The Impact of Similarity Functions on Divisive Analysis Clustering of Tuberculosis Disease

Yulita Molliq Rangkuti[1]*, Mansur AS[2], Agus Junaidi[3], Nurul Maulida Surbakti[1], Rizky Gunawan[4]

yulitamolliq@unimed.ac.id[1*]

Universitas Negeri Medan, Indonesia[1*, 1, 2, 3, 4]

Abstract. Tuberculosis is a bacterial infection that usually affects the lungs but can also affect other parts of the body. It is contagious and spreads through the air when an infected person coughs, sneezes or sings. North Sumatra ranks as one of the three Indonesian provinces that experience the highest rates of both occurrence and death. Tracking Instances of Tuberculosis is crucial for managing and averting the spread of the illness. The Divisive Analysis (DIANA) algorithm is frequently utilized to categorize Tuberculosis cases. DIANA operates as a clustering algorithm that organizes items into sets based on their similarities. The study emphasizes evaluating the effectiveness of various similarity functions. The dataset comprises factors such as mortality rates, infection rates, and recovery rates sourced from the North Sumatra Provincial Health Office and the Central Statistics Agency (BPS). The findings indicated the emergence of four clusters within North Sumatra Province. Furthermore, an assessment was performed employing the Davies Bouldin Index (DBI) to assess the quality of clustering. By comparing various distance metrics (Bray Curtis distance, Chebyshev distance, and Canberra distance), the lowest DBI score was reached with Chebyshev distance, yielding a value of 0.5121. This value reflects a satisfactory level of cluster quality. As a result, the study aids in visualizing the distribution of Tuberculosis cases in North Sumatra Province and provides a basis for data-driven decision-making in addressing the disease.

Keywords: Tuberculosis; DIANA; Davies Bouldin Index; Similarity Function.

1. INTRODUCTION

Tuberculosis is one of the severe public health problems to consider because it is categorized as being able to spread and be contagious [1]. Tuberculosis continues to be a significant threat in the world of health, especially in developing and developed countries. It has also not been eradicated optimally.

Along with the significant increase in the number of patients at RSU. Aisyiyah Ponorogo, data obtained from medical records can represent the population of patient data in North Sumatra Province. From these data, it is hoped that the Provincial Health Office can discover the spread of the disease, especially tuberculosis. Thus, the Provincial Health Office can provide input to the government through the Health Office regarding prevention efforts and health education related to tuberculosis in the regions.

Currently, the pile of patient data in the Provincial Health Office generally only describes patient graphs and statistics, treatment costs, and disease data. Data mining should be implemented from existing medical record data to extract new knowledge or determine specific

ICIESC 2025, September 16, Medan, Indonesia

DOI 10.4108/eai.16-9-2025.2361147

patterns that can be used for analysis, evaluation, or description of certain conditions of patients undergoing treatment. Data mining uses artificial intelligence, statistics, mathematics, and machine learning techniques to extract and identify beneficial information and related knowledge from various large databases [2,3]. With the correct patient medical record data processing technique, the pattern of TB disease spread in the North Sumatra Provincial Health Office can be determined. Meanwhile, the processing of patient medical record data in the North Sumatra Provincial Health Office has not been able to predict the spread of TB disease because it still uses manual methods and spreadsheets. The existing data cannot be validated for measurement. For this reason, this study will process medical record data using data mining with the Divisive analysis (DIANA) Clustering method.

Clustering is a data mining method that uses unsupervised learning to group documents based on their similarities [4]. Devisive Analysis (DIANA) is among the prominent algorithms used for grouping different instances. The approach of the DIANA algorithm is hierarchical and begins by putting every item into a single cluster known as the hierarchy's root, subsequently breaking down that root cluster into multiple smaller [5,6].

Numerous earlier researches have examined how similarity functions affect clustering methods in various scenarios. One such investigation explored the use of Euclidean and Manhattan distances as similarity measures within DIANA for clustering related to the National Socio-Economic Survey (SUSENAS) [6]. The algorithm's use of Euclidean distance resulted in a robust cluster with an MSE of 182.0976, whereas the application of Manhattan distance led to a strong cluster with an MSE of 193.0648. In separate investigation conducted by [8], multiple distance measures were evaluated within PAM algorithm, including Euclidean, Manhattan, Minkowski, and Chebyshev distances, specifically regarding clustering for Dengue Fever. The finding revealed that Chebyshev distance had the highest Silhouette index, making it superior to other distances. A different analysis by [9] examined how different similarity measures affected the K-Medoids clustering algorithm in the context of educational clustering.

Diana's clustering method was started by Kumarahadi et al. (2023) [10], who combined districts/cities in Indonesia based on poverty indicators with validation values using cophenetic and Silhouette coefficient values of 0.90 and 0.71, respectively. Then, Izzuddin and Wijayanto (2024) in [10] analyzed clusters using partitioning clustering methods, namely k-means and Partitioning Around Medoids (PAM), and hierarchical clustering methods, namely Ward and Divisive Analysis (DIANA) can be used in grouping provinces in Indonesia. Based on the model evaluation, the best cluster model was obtained using the ward approach with Principal Component Analysis (PCA) analysis. Suyono et al. (2024) in [12] identified distinct clusters based on stunting prevalence across Indonesian provinces. In one analysis, two clusters emerged: Cluster 1 comprised 32 provinces with low stunting rates, while Cluster 2 included 2 provinces with high prevalence. A separate classification revealed three clusters: 26 provinces with moderate stunting (Cluster 1), 6 with low prevalence (Cluster 2), and 2 with high rates (Cluster 3). A more detailed grouping further divided provinces into four clusters: 21 with moderate stunting (Cluster 1), 5 with low prevalence (Cluster 2), 6 with high rates (Cluster 3), and 2 with extremely high prevalence (Cluster 4). The current DIANA method study employs Euclidean distance from a variety of distance formulas, including Chebyshev, Canberra, and Bray Curtis Distance. This study will alter the Diana approach by replacing the Chebyshev distance formula with the Canberra and Bray Curtis distances to produce more accurate mapping.

2. RESEARCH METOD

This part outlines the framework of the research. Figure 2 illustrates a method for executing DIANA. Three essential steps are required for optimal Tuberculosis mapping in North Sumatra: (1) data preprocessing, (2) DIANA clustering analysis, and (3) evaluation. The figure illustrates three distinct distance measures used in this process: *Chebyshev, Canberra, and Bray Curtis* distances. Validation of DIANA using DBI.

A. Row Data

The information is further backed by comments from a representative of the North Sumatra Provincial Health Office staff member during a discussion. She mentioned that "recently, Tuberculosis Disease has been in the spotlight and has become very worrisome." Furthermore, she clarified that the current approach to the charting regions based on dengue During a discussion, a North Sumatra Provincial Health Office representative emphasized these concerns, stating that 'Tuberculosis has recently gained significant attention and become a serious public health concern.' The official further explained that the province's current dengue mapping system remains manual. This approach tallies infection numbers without providing risk-based zone classifications, limiting the ability to identify high-risk or vulnerable areas for dengue transmission.

Consequently, research is needed to assess the spatial distribution of dengue fever cases across all districts and cities in North Sumatra Province. The study relies on secondary data collected from the North Sumatra Provincial Health Office and the Central Statistics Agency (BPS) spanning 2020 to 2023, incorporating four key variables for analysis.

Table 1. Row data

No	Regency/City	Tuberculosis Case	Recovery	Number of mortalities	Population
1	Medan City	13001	1864	530	2,460,858
2	Pematang Siantar City	883	509	42	270,768
3	Binjai City	798	225	40	295,361
4	Tanjung Balai City	702	113	6	177,640
5	Nias Barat regency	126	0	9	90,585
6	Gunung Sitoli City	432	19	6	136,707

Table 1 presents data on Tuberculosis Disease cases in 33 region or urban areas. This Informatipn was sourced from the North Sumatra Provincial Health Office and spans from January 2020 to Desember 2023. North Sumatra is the focus of attention regarding tuberculosis. Data from the Ministry of Health in 2021 said North Sumatra was ranked sixth among provinces in Indonesia in terms of Tubuerculosisi case [9].

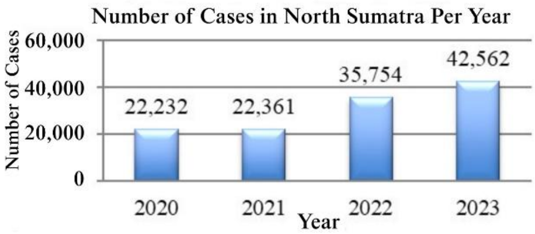

Figure 1. Number of Tuberculosis Cases per Year in North Sumatra Source: North Sumatra Health Service (2024)

Based on the interview with Mr. Japirman Purba, it was revealed that cases of Tuberculosis in 2021 reached 22,361 people. In 2022, there was a significant increase in the number of tuberculosis cases, reaching 35,754 people, an increase of 13,393 from the previous year. Furthermore, in 2023, the cases continued to increase to 42,562 people. The increase in the number of Tuberculosis cases as a whole is the motivation to carry out this research.

Data Preprocessing

In this study, several stages of data preprocessing are carried out: data cleaning, selection, and transformation.

Cleaning Data

This study's data consisted of TB Case Rates, TB Cure Rates, and TB Death Rates. Then, the data cleaning process was carried out because there was empty data. Following up on this, the researcher carried out the data cleaning process on the data by changing each empty data with the median data value [13].

Selection Data

Next, the researcher conducted a statistical analysis using Pearson correlation. Pearson correlation is the most common method used to calculate the correlation between variables. In scikit-learn, Pearson correlation is done by calling the corr() function. The results will be in the form of a correlation matrix as follows

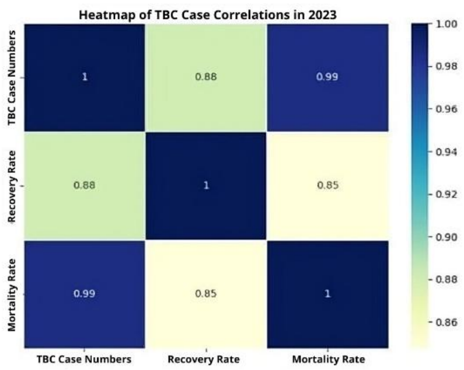

Figure 2. Correlation heatmap graph of TB data variables

Figure 2. The analysis of TB case data reveals a significant positive correlation between TB Case Rate and Recovered Rate, with a value of 0.88, and between TB Case Rate and Mortality Rate, which reaches 0.99. This finding indicates that the increase in TB cases is closely related to the increase in mortality. In addition, there is also a positive correlation of 0.85 between the Recovered Rate and Mortality Rate, indicating that when the recovery rate increases, the mortality rate also tends to increase, possibly due to the high number of cases. The heatmap shows the strength of the relationship between variables, where darker colors indicate a stronger correlation.

B. DIANA Algoritma

This stage creates clustering with the DIANA Algorithm. It selects a district or city as its object and attributes, namely the Positive Number or Number of TB Cases, TB Death Rate, and TB Cure Rate.

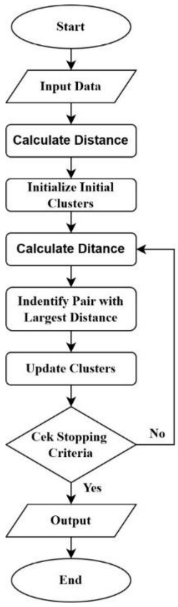

Figure 3. DIANA Flowchart

Simar et al. (2020) in [14] states that the stages of the DIANA algorithm are as follows.

1. State's all data in one cluster, namely $c = 1$ and the number of splinter groups $l = 0$.
2. Calculate the Euclidean distance between objects. Select objects with the highest average dissimilarity or the farthest average distance from other objects.
3. The number of splinter groups increases by one, namely $l = l + 1$.
4. The object selected in step 3 separates from the cluster and forms a new cluster or splinter group to l, namely sgl.
5. The number of clusters increases by one to $c = c + 1$.
6. Calculate the difference for each object i outside the splinter group. If the largest $diff(i)$ is positive, then the object i goes into sgl and repeats step 7. If negative, proceed to step 10.
7. Conducting a test of the number of clusters, namely if $c = n$, the iteration for the polythetic divisive method is complete. If $c \neq n$, then proceed to step 10.

8. Choosing a cluster with the largest diameter or highest dissimilarity between objects.
9. Then, the cluster will be separated by repeating steps 3-10.

The researcher will experiment with the number of clusters by evaluating the cluster results. The evaluation will be carried out using the Davies Bouldin Index (DBI).

C. Davies Bouldin Index (DBI)

DIANA clustering performance using the Davies Bouldin Index (DBI) as the evaluation method has advantages in measuring clustering evaluation because it can produce cohesion and separation values. The DBI is a metric used to measure clustering quality in data analysis. The goal is to measure how well the clusters generated by the clustering algorithm separate different data groups and approach their cluster centers. The lower the DBI value, the better the clustering produced. The DBI formula for calculating clustering quality is as follows.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max\left[\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i . X_j)}\right] i \neq j. \qquad (1)$$

where $\delta(X_i . Y_j)$ defines the distance between clusters X_i and X_j and $\Delta(X_j)$represents the intra-cluster distance (diameter) of cluster $\Delta(X_j)$ and k is the number of clusters of the partition U. In cluster analysis, lower index values indicate higher quality clusters - specifically, those exhibiting tight internal cohesion and well-separated centroids. The optimal cluster count (k) is determined by selecting the configuration that yields the minimal Davies-Bouldin Index (DBI) value [15]. For this study, clustering performance was assessed using DBI values computed through the DIANA algorithm.

D. Similarity Function

Initially developed by G. N. Lance and W. T. Williams between 1966 and 1967, the Canberra Distance measures the distance between two points in vector space when working with raw data. Vashista and Nagar (2017) in [16] noted that this distance measure yields binary outcomes, returning either TRUE or FALSE values [15]

$$d_{can}(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i + y_i|} \qquad (2)$$

where $d_{can}(x, y)$ is Canberra Distance between the data and the centroid., n is the amount of data, x_i is the ith data in the cluster, and y_i is the ith centroid point [8]. Next, The Bray-Curtis Distance, or the Sørensen Distance, represents a similarity metric between samples. According to Bishnoi and Hooda (2020) in [17], when applied to normalized variables with an additive object variable, this distance measure transforms into a modified version of the City Block (Manhattan) Distance. [17].

$$d_{BC}(x, y) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} x_i + y_i} \qquad (3)$$

where $d_{BC}(x, y)$ presents Bray Curtis Distance. While, Chebyshev distance is a technique used to assess the separation between two locations by evaluating the absolute value of the difference in their coordinates. This measure can also be viewed as the greatest distance between any two points, identified by considering the largest difference in values between the x and y coordinates involved. Equation (8) calculates the Chebyshev distance between two objects.

$$d_{che}(x, y) = \max_{i=1}\{|x_i - y_i|\} \qquad (4)$$

where $d_{che}(x, y)$ presents is the Chebyshev distance.

3. RESULTS AND DISCUSSIONS

E. Divisive Analysis (DIANA)

This study employs data mining through DIANA clustering methodology to achieve spatial mapping of objects sharing similar data attributes and characteristics. The research methodology involves:

1. Cluster Configuration: Implementation of the DIANA algorithm with a fixed 4-cluster structure for each annual dataset
2. Distance Metric: Utilization of Euclidean distance as the primary dissimilarity measure
3. Validation Approach: Application of the Davies-Bouldin Index (DBI) for cluster quality assessment
4. Temporal Analysis: Comparative evaluation across four consecutive years (2020-2023)

The DBI validation results for each annual cluster analysis are presented as follows:.

Table 2. DBI Evaluation Results

Function	DBI value
Chebyshev Distance	0.5121
Canberra Distance	0.9809
Bray Curtis Distance	0.8179

Table 2 shows the results of clustering evaluation with DBI for three distances, namely Chebyshev, Canberra, and Bray Curtis distances. DBI measures the quality of clustering, where lower values indicate better clustering results. From Table 2, the lowest DBI value (0.5121) indicates the best clustering quality using Chebyshev distance in 2020-2023. Overall, a DBI closer to zero reflects more optimal clustering results. Thus, DIANA Clustering is good with Chebyshev distance.

Next, calculations using the DIANA method based on the Chebyshev distance were carried out, providing the results shown in Table 3 below.

Table 3. Clustering Results

No	Regency/City	Canberra distance	Bray Curtis distance	Chebyshev distance
1	Asahan Regency	3	4	3
2	Batu Bara Regency	4	4	4
3	Dairi Regency	4	4	4
4	Deli Serdang Regency	2	1	2
5	Humbang Hasundutan Regency	4	2	4
6	Karo Regency	4	4	4
7	Labuhan Batu Regency	4	4	4
8	Labuhan Regency Batu Selatan	4	4	4
9	Labuhan Batu Utara Regency	3	4	4
10	Langkat Regency	3	1	3

11	Mandailing Natal Regency	3	4	3
12	Nias Regency	4	2	4
13	Nias Barat Regency	4	2	4
14	Nias Selatan Regency	4	3	4
15	Nias Utara Regency	4	2	4
16	Padang Lawas Regency	4	4	4
17	Padang Lawas Utara Regency	4	4	4
18	Pakpak Bharat Regency	4	2	4
19	Samosir Regency	4	2	4
20	Serdang Bedagai Regency	3	4	3
21	Simalungun Regency	4	1	3
22	Tapanuli Selatan Regency	4	4	4
23	Tapanuli Tengah Regency	3	4	4
24	Tapanuli Utara Regency	3	4	4
25	Toba Samosir Regency	4	4	4
26	Binjai City	3	4	3
27	Gunungsitoli City	4	3	4
28	Medan City	1	1	1
29	Padangsidimpuan City	4	4	4
30	Pematang Siantar City	3	4	4
31	Sibolga City	4	3	4
32	Tanjung Balai City	4	4	4
33	Tebing Tinggi City	4	4	4

[a] Health Department of North Sumatra

Table 3 shows the distribution of cluster numbers for 33 provinces, by providing colors to mark the clusters. Cluster placement is shown in the form of color. From [18] the cluster division is shown in color. This can be seen from Table 3. Table 3 shows the best regional clustering using the DIANA method with Chebyshev distance. The determination of the best cluster has been determined with DBI, which is 0.5121. From table 3 it can be seen that the green cluster consists of the regencies of West Nias, Dairi, Pakpak Bharat, South Labuhan Batu, Karo, North Padang Lawas, Padang Lawas, South Nias, Batubara, Samosir, Humbang Hasundutan, North Tapanuli, Nias, Central Tapanuli, North Nias, Labuhan Batu, North Labuhan Batu, South Tapanuli, Toba Samosir. and the city of Tebing Tinggi, Tanjung Balai, Padangsidimpuan, Pematang Siantar, Gunung Sitoli, Sibolga. The orange cluster includes South Tapanuli and Mandailing Natal. The yellow cluster consist of the regencies Simalungun, Langkat, Mandailing Natal, Serdang Bedagai, Asahan, and City of Binjai. The orange cluster consist of the Deli Serdang regencies. Finally, the orange cluster consist of Medan City.

4. CONCLUSIONS

The DIANA algorithm is utilized to the chart Tuberculosis cases various districts and cities within North Sumatra Province. There different experiments were performed, using multiple similarity metrics including Chebysev, Bray-curtis, and Canberra. All trials concurred that four was the optimal number of clusters. According to the outcomes assessed with the DBI, Chebyshev Distances emerged as the most effective distance metric, hence this research adopts Chebysev Distance for the distance computations in the DIANA algorithm. This is a new discovery that from the research of Kumarahadi et al. (2023) in [10,11,12] suggest the Euclidean distance as ge preferable choice. This count serves as a basis for area mapping through the color coding of the resulting clusters. The green cluster indicates region with a low incidence of tuberculosis cases; the yellow cluster signifies a moderate level; the orange cluster represents high levels; while the red cluster highlights an extremely high incidence. The green cluster is composed of the regencies of West Nias, Dairi, Pakpak Bharat, South Labuhan Batu, Karo, North Padang Lawas, Padang Lawas, South Nias, Batubara, Samosir, Humbang Hasundutan, North Tapanuli, Nias, Central Tapanuli, North Nias, Labuhan Batu, North Labuhan Batu, South Tapanuli, Toba Samosir. and the city of Tebing Tinggi, Tanjung Balai, Padangsidimpuan, Pematang Siantar, Gunung Sitoli, Sibolga. The orange cluster includes South Tapanuli and Mandailing Natal. In accordance with DBI, the Euclidean distance utilized in the DIANA algorithm forms adequate clusters, whereas alternative similarity metrics yield inferior clusters. Therefore, Euclidean distance is the best similarity function in DIANA to map Tuberculosis data in North Sumatra compared to others. Further research could evaluate cluster outcomes by employing different DIANA clustering techniques alongside the Diana method, assessing the optimal distance metric that include Similarity Cosine, Manhattan, and Mincowski distances.

ACKNOWLEDGMENT

This research is supported/supported entirely by Universitas Negeri Medan (Unimed). We would like to thank our colleagues from Unimed who have provided insights and knowledge that significantly contributed to the this study, even though they might not see eye to eye with every interpretation or conclusion presented in this document. Additionally, we gratefully acknowledge the financial support provided by BOPTN Universitas Negeri Medan (Grant No.292D/UN33.8/PL/2025).

REFERENCES

[1] E. Avcı, Z.D. Muharremoğlu, E.N.N. Bozkurt, S. Kaygusuz, Changing Epidemiology of Tuberculosis and Actions Taken in the World and Türkiye, *Journal of Clinical Practice and Research*, 46(5), 421–430, 2024

[2] J. Han, M. Kamber, and J. Pei, (2017) Data Mining; Concepts and Techniques 3rd ed., USA: Elsevier Inc., 2012.

[3] J. S. Setyo, A. Sudradjat, C. M. Pasien, and P. Kesehatan, (2017) Penerapan Metode C4.5 Terhadap Penyakit Tuberkulosis Paru, Jurnal Kajian Ilmiah., 17(3), 111-118.

[4] R. F. Ramadhan, S. H. Wijoyo, and M. C. Saputra (2023) Penerapan Metode K-Means Clustering pada Ulasan Perumahan PT XYZ di Google Maps untuk Formulasi Strategi Bisnis dengan Analisis SWOT, *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 7(6), 2879-2888

[5] D.S., Mwakapesa, Y. Mao, X. Lan, Y.A. Nanehkaran, (2023) Landslide Susceptibility Mapping Using DIvisive ANAlysis (DIANA) and RObust Clustering Using linKs (ROCK) Algorithms, and Comparison of Their Performance. *Sustainability*, 15, 4218. 1-20 https://doi.org/10.3390/su15054218

[6] J. Wang, W. Feng, M. Huang, S. Feng, D. Du, (2024,) Improvement of Practical Byzantine Fault Tolerance Consensus Algorithm Based on DIANA in Intellectual Property Environment Transactions. *Electronics*, 13, 1634. 1-18 https://doi.org/10.3390/electronics13091634

[7] S. Wijuniamurti, S. Nugroho and R. Rachmawati (2022) Agglomerative Nesting (AGNES) Method and Divisive Analysis (DIANA) Method for Hierarchical Clustering on Some Distance Measurement Concepts, JSDS: *Journal of Statistics and Data Science*, 1(1), 7-11

[8] W. N. Fadillah, Y. M. Rangkuti, and I. M. K. Karo (2024) Analyzing The Effect of Similarity Functions on Partitioning Around Medoids Algorithm for Mapping DHF Disease In North Sumatra, *BAREKENG: Journal of Mathematics and Its Applications*, 18(1) 0413-0426, https://doi.org/10.30598/barekengvol18iss1pp0413-0426

[9] A. Aditya, B. N. Sari, and T. N. Padilah, (2021) Perbandingan pengukuran jarak Euclidean dan Gower pada klaster k-medoids," *Jurnal Teknologi dan Sistem Komputer*, 9(1), 1–7, doi: 10.14710/jtsiskom.2021.13747.

[10] B. M. Kumarahadi, H. Pratiwi, and S. Subanti (2023) Penerapan metode Hierarchical Clustering untuk pengelompokan Kota/ Kabupaten di Indonesia berdasarkan kategori kemiskinan, *Jurnal TIKomSin*, 11(2), 8-12

[11] K. H. Izzuddin, and A. W. Wijayanto (2024) Pemodelan Clustering Ward, K-Means, DIANA, dan PAM dengan PCA untuk Karakterisasi Kemiskinan Indonesia Tahun 2021, *Komputika: Jurnal Sistem Komputer*, 13(1,), 41-53

[12] A. K. Suyono, M. N. Hayati, M. Siringoringo, S. Prangga, and M. Fathurahman (2024) Penerapan Algoritma Divisive Analysis dalam Pengelompokan Provinsi di Indonesia Berdasarkan Prevalensi Stunting, *Jurnal Eksponensial*, 15(2), 119-127

[13] N. S. L. Zulfa, and A. Hadiana (2021) Kajian Data Mining Menggunakan Algoritma K-Means dan K-Medoids dalam Strategi Promosi (Studi Kasus: Universitas Islam Al-Ihya Kuningan). *Jurnal Fakultas Teknik*, 2(2), 57–62. https://repository.uinjkt.ac.id/dspace/bitstream/123456789/71093/1/DATA MINING.pdf

[14] C. Simar, N. Gusriani, and I. Irianingsih (2020) Pengelompokan Kabupaten/Kota di Jawa Barat Tahun 2018 Berdasarkan Indikator Kemiskinan dengan Polythetic Divisive Method. *Kubik: Jurnal Publikasi Ilmiah Matematika*, 5(1), 35–42. https://doi.org/10.15575/kubik.v5i1.8549

[15] K. S. Raju, and D. N. Kumar (2010) *Multicriterion Analysis in Engineering and Management*. PHI Learning Private Limited.

[16] R. Vashistha, and S. Nagar (2017) An intelligent system for clustering using hybridization of distance function in learning vector quantization algorithm. 2017 *Second International Conference on Electrical, Computer and Communication Technologies* (ICECCT), pp. 1-7

[17] S. Bishnoi and B.K. Hooda, (2020), A survey of distance measures for mixed variables, *International Journal of Chemical Studies*, SP-8(4), 338-343, https://doi.org/10.22271/chemi.2020.v8.i4f.10087

[18] D. R. Agustian, and B. A. Dermawan (2022) Analisis Clustering Demam Berdarah Dengue Dengan Algoritma K-Medoids (Studi Kasus Kabupaten Karawang), JIKO (*Jurnal Informatika dan Komputer*), 6(1), 18–26