davies-Bouldin Index (DBI) approach, highlighting the superiority of the K-Means algorithm.

Research [15] This study involves a comparative analysis between the K-Means and K-Medoids algorithms, followed by a validation test of the formed clusters. Employing the Davies-Bouldin Index for cluster analysis, the validity value stands at 0.67 for K-Means Clustering and 1.78 for K-Medoids. Based on the recorded validity values, the K-Means algorithm is selected for application in developing a web-based vehicle fleet cluster due to its higher relevance and lower DBI validity values compared to K-Medoids. Validation of the cluster results in the web application demonstrates a 97% conformity both through the use of the Rapidminer tool and manual calculations.

Research [16] This study is concentrated on two specific allergies: seafood and airborne allergies. The dataset utilized spans from 2011 to 2019, sourced from the Central Statistics Agency. Employing data mining techniques, particularly leveraging the k-medoids clustering method for data processing, this research examines the prevalence of allergies among children across various provinces. This facilitates the identification of provincial groupings based on allergy prevalence, categorizing provinces into three clusters: a low cluster comprising 2 provinces, a medium cluster encompassing 30 provinces, and a high cluster encompassing 2 provinces, predicated on the percentage of allergy prevalence among toddlers in Indonesia. The objective of this study is to furnish insights to health departments, especially community health centers, concerning the categorization of allergic diseases among children in Indonesia, aiming to influence the distribution of anti-allergic immunization throughout the country.

Research [17] undertakes the grouping of cabin crew utilizing data mining clustering techniques during the data processing phase, which involves the elimination of missing values and attribute determination, resulting in a dataset of

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5545

100 instances. Subsequently, at the modeling stage, the most optimal outcomes are achieved through the utilization of the k-means algorithm, forming 4 clusters based on 6 attributes. The evaluation, as indicated by the Davies Bouldin Index (DBI), portrays a value of 0.792 for the k-means algorithm, 0.812 for the x-means algorithm, and 1.700 for the k-medoids algorithm.

Referencing research [18], this study focuses on clustering regions affected by ISPA (Acute Respiratory Infections) in Karawang Regency. These areas are categorized into low, medium, and high groups based on the spread of ISPA. A comparative analysis of distance measures was conducted to determine the most suitable model, assessed through the Davies Bouldin Index (DBI). Employing the Euclidean distance yields a DBI value of 0.088, while the utilization of Chebyshev distance results in a DBI value of 0.116. The efficacy of the K-Medoids algorithm employing Euclidean distance is deemed superior to Chebyshev distance, evidenced by its DBI value, which approaches 0.

K-medoids are superior when the data has outliers, is of different scales, is qualitative or discrete, has a complex or non-linear structure, or is high-dimensional. This advantage makes K-Medoids a better choice for these Devi Cosmetics product sales data types than other methods, such as K-Means. However, K-Means may be more efficient in computing time on large and homogeneous datasets. The main challenge in this research is to determine the most optimal distance metric for the K-Medoids method among the various available options, including Manhattan Distance, Jaccard Similarity, Dynamic Time Warping Distance, Cosine Similarity, Chebyshev Distance, Canberra Distance, and Euclidean Distance. This research compares these distance metrics in the K-Medoids algorithm to identify the best one based on the smallest Davies-Bouldin Index (DBI) value. Optimal grouping results can be used to make appropriate decisions regarding the sale of cosmetic products at Devi Kosmetik.

This research is innovative in that it compares seven distance computation approaches using different k-tests to optimize clustering with the K-Medoids algorithm, and then finds the ideal clusters by minimizing the DBI value. Prior studies concentrated on maximizing the number of clusters by utilizing solely Chebyshev and Euclidean distances. To improve upon it and produce more ideal grouping results, this study adds a number of new distance measurements. According to the research's consequences, the conclusions can be used to enhance the data grouping procedure, enabling more precise decision-making based on the clustered data.

MATERIALS AND METHODS

Research Framework

A research framework, consisting of phases or stages, is constructed in order to accomplish the study objectives. The research framework that the author has put up is as follows:

1. Data Collection: This initial step involves gathering data pertinent to your research topic. Depending on the research objectives, the data can be in the form of images, text, numbers, or other types.
2. Data Normalization: Data normalization aims to ensure that each feature makes a balanced contribution when calculating the distance between data points. The first step in the data normalization process using the k-medoids method is to collect the data to be grouped and examine its characteristics, including the type of feature, whether numeric, categorical, or a combination of both, as well as the range of values for each feature to find out whether there are features with a vast scale. Different. Depending on the characteristics of the data, you can choose a suitable normalization method, such as Min-Max Scaling, which reorders the data into the range [0, 1], or Z-score Normalization, which makes the data have a mean of 0 and a standard deviation 1.
3. In this stage, the data are modified to fit a predetermined scale. When employing distance measurements like the Manhattan Distance, Chebyshev Distance, Canberra Distance, and Euclidean Distance, normalization is essential since the data scale can have a big impact on these metrics.
4. Distance Calculation with Various Metrics: In this step, you will compute the distances between pairs of data points using different metrics, including Manhattan Distance, Jaccard Similarity, Dynamic Time Warping Distance, Cosine Similarity, Chebyshev Distance, Canberra Distance and Euclidean Distance. This provides insights into the similarities and differences between data pairs from various perspectives.
5. Clustering Results: You will group the data based on the calculated distance metric. Data points that are similar or close together will be combined into clusters or groups. Devi Cosmetics product sales data grouping uses the K-Medoids grouping method using distance calculations: Manhattan Distance, Jaccard Similarity, Dynamic Time Warping Distance,

Cosine Similarity, Chebyshev Distance, Canberra Distance and Euclidean Distance.

6. DBI Evaluation: The Davies-Bouldin Index (DBI) is used to evaluate the quality of the resulting clusters. This metric helps assess the effectiveness of clustering. The smaller the DBI value, the better the grouping.
7. Conclusion: This is the final part of your research, where you will summarize your findings, provide an interpretation of the grouping results, evaluate the DBI, and provide a conclusion on whether your research objectives were achieved. You can also discuss the implications of your findings in the context of your problem.

Data Normalization

The study's dataset, which includes 55 items, tracks Devi Cosmetics' sales of cosmetic products from January to April of 2022. Equation (1) is the formula used to achieve data normalization:

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

where:
x' = normalization result, x = data to be normalized, a = smallest data from dataset and b = largest data from dataset. The normalized sales data for Devi Cosmetics from January to April 2022 is presented in Table 1:

Table 1. The function of Power Supply Components

No	January 2022	February 2022	March 2022	April 2022
1	0.1111	0.1944	0.3056	0.0278
2	0.3611	0.3333	0.4167	0.2222
3	0.0556	0.1944	0.2778	0.0833
4	0.2778	0.3889	0.1667	0.0556
5	0.9444	0.1667	0.1667	0.3056
6	0.7222	0.1944	0.2222	0.0556
7	0.2222	0.0833	0.3333	0.3333
8	0.0556	0.2500	0.0833	0.2778
9	1.0000	0.3056	0.2500	0.3611
10	0.0000	0.1111	0.0833	0.0833
11	0.1944	0.1111	0.3333	0.0000
12	0.1667	0.1389	0.3333	0.0278
13	0.1944	0.0556	0.0000	0.0833
14	0.0556	0.2500	0.2222	0.1944
15	0.1944	0.0278	0.1667	0.1667
16	0.0000	0.1389	0.3056	0.0556
17	0.1111	0.1944	0.0000	0.0278
18	0.0278	0.0278	0.0556	0.0278
19	0.0278	0.1389	0.0556	0.0278
20	0.0833	0.2222	0.1111	0.0000
21	0.5000	0.4167	0.5556	0.1944
22	0.5833	0.6389	0.6389	0.2222
23	0.0000	0.1389	0.3333	0.5833
24	0.1111	0.0000	0.1944	0.0278
25	0.0278	0.2500	0.0556	0.0833
26	0.6389	0.4444	0.4444	0.2778
27	0.3056	0.1111	0.0556	0.0278
28	0.1111	0.1944	0.1111	0.0556
29	0.3333	0.0833	0.1667	0.1111
30	0.2222	0.3056	0.4444	0.2222
31	0.2778	0.3611	0.2500	0.1389
32	0.0556	0.2222	0.2778	0.1111
33	0.1389	0.1111	0.1111	0.0556
34	0.0556	0.0278	0.0278	0.0278
35	0.5556	0.5000	0.6944	0.2500
36	0.2500	0.1111	0.1111	0.0833
37	0.1389	0.3056	0.2222	0.1111
38	0.2500	0.1111	0.1389	0.2222
39	0.1111	0.1944	0.1389	0.1111
40	0.1111	0.1389	0.2778	0.0833
41	0.1389	0.2222	0.0278	0.0833
42	0.3611	0.4167	0.2778	0.1111
43	0.0556	0.0000	0.0556	0.0556
44	0.4444	0.2222	0.4167	0.1667
45	0.1667	0.0833	0.2500	0.0556
46	0.4167	0.3333	0.1111	0.3333
47	0.0833	0.0833	0.2778	0.0556
48	0.3056	0.5000	0.1667	0.3333
49	0.3333	0.4444	0.4167	0.1667
50	0.2778	0.2778	0.1111	0.1389
51	0.1667	0.3333	0.4444	0.1667
52	0.0278	0.1667	0.0000	0.0000
53	0.3056	0.3056	0.0556	0.3889
54	0.2500	0.0833	0.1389	0.0556
55	0.3056	0.6111	0.0556	0.3056
56	0.1944	0.1667	0.0833	0.0833

Source: (Research Results, 2024)

RESULTS AND DISCUSSION

This research introduces a novel approach to identifying the optimal clusters by comparing seven distance metrics within the K-Medoids algorithm using Davies-Bouldin Index (DBI) values. The Davies-Bouldin Index (DBI) is the basis for cluster evaluation because it combines two essential aspects in clustering: cohesiveness within clusters and separation between clusters. Within-cluster compactness refers to how close or homogeneous the data is in one cluster, while inter-cluster separation measures how far apart the clusters are. A smaller DBI value indicates a better cluster because it indicates that the data in the cluster is more compact and centred around its medoid or centroid, while the clusters are more separated. This combination produces a more precise, well-defined, and reliable cluster structure for further analysis. Thus, DBI helps ensure that the clustering algorithm has produced optimal clusters in terms of homogeneity and differentiation between clusters. Each distance metric is tested with k values ranging from 2 to 9. The results of this optimization can be utilized for effective data grouping, aiding in optimal decision-making.

The study evaluates seven distance metrics for the purpose of grouping cosmetic sales. The DBI is used to evaluate these metrics, and the cluster

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5545

with the lowest DBI value is considered ideal. These distance measures have the following formulas:

a. Euclidean Distance

Euclidean Distance [9], serves as a metric embedded within Euclidean geometry, utilized for quantifying the spatial separation between two points within a dimensional space. It delineates the magnitude of the straight line segment connecting these points. Specifically, in a two-dimensional context, the Euclidean Distance between points (x1, y1) and (x2, y2) is ascertained by employing the formula expounded in Equation (2).

$$dij = \sqrt{\sum_{k=1}^{m} xij - cij^2} \quad (2)$$

Information:

Dij = represents the spatial separation between data values and cluster center values, m = denotes the number of data dimensions. Xij = data values within the k-th dimension, Xjk = denotes the cluster center values within the same dimension [19].

The test results show that the results of the DBI calculation use Euclidean distance calculations for various numbers of clusters (k) in a dataset. DBI is used as an evaluation metric in clustering to measure the quality of the clusters produced by the clustering algorithm. Lower DBI values indicate better or denser clustering. Based on testing, the lowest DBI value is when the number of clusters is 2, namely 611,223. A lower DBI value indicates better clustering. In this context, clustering with 2 clusters in the Euclidean distance calculation gives the best results based on DBI.

b. Canberra Distance

The Canberra Distance [20] serves as a metric for gauging dissimilarities between two vectors or points within a multidimensional space. Widely employed in data analysis, particularly in scenarios involving high-dimensional data with diverse attributes, it stands out as an alternative to distance metrics like the Euclidean or Manhattan distances. Notably, the Canberra Distance adjusts for variations in attribute value scales and magnitudes. Its formula, represented by Equation (3), is as follows:

$$dij = \sqrt{\sum_{k=1}^{m} \frac{|aik - ajk|}{|aik| - |ajk|}} \quad (3)$$

Information:

dij = difference level, m = number of vectors, aik = input image vector and ajk = comparison image vector [21].

The test results using Canberra distance calculations show the results of DBI calculations for various numbers of clusters (k) in a dataset. All DBI values for each number of clusters from k=2 to k=9 are infinite (∞). DBI is supposed to provide numerical values that evaluate the quality of clusters in data clustering. All DBI values for various numbers of clusters are infinite (∞), which indicates there is a problem or error in the calculation or use of DBI. An infinite value (∞) in the context of DBI usually indicates some scenario in which the measurement or calculation of the distance between clusters or data points cannot be performed correctly.

c. Chebychev Distance

Chebyshev Distance, also recognized as Supremum Distance or Infinity Norm, as introduced by Gao [22], stands as a metric employed for quantifying the maximal spatial separation between two points within a multidimensional space. This metric evaluates the most prominent distinction between the corresponding components of two vectors or points. The formulation for Chebyshev Distance, as depicted in Equation (4):

$$dij = max_k \, |X_{ij} - X_{ik}| \quad (4)$$

The provided expression is employed to compute the disparity or dissimilarity between elements in two rows (indexed as i and k) of a matrix X. This disparity is determined by identifying the maximum value (largest value) of the discrepancy in values between elements within the same column (indexed as j) across both rows [23].

The test shows the results of the Davies-Bouldin Index (DBI) calculation using Chebyshev distance calculations for various numbers of clusters (k) in a dataset from k=2 to k=9. DBI is used as an evaluation metric in clustering to measure the quality of the clusters produced by the clustering algorithm. In this test, the lowest DBI value is k=2, namely 166.632.

d. Cosine Similarity

Cosine Similarity, as delineated by Singh [24], is a metric utilized for quantifying the degree of similarity between two vectors within a multidimensional space, particularly relevant in the domains of data analytics and text processing. This metric assesses the angle between two vectors rather than their Euclidean distance. Cosine Similarity ranges from -1 to 1, with higher values

indicating greater similarity between the vectors. The formula for Cosine Similarity is presented as specified in Equation (5).

$$Cosine\ Similarity = \frac{A,B}{||A||,||B||} \quad (5)$$

Information:

A = is the weight of each feature in vector A and B = is the weight of each feature in vector B [25].

The results of the Davies-Bouldin Index (DBI) calculation use cosine similarity distance calculations for various numbers of clusters (k) in a dataset from k=2 to k=9 with the lowest DBI value being k=2 with a DBI value=343,885. DBI is used as an evaluation metric in clustering to measure the quality of the clusters produced by the clustering algorithm. Lower DBI values indicate better or denser clustering.

e. Dynamic Time Warping Distance

A method for measuring the similarity between two temporal data sequences or time series, which may have different lengths or temporal distortions, is called Dynamic Time Warping (DTW) Distance [10]. DTW is not like typical linear comparison methods in that it is an algorithmic tool for aligning two temporal trajectories that may show different rates or patterns of change. Equation (6) has the formulation for dynamic time warping provided.

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

Information:

m = the number of variables A and B, A1 = the ith data matrix A and B1 = the ith data matrix B [26].

The results of the Davies-Bouldin Index (DBI) calculation using Dynamic Time Warping (DTW) distance calculations for various numbers of clusters (k) in a dataset show that the lowest DBI value is k=2, namely 529,982. Because a lower DBI value indicates better clustering, in this context, clustering with 2 clusters provides the best results based on the DBI value.

f. Jaccard Similarity

Jaccard Similarity [27], constitutes a metric for evaluating the similarity between two sets. This metric quantifies the degree of overlap between elements of the two sets relative to the total number of unique elements in both sets. Jaccard Similarity is widely applicable in data analytics contexts involving sets, such as text mining, cluster analysis, and content-based recommendation systems. The formula for computing Jaccard Similarity between two sets, denoted as A and B , is formally defined in Equation (7).

$$J(x, y) = \frac{\sum_i^p x_i y_i}{\sum_{j=1}^p x_i^2 + \sum_{j=1}^p y_i^2 - \sum_{j=1}^p x_i y_i} \quad (7)$$

Information:

x = the value of the key and y = the value of the document [28].

All DBI values for each number of clusters from k=2 to k=9 are infinite (∞). DBI is supposed to provide numerical values that evaluate the quality of clusters in data clustering. All DBI values for various numbers of clusters are infinite (∞), which indicates there is a problem or error in the calculation or use of DBI. An infinite value (∞) in the context of DBI usually indicates some scenario in which the measurement or calculation of the distance between clusters or data points cannot be performed correctly.

g. Manhattan Distance

Manhattan Distance [29], which is a multidimensional measure of the distance between two places, is sometimes referred to as City Block Distance or L1 Distance. The term "Manhattan" refers to the city's street grid in New York City, in which transportation between locations requires following pathways perpendicular to the coordinate axes. The formula for Manhattan Distance is shown in Equation (8):

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

The difference or distance between vectors I and j is calculated using this formula. The difference (absolute value) between the respective components of the two vectors is added up to determine this distance [30].

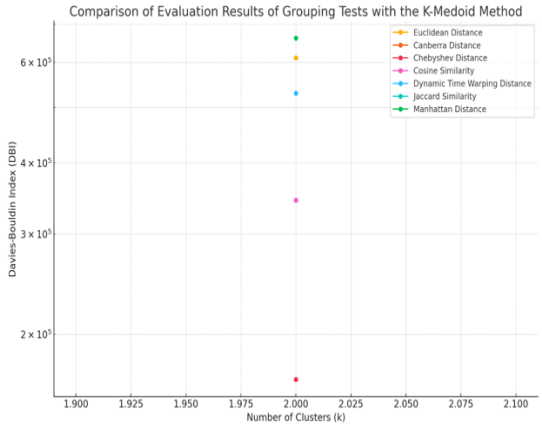
Test results using Jaccard Similarity distance calculations show the results of DBI calculations for various numbers of clusters (k) in a dataset. The results of the Davies-Bouldin Index (DBI) calculation using Manhattan distance calculations for various numbers of clusters (k) in a dataset from k=2 to k=9 show that the DBI value has the lowest value, k=2, namely 662,031. Since lower DBI values indicate better clustering, in this context, clustering with 2 clusters provides the best results based on DBI.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5545

The results of the Davies-Bouldin Index (DBI) calculation using Manhattan distance calculations for various numbers of clusters (k) in a dataset from k=2 to k=9 show that the DBI value has the lowest value, k=, namely 662,031. Since lower DBI values indicate better clustering, in this context, clustering with 2 clusters provides the best results based on DBI.

From the test results above, it can be explained according to Figure 1 below:



Source: (Research Results, 2024)

Figure 1. Comparison of Evaluating Results of Grouping Tests with the K-Medoid Method

The following graph compares the evaluation results of grouping tests using the K-Medoid method in sales of Devi Cosmetics cosmetic products with variations in several distance calculations. Using multiple distance metrics, this graph depicts the Davies-Bouldin Index (DBI) values for various clusters (k) from k=2 to k=9. The graphic explanation in Figure 1 is:

1. Euclidean Distance: The lowest DBI value is at k=2 with a value of 611.223, indicating the best cluster based on this metric.
2. Canberra Distance: All DBI values from k=2 to k=9 are infinite (∞), indicating a problem in the calculation or use of DBI.
3. Chebyshev Distance: The lowest DBI value is at k=2, 166.632, indicating the best cluster based on this metric.
4. Cosine Similarity: The lowest DBI value is at k=2 with a value of 343.885, indicating the best cluster based on this metric.
5. Dynamic Time Warping Distance: The lowest DBI value is at k=2, 529.982, indicating the best cluster based on this metric.
6. Jaccard Similarity: All DBI values from k=2 to k=9 are infinite (∞), indicating a problem in the calculation or use of DBI.
7. Manhattan Distance: The lowest DBI value is at k=2 with a value of 662.031, indicating the best cluster based on this metric.

Lower DBI values indicate better clusters because they indicate cohesiveness within clusters and better separation between clusters. Based on the above results, most of the distance metrics show k=2 as the optimal number of clusters, except for the Canberra and Jaccard metrics, which show infinite DBI values (∞), indicating a problem in the calculation.

The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering results, where lower DBI values indicate better performance due to more compact and well-separated clusters, with a value of 0 being ideal for perfect clustering. In the analysis of the results, most distance metrics identified k=2as the optimal number of clusters, suggesting that dividing the data into two clusters achieves the best separation and cohesiveness. This generally indicates that the dataset naturally divides into two distinct groups. However, the Canberra and Jaccard metrics resulted in infinite DBI values (∞), highlighting significant issues in the calculation, which could be due to zero distance problems, inappropriate metrics for the data type, or handling sparse or binary data. These findings suggest a need to reassess the data preprocessing steps and consider using different distance metrics more suited to the data type, especially if it is sparse, binary, or has unique characteristics that might cause calculation issues. Addressing these potential problems can lead to a more robust clustering analysis.

CONCLUSION

This research uses a dataset that records sales of cosmetic products at Devi Cosmetics from January to April 2022, which includes 55 items. Before being used in research, the data underwent normalization. This investigation compares the effectiveness of calculating seven distances. One limitation of this study is its reliance on a more extensive data set to improve the optimization of results. Several existing distance calculations show that the Chebychev Distance calculation is the best distance calculation using the K-Medoids grouping method for grouping sales of cosmetic products at Devi Cosmetics with a total of k = 2 and a DBI value of 166,632. This grouping has the smallest DBI value compared to other distance calculations. This shows that the grouping results with Chebychev are more precise than those of other distance calculations. Future research could explore alternative

methodologies to achieve optimal clustering, thereby improving the accuracy of data clustering.

REFERENCE

[1] D. Laila Sari, M. Saputra, and H. Gemasih, "Penerapan Data Mining Dalam Proses Prediksi Perceraian Menggunakan Algoritma Naive Bayes Di Kabupaten Aceh Tengah," *Jurnal Teknik Informatika Dan Elektro*, vol. 4, no. 1, pp. 23–35, 2022, doi: https://doi.org/10.55542/jurtie.v4i1.112

[2] S. D. Nirwana, M. I. Jambak, and A. Bardadi, "Perbandingan Algoritma K-Means Dan K-Medoids Dalam Clustering Rata-Rata Penambahan Kasus Covid-19 Berdasarkan Kota/Kabupaten Di Provinsi Sumatera Selatan," *JSiI (Jurnal Sistem Informasi)*, vol. 9, no. 2, pp. 126–131, 2022, doi: https://doi.org/10.30656/jsii.v9i2.5127

[3] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends,* vol. 2, no. 1, pp. 20–28, 2021, doi: https://doi.org/10.38094/jastt20165

[4] 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, p. 219, 2021, doi: https://doi.org/10.30865/jurikom.v8i6.3655

[5] M. H. Santoso, "Application of association rule method using apriori algorithm to find sales patterns: Case study of Indomaret Tanjung Anom," *Brilliance: Research of Artificial Intelligence*, vol. 1, no. 2, pp. 54–66, 2021.

[6] H. Putra and N. Ulfa Walmi, "Penerapan Prediksi Produksi Padi Menggunakan Artificial Neural Network Algoritma Backpropagation," *Jurnal Nasional Teknologi Dan Sistem Informasi*, vol. 6, no. 2, pp. 100–107, 2020, doi: https://doi.org/10.25077/teknosi.v6i2.2020.100-107

[7] L. Liu, Y. Dong, W. Lang, H. Yang, and B. Wang, "The Impact of Commercial-Industry Development of Urban Vitality: A Study on the Central Urban Area of Guangzhou Using Multisource Data," *Land*, vol. 13, no. 2, p. 250, 2024.

[8] M. Wahyudi and L. Pujiastuti, "Komparasi K-Means Clustering dan K-Medoids dalam Mengelompokkan Produksi Susu Segar di Indonesia," *Jurnal Sistem Informasi,* vol. 4, no. 2, pp. 243–254, 2022, doi: https://doi.org/10.30812/bite.v4i2.2104

[9] T. A. Terlep, M. R. Bell, T. M. Talavage, and D. L. Smith, "Euclidean distance approximations from replacement product graphs," *IEEE Transactions on Image Processing,* vol. 31, pp. 125–137, 2021.

[10] 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, 2023.

[11] F. Faisal, L. A. Giopani, M. Fitriah, Z. C. Dwynne, and S. Syahidatul, "Comparison of K-Means and K-Medoids Algorithms for Temperature Grouping in Riau Province," *Jurnal Teknologi dan Sistem Informas*i, vol. 2, no. 2, pp. 128–134, 2022.

[12] C. G. Demartini, L. Sciascia, A. Bosso, and F. Manuri, "Artificial Intelligence Bringing Improvements to Adaptive Learning in Education: A Case Study," *Sustainability*, vol. 16, no. 3, p. 1347, 2024.

[13] B. A. Setiawan and Sulastri, "Perbandingan Clustering Optimalisasi Stok Barang Menggunakan Algoritma K–Means dan Algoritma K–Medoids," *Jurnal Sistem dan Informatika,* vol. 978–979, 2021.

[14] T. S. Syamfithriani, N. Mirantika, and R. Trisudarmo, "Perbandingan Algoritma K-Means dan K-Medoids Untuk Pemetaan Daerah Penanganan Diare Pada Balita di Kabupaten Kuningan," *Jurnal Sistem Informasi Bisnis,* vol. 12, no. 2, pp. 132–139, 2023, doi: https://doi.org/10.21456/vol12iss2pp132-139

[15] A. Supriyadi, A. Triayudi, and I. D. Sholihati, "Perbandingan Algoritma K-Means Dengan K-Medoids Pada Pengelompokan Armada Kendaraan Truk Berdasarkan Produktivitas," *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, vol. 6, no. 2, pp. 229–240, 2021, doi: https://doi.org/10.29100/jipi.v6i2.2008

[16] H. Ningrum, E. Irawan, and M. R. Lubis, "Implementasi Metode K-Medoids Clustering Dalam Pengelompokan Data Penyakit Alergi Pada Anak," *Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika),* vol. 6, no. 1, p. 130, 2021, doi: https://doi.org/10.30645/jurasik.v6i1.277

[17] A. J. Wahidin and D. I. Sensuse, "Perbandingan Algoritma K-Means, X-Means Dan K-Medoids Untuk Klasterisasi Awak Kabin Lion Air," *Jurnal ICT: Information Communication & Technology*, vol. 20, no. 2, pp. 298–302, 2021,

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5545

doi: https://doi.org/10.36054/jict-ikmi.v20i2.387

[18] M. N. P. Pamulang, M. N. Aini, and U. Enri3, "Komparasi Distance Measure Pada K-Medoids Clustering untuk Pengelompokkan Penyakit ISPA," *Edumatic: Jurnal Pendidikan Informatika,* vol. 5, no. 1, pp. 99–107, 2021, doi: https://doi.org/10.29408/edumatic.v5i1.3359

[19] 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, p. 109282, Jul. 2022, doi: https://doi.org/10.1016/j.ecolind.2022.109282

[20] A. Li, C. Fan, F. Xiao, and Z. Chen, "Distance measures in building informatics: An in-depth assessment through typical tasks in building energy management," *Energy and Buildings*, vol. 258, p. 111817, 2022, doi: https://doi.org/10.1016/j.enbuild.2021.111817

[21] 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, 2020, doi: https://doi.org/10.1155/2020/8848031

[22] G X. Gao and G. Li, "A KNN Model Based on Manhattan Distance to Identify the SNARE Proteins," *IEEE Access,* vol. 8, pp. 112922–112931, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3003086

[23] 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, doi: https://doi.org/10.31326/jisa.v5i1.1097

[24] 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, doi: https://doi.org/10.35940/ijeat.e9666.069520

[25] 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, doi: https://doi.org/10.1080/08839514.2020.1723868

[26] 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, p. 102405, Jun. 2021, doi: https://doi.org/10.1016/j.jag.2021.102405

[27] T. Z. Baharav, G. M. Kamath, N. T. David, and I. Shomorony, "Spectral jaccard similarity: a new approach to estimating pairwise sequence alignments," *Patterns*, vol. 1, no. 6, 2020.

[28] 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, doi: https://doi.org/10.1093/bioinformatics/btaa956

[29] M. Ivaškevičius, "Influence of urban shape (as memory) on social capital," *Ph.D. dissertation, Kauno technologijos universitetas*, 2021.

[30] N. Li and S. Wan, "Research on Fast Compensation Algorithm for Interframe Motion of Multimedia Video Based on Manhattan Distance," *Journal of Mathematics*, 2022, doi: https://doi.org/10.1155/2022/3468475