Comparison of Distance Measurements Based on k-Numbers and Its Influence to Clustering

Deny Jollyta[1], Prihandoko[2], Dadang Priyanto[3], Alyauma Hajjah[1], Yulvia Nora Marlim[1]
[1]Institut Bisnis dan Teknologi Pelita Indonesia, Riau, Indonesia
[2]Universitas Gunadarma, Depok, Indonesia
[3]Universitas Bumigora, Mataram, Indonesia

Article Info

Article history:

Received June 16, 2023
Revised September 02, 2023
Accepted October 17, 2023

Keywords:

Distance Measurements
Clustering
Numerical Measurements
Optimal Cluster

ABSTRACT

Heuristic data requires appropriate clustering methods to avoid casting doubt on the information generated by the grouping process. Determining an optimal cluster choice from the results of grouping is still challenging. This study aimed to analyze the four numerical measurement formulas in light of the data patterns from categorical that are now accessible to give users of heuristic data recommendations for how to derive knowledge or information from the best clusters. The method used was clustering with four measurements: Euclidean, Canberra, Manhattan, and Dynamic Time Warping and Elbow approach for optimizing. The Elbow with Sum Square Error (SSE) is employed to calculate the optimal cluster. The number of test clusters ranges from k = 2 to k = 10. Student data from social media was used in testing to help students achieve higher GPAs. 300 completed questionnaires that were circulated and used to collect the data. The result of this study showed that the Manhattan Distance is the best numerical measurement with the largest SSE of 45.359 and optimal clustering at k = 5. The optimal cluster Manhattan generated was made up of students with GPAs above 3.00 and websites/vlogs used as learning tools by the mathematics and computer department. Each cluster's ability to create information can be impacted by the proximity of qualities caused by variations in the number of clusters.

Corresponding Author:

Deny Jollyta, +628127585546
Faculty of Computer Science,
Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Indonesia,
Email: deny.jollyta@lecturer.pelitaindonesia.ac.id.

How to Cite:
D. Jollyta, P. Prihandoko, D. Priyanto, A. Hajjah, and Y. Nora Marlim, ”Comparison of Distance Measurements Based on k-Numbers and Its Influence to Clustering”, *MATRIK: Jurnal Managemen,Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 1, pp. 93-102, Nov, 2023.

1. INTRODUCTION

The problem that most often arises in processing heuristic data is a group that is not appropriate due to the existence of a number of data variants [1, 2]. A clustering algorithm assigns large data to a group (cluster) with the same properties [3], and in its application, it can be combined with other algorithms [4, 5]. Often, information generated from groupings does not present group members that fit the criteria. Some certain variables also influence the results of grouping, such as data variants, number of data, types of measurements used, and the number of clusters that are best taken as a benchmark of information. Measurements and a number of clusters generally cause information doubt. This is as reason for comparing distance measurements to see conformity in data usage. The data is tested on different clusters to determine the effect on the grouping formed. This study aims to compare and analyze the four numerical measurement formulas in light of the data patterns that are now accessible to give users of heuristic data recommendations from the best clusters, namely Euclidean, Canberra, Manhattan, and Dynamic Time Warping. Strong attribute relationships can form more solid and accountable group information.

Numerical measurement techniques have been widely used to help heuristic data grouping. In the study [6], Euclidean groups three different types of data, and this measurement has trouble identifying data without normalization. In the other research [7], Euclidean points were used to measure face recognition. Euclidean can classify 85% of existing facial images but failure of face recognition on the contrary. The Euclidean distance performed less well than Mahalano and Canberra in the two tests stated above, with an average performance for a smaller sample size. It impacts the grouping outcomes because Euclidean cannot recognize data with the same criteria.

The Canberra Distance approach used link value calculation from hydrological characteristics of the catchment in the networks field [8]. In a study [6], the Silhouette Index was used to find the best clusters, and the inter-centroid grouping results for the K-Means procedure were performed in Canberra, Euclidean, and Manhattan. The findings demonstrate that Canberra produces superior clustering compared to the other two measures. Research [9] also compares Canberra's performance, and the best clusters are discovered by using SSE and the Silhouette Index. Canberra performs worse than Euclidean when measured against the K-Nearest Neighbor method [10]. In subsequent research, the Manhattan Distance is frequently employed for face recognition with a similarity level of up to 70% [11], and based on the resulting cluster, it has data that shouldn't be in that cluster. Besides that, Sunardi and friends used Manhattan calculations for human face recognition [12].

Various similarity techniques in clustering have provided solutions to various daily problems. Besides the three distance measurements, Dynamic Time Warping Distance (DTWD) also solves many problems by similarity. In general, DTWD employs time series data, while some don't. For instance, similarity issues in facial recognition detection systems [13] and other studies [14], DTWD is used to identify Convolutional Neural Network limits on air quality forecasting systems. The results of grouping the data have not considered whether the results are the best information from the resulting groupings.

Analysis is still required to determine how well the four distance measures deliver the best cluster outcomes based on research of the four distance measurements. The difference between this research and previous research is finding the optimal clusters using Elbows from categorical data for numerical measurements. Categorical data was used because clustering is not usually applied to data in the form of numbers [15]. The best clusters are obtained from the Elbow method with the Sum of Square Error (SSE) [16]. Typically, the Euclidean distance is calculated using the Elbow and SSE methods and several k tests [17, 18]. This study contributes by comparing numerical distance measurements based on the best clusters using categorical data and demonstrating how the analysis of cluster results might influence grouping results. This study demonstrated that Manhattan's performance in providing information through optimal clusters remains superior even when SSE calculations and categorical data are used. The information obtained is more targeted based on criteria. The results of this study are expected to help users acquire the greatest information from optimal cluster results by using appropriate distance measurements.

2. RESEARCH METHOD

The methodology was developed according to research needs. Preparing heuristic data as training data is the initial stage of research. The research [19] showed that training data preparation was inseparable from the Knowledge Discovery in Database (KDD) stages in data mining. This study used 300 data from an examination of 4 departments at 4 universities in Pekanbaru, which is only data worthy of participating in the training process. Moreover, innovation in the framework of research is shown in Figure 1.

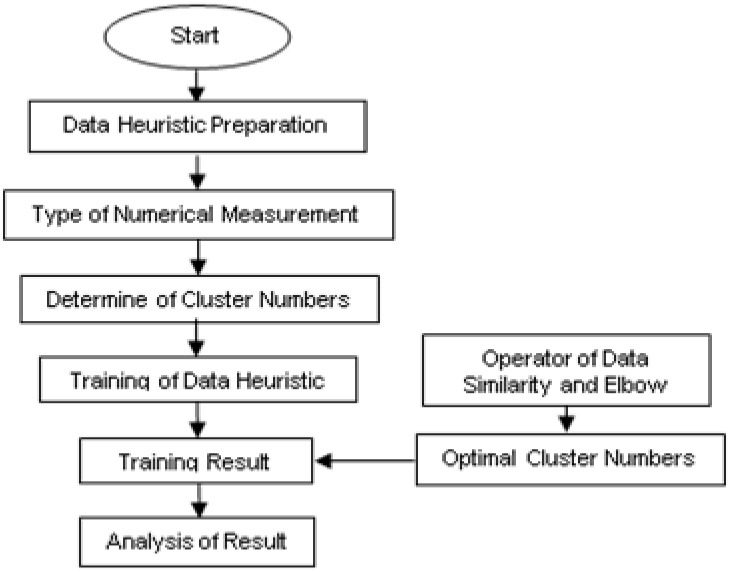
Start
Data Heuristic Preparation
Type of Numerical Measurement
Determine of Cluster Numbers
Training of Data Heuristic
Operator of Data Similarity and Elbow
Training Result
Optimal Cluster Numbers
Analysis of Result

Figure 1. Research Methodology

In more detail, data is trained using numeric distance formulas by changing the number of clusters. Those distance measurements are tested alternately using the RapidMiner application. Each change in the number of clusters is recorded as a comparison at the end of the test. The training results are analyzed to ensure that the information formed best compares with the optimal number of clusters. In some applications, clustering is referred to as data segmentation because it groups a large set of data sets by looking for data that has similarities and different characteristics in other groups [20]. Data is grouped according to their natural characteristics.

2.1. Numerical Measurements

The following discussion considers the four distance measurement techniques.

a. Euclidean Distance

Euclidean Distance is simply the distance between the points [21]. If the distance between two points is $(x1, y1)$ and $(x2, y2)$, then the calculated using equation (1). Where $d(i, j)$ is data separation from the center (i) to (j), xni is the attribute's i data on the k data, and xnj is the j data on the k data attribute.

$$d(i,j) = \sqrt{(x_{1i} - x_{1j}^2) + (x_{2i} - x_{2j})^+ \ldots + (x_{ki} - x_{kj})^2} \tag{1}$$

b. Canberra Distance

Canberra Distance is a metric function often used for distribution around an origin area. Canberra Distance examines the sum of a series of fraction differences between the coordinates of a pair of objects [11]. Canberra Distance works by dividing absolute differences between variables from two objects with the sum of absolute variables values and the formula using equation (2). Where, d is distance, yi is $(y1, y2, \ldots yn)$, and yj is $(y1, y2, \ldots yn)$.

$$d^{CAD}(i,j) = \Sigma_{k=0}^{n-1} \frac{\mid y_{i,k} - y_{j,k} \mid}{\mid y_{i,k} \mid + \mid y_{j,k} \mid} \tag{2}$$

c. Manhattan Distance

Manhattan Distance is a measurement produced based on the sum of the difference between the two objects, and the results obtained are absolute [11]. Manhattan Distance calculates distances in a perpendicular manner using equation 3. Where $d(i, j)$ is data separation from the center (i) to (j), x_{ni} is the attribute's i data on the n data, and xnj is the j data on the n data attribute.

$$d_(i,j) = \mid x_{i,1} - x_{j,1} \mid + \mid x_{i,2} - x_{j,2} \mid + \Lambda + \mid x_{i,n} - x_{j,n} \mid \tag{3}$$

d. Dynamic Time Warping Distance (DTWD)
Dynamic Time Warping Distance is mostly used for time series data. DTWD is famous for its ability to manage time distortions by realigning time series when comparing them [14]. As a similarity measure, DTWD is well-known to find the best distance calculation using equation (4) [22]. Where K is a sequence of index pairs, and (q,c) is the set of all admissible paths.

$$DTW(q,c) = min\{\sqrt{\Sigma_{k-1}^{K} W_k} \tag{4}$$

e. Optimal Cluster Method
In some applications, clustering is referred to as data segmentation because it groups a large set of data sets by looking for data that has similarities and different characteristics in other groups [3]. Data grouping is done using a predetermined algorithm, and then the data will be processed by the algorithm to be grouped according to their natural characteristics. Generally, the number of clusters is determined by comparing the results of different clusters based on variations in the number of clusters. Research [23] shows that a combination of a number of distance formulas can determine the number of clusters. Various approaches, such as Elbow, are applied to get the optimal number of clusters. In this paper, determining the optimal number of clusters for each numerical measurement uses a similarity data operator in the RapidMiner application by adopting the Elbow method to then calculate the Sum of Square Error (SSE) with Equation 5 [24]. Where xi is the attribute value of data to i, and ck is the attribute value of the cluster center point to i.

$$SSE = \Sigma_{K=1}^{K} \Sigma_{x_i e S_K} \mid\mid x_i - c_k \mid\mid_2^2 \tag{5}$$

3. RESULT AND ANALYSIS

3.1. Data Training

Training data is heuristic data obtained through surveys. This study uses a database of students who utilize a number of social media as learning media in lectures, followed by the acquisition of a student's GPA as a result of the lecture. Table 1 shows the data in question.

Table 1. Data of Student Social Media

No	Name	Department	Media	Activity	GPA
1	A1	Computer	Vlog	Practice	3.20
2	A2	Accounting	Website	Theory	3.56
3	A3	Computer	Youtube	Practice	3.40
4	A4	Computer	Website	Practice	3.56
5	A5	Hospitality	Website	Practice	3.30
6	A6	Mathematics	Vlog	Practice	2.95
...					
300	A300	Computer	Website	Practice	3.29

Table 1 consists of four attributes that will be trained: department, media, activity, and GPA. Data were collected from five departments from four universities in Pekanbaru, which were taken randomly. Computer, mathematics, accounting, management, and hospitality departments are arranged as a questionnaire. 50 student data represent each department. The initial data attribute contains non-numeric data or categorical data except GPA, so converting data to numeric categorical is necessary. In addition, the department consists of four instances, media consists of five instances, activity consists of two instances, and the GPA is divided into five instances. The conversion is shown in Tables 2, Table 3, Table 4, and Table 5.

Table 2. Department Conversion

Department	Conversion
Computer	1
Mathematics	2
Accounting	3
Management	4
Hospitality	5

Table 3. Media Conversion

Media	Conversion
Vlog	1
Website	2
Youtube	3
Instagram	4
Twitter	5

Table 4. Activity Conversion

Activity	Conversion
Theory	1
Practice	2

Table 5. GPA Conversion

GPA	Conversion
0.00 1.50	1
1.51 2.00	2
2.01 2.50	3
2.51 3.00	4
3.01 4.00	5

3.2. Comparison of Distance Analysis

The test used the same amount of data but for four different numerical measurements and 9 times the change in the number of clusters. The initial number of clusters is 2, followed by 3 to 10. Using Equations (1), (2), (3), and (4), a good cluster is a cluster that has high homogeneity between members in one cluster (within the cluster) and high heterogeneity between clusters (between clusters). Likewise, if the number of clusters is formed more and more, the distance between clusters will be smaller, as shown in Table 6.

Table 6. Average Within Cluster Distance

No. Cluster	Euclidean Distance	Canberra Distance	Manhattan Distance	Dynamic Time Warping Distance
2	183.001	72.311	312.422	79.89
3	147.019	46.009	213.3	65.816
4	107.700	27.199	137.047	52.002
5	71.857	19.023	106.153	41.844
6	60.922	15.102	96.513	37.3
7	50.545	11.826	77.287	34.87
8	46.558	9.724	68.46	29.869
9	38.910	6.971	48.453	34.167
10	34.724	5.414	34.827	30.553

The distance between clusters, as shown in Table 6, is caused by sharing data into more groups so that distance measurements tend to form members with even smaller distances. The results of training using RapidMiner show varying distances between clusters. The smaller the cluster member's distance from the cluster center point, the more cluster members. For example, Table 7 displays the number of members in each cluster for distance measurements using Euclidean Distance.

Table 7. Member's Distance of Euclidean Distance

Numerical Measurement	No Cluster	Distance within Cluster	Amount of Cluster Members
Euclidean Distance	2	183.001	cluster_0 (207), cluster_1 (93)
	3	147.019	cluster_2 (120), cluster_0 (105), cluster_1 (75)
	4	107.700	cluster_2 (110), cluster_3 (43), cluster_1 (72), cluster_0 (75)
	5	71.857	cluster_4 (67), cluster_3 (43), cluster_1 (72), cluster_2 (60), cluster_0 (58)
	6	60.922	cluster_3 (67), cluster_5 (57), cluster_0 (72), cluster_2 (46), cluster_1 (36), cluster_4 (22)
	7	50.545	cluster_0 (51), cluster_6 (53), cluster_5 (29), cluster_2 (72), cluster_3 (44), cluster_4 (32), cluster_1 (19)
	8	46.558	cluster_0 (71), cluster_4 (33), cluster_7 (23), cluster_5 (33), cluster_3 (40), cluster_2 (59), cluster_6 (26), cluster_1 (15)
	9	38.910	cluster_8 (19), cluster_0 (46), cluster_5 (48), cluster_2 (59), cluster_7 (32), cluster_3 (38), cluster_4 (34), cluster_6 (19), cluster_1 (5)
	10	34.724	cluster_9 (13), cluster_0 (56), cluster_5 (51), cluster_4 (33), cluster_7 (27), cluster_8 (38), cluster_3 (36), cluster_2 (29), cluster_1 (7), cluster_6 (10)

Each cluster has a different distance. The more clusters you want to set up, the smaller the distance between clusters. Therefore, the homogeneity between members in the cluster produces varied grouping information. For example the measurement of Euclidean Distance in the number of clusters 6 produces information in Table 8.

Table 8. Euclidean Distance Cluster Results in k=6

Cluster	Amount of Cluster Members	Information
3	67	This cluster contains 85% of Computer students who use media websites and YouTube to support learning. References and tutorials are needed to understand the material. The rest use vlogs for practice with the acquisition of GPA> 3.
5	57	This cluster brings the distance between vlog media and website attributes Accounting and Mathematics students use. 61% of the cluster members use the website to obtain articles from the best journals. These two media are more widely used for theory with the acquisition of GPA> 3. However, there are still students with a GPA of 2.59
0	72	Most of the students of Management and Hospitality use vlogs and website media in lecture activities. They used both media to make material demonstrations through references. Students who enter this group have GPA> 3
2	46	Mathematics students dominate this group. Almost 90% use YouTube media, in theory. The use of YouTube is more about understanding how to apply the formula. Others are Accounting students who all use YouTube for theory with a GPA> 3.
1	36	Students of Management and Hospitality use YouTube to understand the material. The use of this media is balanced for theory and practice with the average GPA> 3
4	22	This cluster contains Accounting, Management, and Hospitality students who use Twitter and instagram to support lectures. In addition, Instagram is used by Hospitality students to showcase and sell the results of activities such as cooking and stitching. The average student in this cluster has a GPA> 3.

3.3. The Optimal Cluster and Its Influence on Clustering

Based on clustering that resulted from increasing the number of clusters, it is necessary to determine the most optimal number of clusters to represent the desired grouping. This study obtained it from training using the RapidMiner application by utilizing data similarity operators and Elbow methods in the SSE formula (5) to obtain the largest distance difference. The use of this data similarity operator is adjusted to the distance measurement technique used. Difference is taken from distance data between clusters found in Table 6. Because the difference results start from the second cluster number, where k = 2, the data shown starts from k = 3, and the optimal number of clusters from each measurement can be seen in Table 9.

Table 9. The Optimal Cluster of Numerical Measurements

No. Cluster	Euclidean	Canberra	Manhattan	Dynamic Time Warping
k=2 (Initiation)	0	0	0	0
k=3	35.982	26.302	99.122	14.074
k=4	39.319	18.81	76.253	13.814
k=5	35.843	**8.176**	**30.894**	10.158
k=6	**10.935**	3.921	9.640	**4.544**
k=7	10.377	3.276	19.226	2.430
k=8	3.987	2.102	8.827	5.001
k=9	7.648	2.753	20.007	4.298
k=10	4.186	1.557	13.626	3.614

Using Table 9, the best cluster is determined by the greatest SSE value after comparing the variances of each cluster. For instance, the SSE value for Euclidean is 3.337 for k=3 and k=4. For Euclidean at k=4 and k=5, the SSE value is 3.476. as well as. The optimal k is determined to be the one with the highest SSE value. An optimal depiction of cluster positions for each measurement is shown in the following Figure 2.

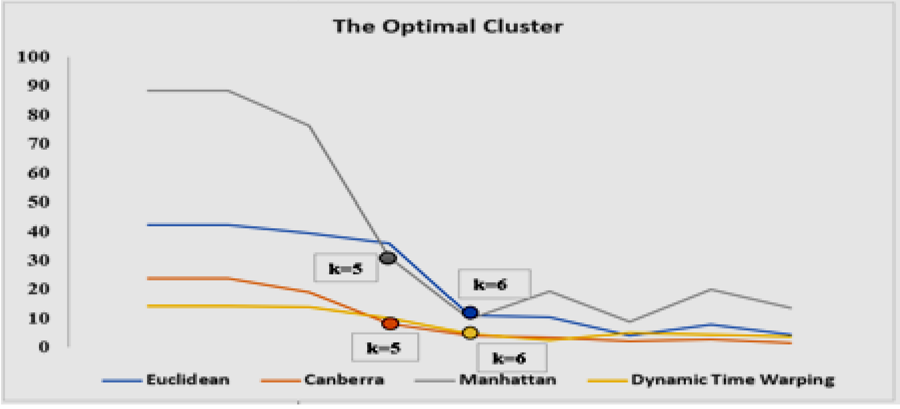

Figure 2. The optimal cluster

Determination of the optimal number of clusters in Figure 2 forms the basis of grouping data. The relationship between members of the cluster would produce different information. The more the number of clusters, the fewer the number of members, and the resulting information will be different. Based on the information above, the best measurement is Manhattan. Manhattan distance has the biggest SSE at k=5. According to Table 9, SSE k=5 is derived from 76.253-30.894=45.359. Meaning that its cluster is better than others, and this SSE is the biggest from other measurements. The best information on Manhattan from k=5 can be shown in Table 10.

Table 10. Manhattan Distance Cluster Result in k=5

Cluster	Amount of Cluster Members	Information
0	88	This cluster members are grouped based on GPA. All the members have a GPA $>$ 3.00 and use vlog media, YouTube, and websites to understand lecture material more.
1	43	More than 50% of members used Instagram, Twitter, and YouTube for the proximity of the lectures in theory.
2	67	The cluster consists of Management and Hospitality department students. They used vlogs, websites, and YouTube to make references and tutorials. More than 97% of students have GPA $>$ 3.00
3	37	21.6% of students in this cluster have a GPA $<$ 3.00 and are from the Mathematics department. Using YouTube is not recommended.
4	65	Theory activity used vlogs and website media to support Computer department students' learning. All computer concepts are described clearly, and students have an average GPA $>$ 3.

Generally, the results of training data with a greater number of k produce better information or knowledge. Information users tend to feel that the grouping formed does not align with expectations. With the greater number of k, the grouping of data is increasingly clear and represents the wishes of the information user. This applies to all numerical measurements tested. The test used the same amount of data but for four different numerical measurements and 9 times the change in the number of clusters. The initial number of clusters is 2, followed by 3 to 10. Using Equations (1), (2), (3), and (4), a good cluster is a cluster that has high homogeneity between members in one cluster (within the cluster) and high heterogeneity between clusters (between clusters). Likewise, if the number of clusters is formed more and more, the distance between clusters will be smaller, as shown in Table 6. Overall, the fourth comparison of the distance formula gives different information. Comparison also provides more knowledge for data processors in using the appropriate distance formula, as shown in Table 11.

Table 11. Comparison Result of Numerical Distances

Indicator	Euclidean Distance	Canberra Distance	Manhattan Distance	Dynamic Time Warping Distance
Within Cluster Distance	183.001	72.311	312.422	79.89
Biggest Elbow	k=5 to k=6 is 24.908	k=4 to k=5 is 10.634	k=4 to k=5 is 45.359	k=5 to k=6 is 5.614
Optimal Cluster	6	5	5	6
Time	Quick Response (0.02)	Quick Response (0.01)	Quick Response (0.01)	Slow Response (0.06)

Table 11 shows that Manhattan Distance has the biggest Elbow than others. These results are obtained by calculating the SSE that forms the farthest Elbow. In addition, Manhattan processed data in under 0.01 seconds according to the test time indicator. This study demonstrates that the four distance measurements achieve the optimum membership utilizing Elbow and SSE for ideal clusters. This differs from previous studies such as [6] and [25] that used Silhouette and DBI to obtain the best clusters. Each measurement produces the most optimal cluster member. An optimal cluster's members exhibit strong connections and are connected to one another, resulting in information that is particularly useful to interested parties. For instance, social media is the most beneficial to student learning. So lecturers can disseminate lecture materials using social media.

4. CONCLUSION

Using numerical measurements in heuristic data can form the desired group of data easily. The comparison results of this distance formula show that Manhattan Distance has the biggest Elbow of 45.359 with the optimal cluster at k = 5. This indicates that of the three, Manhattan is the best numerical measurement. The findings of this study also demonstrate that accurate information can be derived from numerical measurement computations using categorical data. The variants and quantities of heuristic data can create information in more detail in training with different amounts of k. The difference in the number of clusters was quite influential on the grouping results, namely information that is more detailed and following users' needs of information caused by the closeness of the relationship between attributes. However, the optimal number of clusters needs to be determined because the increase in the cluster numbers during training creates doubts about the information produced. Determination of the optimal number of clusters can direct information users to the certainty of grouping for different data. Regarding raising student GPAs, information created by the Manhattan optimal cluster can persuade users that social media vlogs and websites truly benefit students with exact sciences. This can make it easier for universities or communities to direct the right use of social media to aid student learning. In the next evaluation, it is necessary to note the linkages of attributes from heuristic data that are used because it can ignore the correlation between attributes. This certainly can change the group members formed and the information produced.

5. ACKNOWLEDGEMENTS

The Acknowledgment is addressed to Institut Bisnis dan Teknologi Pelita Indonesia, Universitas Gunadarma, and Universitas Bumigora, which supports the completion and publication of this research.

REFERENCES

[1] R. Suwanda, Z. Syahputra, and E. M. Zamzami, "Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K," in *Journal of Physics: Conference Series*, vol. 1566, no. 1, 2020, p. 7.

[2] Sapriadi, Sutarman, and E. B. Nababan, "Improvement of K-Means Performance Using a Combination of Principal Component Analysis and Rapid Centroid Estimation," in *Journal of Physics: Conference Series*, vol. 1230, no. 1, 2019, p. 8.

[3] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Computer Science*, vol. 2, no. 5, pp. 1–22, 2021.

[4] 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. -, pp. 1–9, 2020.

[5] M. Zubair, M. A. Iqbal, A. Shil, M. J. Chowdhury, M. A. Moni, and I. H. Sarker, "An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling," *Annals of Data Science*, vol. June, no. June, pp. 23–25, 2022.

[6] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," in *Journal of Physics: Conference Series*, vol. 1566, no. 1, 2020, p. 8.

[7] H. Wu, Y. Cao, H. Wei, and Z. Tian, "Face Recognition Based on Haar like and Euclidean Distance," in *Journal of Physics: Conference Series*, vol. 1813, no. 1, 2021, pp. 2–8.

[8] P. Istalkar, S. L. Unnithan, B. Biswal, and B. Sivakumar, "A Canberra distance-based complex network classification framework using lumped catchment characteristics," *Stochastic Environmental Research and Risk Assessment*, vol. 35, no. 6, pp. 1293–1300, 2021.

[9] M. Raeisi and A. B. Sesay, "A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 10, no. August, pp. 86 286–86 297, 2022.

[10] K.-n. Neighbor, A. F. Pulungan, M. Zarlis, and S. Suwilo, "Performance Analysis of Distance Measures in K-Nearest Neighbor," in *ICMASES 2019*, 2020, p. 9.

[11] A. Fadlil and N. Tristanti, "Comparative Analysis of Euclidean , Manhattan , Canberra , and Squared Chord Methods in Face Recognition," vol. 37, no. 3, pp. 593–599, 2023.

[12] Sunardi, Abdul Fadlil, and Novi Tristanti, "The Application of The Manhattan Method to Human Face Recognition," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 6, pp. 939–944, 2022.

[13] D. Deriso and S. Boyd, "A general optimization framework for dynamic time warping," *Optimization and Engineering*, vol. June, no. 0123456789, p. 22, 2022.

[14] E. Eslami, Y. Choi, Y. Lops, A. Sayeed, and A. K. Salman, "Using wavelet transform and dynamic time warping to identify the limitations of the CNN model as an air quality forecasting system," *Geoscientific Model Development*, vol. 13, no. December, pp. 6237–6251, 2020.

[15] P. Lippe and E. Gavves, "L Atent N Ormalizing F Lows for," in *conference paper at ICLR 2021*, no. -, 2021, p. 27.

[16] C. Guyeux, S. Chrétien, G. B. Tayeh, and J. Demerjian, "Introducing and Comparing Recent Clustering Methods for Massive Data Management in the Internet of Things," *Journal of Sensor and Actuator Network*, vol. 8, no. 56, pp. 1–25, 2019.

[17] R. Bond and P. Biglarbeigi, "Data-driven versus a domain-led approach to k-means clustering on an open heart failure dataset," *International Journal of Data Science and Analytics*, vol. 15, no. 1, pp. 49–66, 2023.

[18] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," *Accounting, Auditing and Finance*, vol. 2020, no. 1, pp. 5–8, 2020.

[19] M. A. Jassim and S. N. Abdulwahid, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," in *IOP Conference Series: Materials Science and Engineering*, vol. 1090, no. 1, 2021, p. 012053.

[20] J. Han and M. Kamber, *Data Mining: Concepts and Techniques (2nd edition)*, 2006, vol. 54, no. Second Edition.

[21] M.-f. O.-d. Algorithm, R. Laher, A. Grant, F. Fang, W. Chen, Z. Tian, L. Zhang, and Y. Yang, "An outlier detection algorithm based on maximum and minimum distance," in *ICEECT*, 2021, p. 6.

[22] H. S. Lee, "Application of dynamic time warping algorithm for pattern similarity of gait," *Journal of Exercise Rehabilitation*, vol. 15, no. 4, pp. 526–530, 2019.

[23] D. Bertsimas, A. Orfanoudaki, and H. Wiberg, *Interpretable clustering : an optimization approach.* Springer US, 2021, vol. 110, no. 1.

[24] R. D. Dana, D. Soilihudin, and R. D. Priyatna, “Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method,” in *1st International Conference of SNIKOM 2018*, 2019, p. 7.

[25] S. Gultom, S. Sriadhi, M. Martiano, and J. Simarmata, “Comparison analysis of K-Means and K-Medoid with Ecluidience Distance Algorithm, Chanberra Distance, and Chebyshev Distance for Big Data Clustering,” *IOP Conference Series: Materials Science and Engineering*, vol. 420, no. 1, p. 8, 2018.