JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 6 No. 1 (2025) 262-277
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Comparative Analysis of Distance Functions on DBSCAN Algorithm: Mapping Malnourished Toddlers in Medan City, Indonesia

Ichwanul Muslim Karo Karo[1], Mohd Farhan Md Fudzee[2]*, Shahreen Kasim[2], Azizul Azhar Ramli[2], Jemal H. Abawajy[3], Mohammad Syafwan Arshad[4]

[1] *Computer Science, Faculty of Mathematics and Natural Science, Medan State University, Medan, 20221, INDONESIA*
[2] *Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400, Johor, MALAYSIA*
[3] *School of Information Technology, Deakin University, Geelong VIC 3220 AUSTRALIA*
[4] *MZR Global Sdn Bhd, Shah Alam, MALAYSIA*

*Corresponding Author: farhan@uthm.edu.my
DOI: https://doi.org/10.30880/jscdm.2025.06.01.017

Article Info

Received: 12 September 2024
Accepted: 28 April 2025
Available online: 30 June 2025

Keywords

Malnutrition toddler, DBSCAN, distance functions, silhouette index, *Minpts*

Abstract

Medan City is one of Indonesia's largest cities and faces fundamental challenges in addressing malnourished toddlers. It had a stunting prevalence of 19.9% in 2022. The high rates necessitate a practical approach to identifying and managing high-risk areas. This study aims to map districts in Medan City based on the spatial data of public health center locations and malnutrition data for toddlers, using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. DBSCAN is a popular clustering algorithm because of its ability to group data based on density and detect outliers as noise. However, using the Euclidean distance function in DBSCAN may not be appropriate for all geospatial cases. The novelty lies in comparing five distance functions (Euclidean, Manhattan, Minkowski, Cosine, Chebyshev) within DBSCAN to determine which produces the most meaningful clustering in a geospatial health context. The study shows that DBSCAN with the Chebyshev distance function cannot effectively map the malnutrition problem in toddlers, as indicated by a Silhouette index (SI) value below 0.25. The clustering quality using Minkowski and Cosine distance functions in DBSCAN is not superior to that of the classical DBSCAN, with all three producing weak and unclear structures. The most effective mapping results come from using the Manhattan distance function in DBSCAN, which yields an SI value of 0.51045, two clusters, and optimal parameters of Minpts = 6–9 and ε = 6.98–7.8. The first cluster includes two districts (Medan Labuhan and Marelan), while the remaining districts form the second cluster. The analysis of different distance functions provides new insights into how selecting the appropriate distance measure can influence clustering quality in a geospatial context with DBSCAN. The similarity of the clusters is expected to inform decision-making in addressing toddler malnutrition issues in Medan City.

1. Introduction

Medan, the third-largest and most developed city in Indonesia [1], still struggles with the fundamental problem of malnourished toddlers. The progress of Medan City is directly proportional to the number of malnutrition cases that increase every year. The stunting prevalence in Medan City in 2022 is still at 19.9%. The number explains that there are 23,725 malnourished toddlers out of 119,225 toddlers [2]. Obviously, these conditions pose a threat to children's health in the short and long term. Short-term impacts include thinness, wasting, stunting, and even death. Meanwhile, in the long term, it can also trigger a generation of idiots. The local government is mitigating the problem of malnutrition by appointing a public health center (*Puskesmas*) in each district as the main service for handling malnutrition [3], [4]. A *Puskesemas* covers a district area. Mapping these areas can generate useful information to inform the development of a policy aimed at reducing malnutrition rates. Therefore, it is necessary to map malnourished toddler cases by area to develop an effective intervention plan. However, mapping a malnourished toddler is not easy because clinical and managerial issues, vector control, preventive measures, and surveillance must be considered.

Clustering is a powerful data analysis tool to detect clusters, groups, or map a set of physical or abstract objects based on their similarity[5]. Clustering has been widely implemented in the health area, such as Covid-19 mapping [5], [6], [7], [8], [9], mapping stunting [10], [11], [12], malnutrition cases [13], [14], [15], [16] and other health issues. The mapping results of the clustering task will be useful in planning and handling health problems based on the information gathered. There are many clustering algorithms, such as K-Means [17], PAM [18], DIANA [19], DBSCAN [20], etc.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is one of the popular and best clustering algorithms used to group data based on density in the data space [6], [20], [21]. This algorithm does not require a predetermination of the number of clusters and is able to find clusters with arbitrary shapes [22], as well as being able to identify outlier data as noise. Therefore, the DBSCAN algorithm is useful in many mapping applications. The DBSCAN algorithm groups data in a cluster based on two important parameters: Epsilon (ε) and Minimum point (*MinPts*). *Minpts* is the minimum number of items required in a cluster. Epsilon (ε) is the distance value used as the basis for defining the neighborhood of a data point. The classical DBSCAN algorithm uses the Euclidean distance function to determine ε and data classification (core point, border point, and outlier). Additionally, the DBSCAN algorithm is often used to cluster non-spatial cases that do not involve spatial elements. Philosophically, there is a bias in the utilization of the DBSCAN algorithm. Moreover, for some geospatial datasets, the Euclidean distance is not necessarily representative [9].

This study aims to map the district area in Medan City based on the spatial information of Puskesmas locations and data on malnourished toddlers. The mapping process uses DBSCAN clustering. The DBSCAN algorithm has advantages in spatial data clustering, particularly in complex urban environments like Medan City. Unlike partition-based algorithms such as K-Means, DBSCAN does not require prior knowledge of the number of clusters [22], which is beneficial when analyzing real-world data with unknown or uneven distributions. Its ability to identify clusters of arbitrary shape and handle noise or outliers makes it well-suited for spatial datasets, where malnutrition cases may be concentrated irregularly across districts. Moreover, DBSCAN is a density-based algorithm, allowing it to group high-risk areas based on the concentration of malnutrition cases and ignore isolated instances as noise. This characteristic aligns well with the study's objective, which is to map meaningful clusters of malnutrition prevalence while accounting for both spatial proximity and health severity. These capabilities make DBSCAN a robust and effective tool for uncovering patterns in public health surveillance data. The output of the clustering results is in the form of district group information that can be used as a reference material to alleviate malnourished toddlers. This research also analyzes various distance functions so as to produce the best cluster.

These are the main contributions of this study. The study applies the DBSCAN clustering algorithm to map malnourished toddler cases across Medan City, Indonesia, utilizing both spatial (location of health centers) and non-spatial (malnutrition stats) data. The novelty lies in evaluating five distance functions (Euclidean, Manhattan, Minkowski, Cosine, and Chebyshev) within DBSCAN to observe which provides the best and most relevant clustering in a geographic health location. The study finds that Manhattan distance yields the best clustering results (Silhouette Index = 0.51045), outperforming the classical approach and others. It demonstrates that the choice of distance function has a significant impact on the clustering outcome of DBSCAN in spatial health data analysis. It demonstrates that the choice of distance function has a significant impact on the clustering outcome of DBSCAN in spatial health data analysis. Practically, the implication of two districts in the highest-risk cluster offers actionable insights for local government and healthcare providers to prioritize interventions.

The rest of this paper is structured into three sections. Section 2 analyzes recent research on the DBSCAN algorithm and its applications, with a focus on health-related data clustering, and discusses the various distance functions commonly used in clustering. Section 3 outlines the proposed method, including data collection, preprocessing, and the modified DBSCAN algorithm. Section 4 shows and analyzes the experimental findings, evaluating the effectiveness of various distance functions in clustering malnourished patients. Finally, the last

section summarizes the findings and proposes future directions for enhancing geospatial clustering through the optimization of distance functions.

2. Related Studies

2.1 Related Study of DBSCAN Algorithm

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is a group of density-based clustering algorithms[6]. This algorithm is able to identify a number of clusters without user intervention, which is also what distinguishes it from partition-based or hierarchical clustering algorithms. The DBSCAN algorithm is able to find clusters with irregular shapes and identify outliers (noise). The DBSCAN algorithm is also designed for spatial data, namely data that has elements of location information, such as geo-location, area, or the like. However, the Euclidean distance function on classical DBSCAN to determine Epsilon (ε) is not entirely reliable and relevant in particular cases.

The DBSCAN algorithm is frequently used to analyze various geospatial cases, including mapping areas, detecting hotspots, and identifying high-density zones. Firstly, this paragraph presents some implementations of the DBSCAN algorithm for mapping or clustering diseases. A study introduces a cluster-based method for dengue contingency planning by grouping patient cases based on environmental and demographic factors[19]. The plan is mapped out by selecting appropriate measures for each area using clustering techniques. Silhouette scoring is employed to select the best features, determine the optimal clustering method, and evaluate cluster severity. Cluster severity is categorized into levels (low, medium, high, and extreme) and is aligned with plans specific to the village and the various season types. The study by [23] focused on understanding rabies outbreaks in herbivores through temporal analysis with moving averages and spatial diffusion mapping. Spatial clustering was analyzed using nearest neighbor analysis and ST-DBSCAN to identify outbreak patterns. Another study by [9] compared DBSCAN and K-Means algorithms for clustering hotspot zones of Covid-19 positive patients. That is an effort to control the spread of the pandemic. Improvements or modifications do not accompany the implementation of both algorithms. In other words, the distance function in both algorithms uses Euclidean distance. Both algorithms successfully grouped into four clusters. In similar cases, a study by [6] using the DBSCAN algorithm to group countries with similar COVID-19 case patterns. To ensure the study's results can be effectively applied as recommendations for handling cases, the optimal Eps value range is 0.1-0.2, and the Minpts value is 3 or 4. The best number of clusters is 3 clusters with a Silhouette Index value of 0.3624. Obviously, the quality of the resulting cluster is still weakly structured.

This paragraph presents research related to the improvement or modification of DBSCAN, particularly in terms of algorithm optimization. The study by [20] improved DBSCAN algorithm based on neighbor similarity. It utilizes Cover Tree to retrieve the neighbors of each point in parallel and uses the triangle rule (triangle inequality) to reduce many unnecessary distance calculations. From experiments conducted on large-scale datasets, the proposed algorithm is shown to significantly improve the performance of the original DBSCAN and outperform other major enhancements of DBSCAN. The proposed algorithm has two advantages: first, it is faster; and second, the results are more accurate compared to rho-approximate DBSCAN, which is the fastest version of DBSCAN. Another modified idea is analyzing the distance function on the DBSCAN algorithm [24]. It analyzes the comparison between eight similarity measures in DBSCAN for moving object trajectories. The distance functions analyzed are Euclidean, L1, Hausdorff, Fréchet, Dynamic Time Warping (DTW), Longest Common SubSequence (LCSS), Edit Distance on Real signals (EDR), and Edit Distance with Real Penalty (ERP) on three different datasets with diverse characteristics. The evaluation results show that the selection of an appropriate distance function depends on the data and motion parameters. However, overall, Euclidean distance shows superiority in terms of purity index, while EDR distance provides better performance in terms of spatial and spatiotemporal quality of the clusters. In terms of computation time and scalability, the Euclidean, L1, and LCSS distance functions are the most efficient. Table 1 is a summary of several studies that discuss the application of the DBSCAN algorithm.

Table 1 *Summary of related study*

Ref.	Objective of Study	Edit the Value of DBSCAN	Result/Evaluation
[19]	Cluster dengue patient cases based on environmental and demographic factors, and the cluster is categorized into low, medium, high, or extreme levels.	-	The study used data from 15,000 cases from Semarang City, Indonesia (2016-2020). The evaluation was conducted using silhouette scores. K-Means mapped cluster severity and matched clusters to suitable contingency policies, achieving a high silhouette score compared to DBSCAN and agglomerative clustering, with three optimal clusters.
[23]	This study focused on understanding rabies outbreaks in herbivores through temporal analysis with moving averages and spatial diffusion mapping.	Combine nearest neighbor analysis and Spatio-Temporal-DBSCAN.	Clusters of outbreaks were identified in the regions of Araguaína and Palmas (nearest neighbor index = 0.555; $p < 0.05$).
[9]	This study categorizes hotspot zones of COVID-19 positive patients as an effort to control the pandemic spread.	-	The DBSCAN method separates the data into 4 main clusters and noise points with ε = 0.45 and *Minpts*=20.
[6]	This study categorizes countries with COVID-19 case patterns to obtain case-control recommendations.	-	Best ε is 0.1 – 0.2; value of *Minpts* =3 and 4. The best number of clusters is 3 with a Silhouette Index value of 0.3624.
[20]	This study tries to improve the DBSCAN algorithm's performance by filtering many unnecessary distance computations.	Improved DBSCAN algorithm based on neighbor similarity.	Improved DBSCAN has two advantages: it is faster, and the resulting clusters are more accurate.
[24]	This study compares eight similarity measures in the density-based clustering of moving objects' trajectories.	Replacing distance functions (Euclidean, L1, Hausdorff, Fréchet, Dynamic Time Warping (DTW), Longest Common SubSequence (LCSS), Edit Distance on Real signals (EDR), and Edit Distance with Real Penalty (ERP)) on DBSCAN to cluster three different datasets.	Euclidean distance function performed better based on purity index, EDR distance function performed better for spatial and spatio-temporal clusters. Euclidean, L1, and LCSS distance functions are the most efficient in terms of computation time and scalability.

2.2 Distance Metric

Suppose there are two objects, A and B. The Distance function is the most commonly used approach to measure how close or similar the two objects are. A practical distance function improves the performance of our machine learning model [24], [25], [26]Whether that's for classification tasks or clustering. Knowing when to use which distance measure can help you go from a poor classifier to an accurate model. There are five popular and most used distance functions in data science: Euclidean, Manhattan, Minkowski, Cosine, and Chebyshev. An illustration of how they work can be seen in Fig. 1, and the brief is shown in Table 2.

J. of Soft Computing and Data Mining Vol. 6 No. 1 (2025) p. 262-277

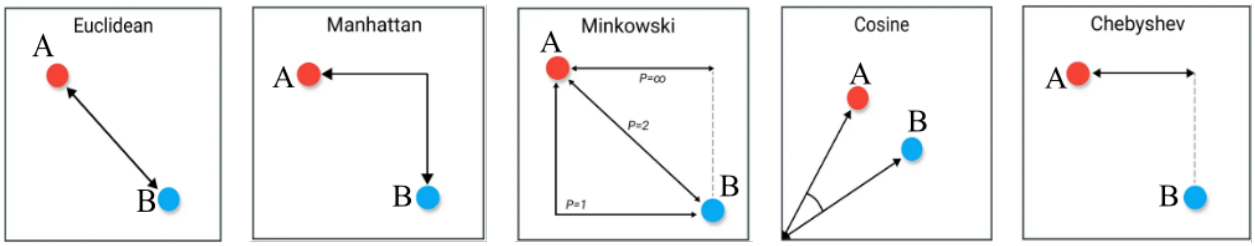

Fig. 1 *Distance functions illustration*

Table 2 *Brief of distance functions*

Distance Metric	Description	Formula
Euclidean	The most common function to measure the proximity between two vectors in dimensional space. Euclidean distance works by drawing a "straight line" distance between two objects	$\boldsymbol{dist_{Euclidean}}\ (\boldsymbol{A},\boldsymbol{B}) = \sqrt{\sum_{i=1}^{n}(\boldsymbol{A_i} - \boldsymbol{B_i})^2}$
Manhattan	Also known as Taxicab Distance or L1 Norm to calculate the distance in vector space [26]. Unlike the Euclidean distance, the Manhattan distance calculates the distance between two points by summing the absolute projection lengths of their differences on the coordinate axes.	$\boldsymbol{dist_{Manhattan}}\ (\boldsymbol{A},\boldsymbol{B}) = \sum_{i=1}^{n} \lvert \boldsymbol{A_i} - \boldsymbol{B_i} \rvert$
Minkowski	This allows distance measurement in a more flexible vector space by setting parameter p, the parameter determines the type of metric. If p = 1, it becomes Manhattan distance, else if p = 2, it becomes Euclidean Distance, else $p \to \infty$, it become Chebyshev distance.	$\boldsymbol{dist_{Minkowski}}(\boldsymbol{A},\boldsymbol{B}) = \sqrt[p]{\sum_{i=1}^{n}\lvert \boldsymbol{A_i} - \boldsymbol{B_i} \rvert^p}$
Cosine	Cosine distance is used to determine how different two objects based on cosine angle of the two vectors, rather than geometric distance.. If the vectors are perpendicular to each other, Cosine similarity is 0; if they are identical, the cosine value is 1. It is useful in many applications, especially when orientation and text similarity.	$\boldsymbol{dist_{Cosine}}\ (\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A.B}}{\lvert\lvert \boldsymbol{A} \rvert\rvert\ \lvert\lvert \boldsymbol{B} \rvert\rvert}$
Chebyshev	also known as L∞ Norm or Maximum Distance, is a distance metric that measures the farthest distance along a single coordinate axis in vector space. Chebyshev distance is often used in scenarios where movement is only allowed in straight lines parallel to the coordinate axes. In this context, Chebyshev distance determines the maximum number of steps required to move from one point to another	$\boldsymbol{dist_{Chebyshev}}(\boldsymbol{A},\boldsymbol{B}) = \mathbf{max} \lvert \boldsymbol{A_i} - \boldsymbol{B_i} \rvert$

3. Proposed Method

The research flowchart for this study is shown in Fig. 2. There are four key processes involved in achieving the objective: data collection, preprocessing, clustering, and evaluation. The novelty of data can be a contribution to research. The data used is collected from various sources, making it unique and previously unutilized. Data preprocessing is a crucial step in data analysis and machine learning, involving the preparation of raw data to be transformed into a suitable and optimal format for clustering models. This process encompasses various techniques for handling data imperfections. The next stage is mapping malnutrition cases in Medan City with the DBSCAN algorithm. Additionally, this stage analyzes the distance function of the algorithm to produce the best cluster. The final stage is the evaluation process. Each cluster generated from various experimental scenarios will be evaluated on the quality of the resulting cluster with the Silhouette index method.

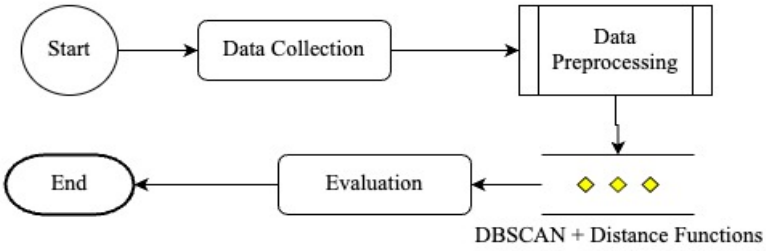

Fig. 2 *Flowchart research*

3.1 Data Collection

This study uses two types of data: spatial and non-spatial. Spatial data is data that has an element of location, in this case, represented by the location of the Puskesmas in each neighborhood of Medan City, Indonesia, which handles cases of malnourished toddlers. Non-spatial data refers to the condition of malnourished toddlers in the area covered by the Puskesmas, including the number of malnourished toddlers, the number of patients who died from malnutrition, and the number of patients who recovered. Both datasets are stored in memory (Shown in Table 3).

Table 3 *Dataset of malnutrition cases in Medan City*

Districts	Longitude	Latitude	Malnutrition Cases		
			Amount	Died	Recovery
Medan Tuntungan	3.51768116543486	9.86167687233008	3	3	0
Medan Selayang	3.55179591215769	9.86390768172166	81	50	0
Medan Johor	3.54980344136439	9.87071134836160	6	2	0
Medan Amplas	3.54512457917472	9.87116752661576	56	30	1
Medan Denai	3.57117920043062	9.87252523372700	22	14	1
Medan Tembung	3.60829810230488	9.86971369195422	18	1	4
Medan Kota	3.56216765996960	9.86952126761048	35	24	2
Medan Area	3.57842095476121	9.87026316247733	7	7	0
Medan Baru	3.56068632640928	9.86634267264142	0	0	0
Medan Polonia	3.56920876059754	9.86684968907061	43	21	0
Medan Maimun	3.54609820187685	9.86836191124246	82	23	2
Medan Sunggal	3.57657693933562	9.86123021842964	30	16	0
Medan Helvetia	3.61230901447922	9.86333009650401	2	0	0
Medan Barat	3.60783249178420	9.86692205883956	2	0	1
Medan Petisah	3.59481152520940	9.86624399413861	30	18	0
Medan Timur	3.61843390597287	9.86933804221675	24	10	2
Medan Perjuangan	3.60281702598655	9.87106461200089	70	20	0
Medan Deli	3.67674677429044	9.86643681109499	62	24	0
Medan Labuhan	3.72336323863162	9,86784015213942	105	44	3
Medan Marelan	3.71239240624435	9.86504842161186	74	39	4
Medan Belawan	3.78234032597988	9.86924560572178	61	26	0

The spatial data in this research uses the representation of *Puskesmas* locations in each district (D_i). In case there are 21 districts in Medan City, then $D = \{D_1, D_2, \ldots D_{21}\}$. *Puskesmas* are at the forefront of data collection and malnutrition eradication in Indonesia[3], [4]. Data on toddler malnutrition at the lowest level is collected at *Puskesmas*, not at other institutions. Further, the spatial data model for each district can be defined as a matrix, $D_i = \{Longitude_i, Latitude_i\}$, shown in Table 2. The spatial data is obtained from the location coordinates of the health center on Google Maps.

The non-spatial data in this research is represented by malnutrition condition data from each sub-district in Medan City in 2023. There are three non-spatial variables; amount, died and recovery. Furthermore, the non-spatial data model for each district can be defined as a matrix, $D_i = \{\text{amount}_i, died_{i,}, recovery_{i,}\}$, shown in Table 2. The non-spatial data was sourced from the Medan City Health Office database, Indonesia.

3.2 Data Preprocessing

The data that has been collected will be normalized using the Z-score. Z-score normalization is also known as the standardized value (shown in Equation (1)) [27]. This technique is commonly used in data mining to identify outlier data [27], [28]. Z-score normalization involves the process of transforming data to create a new range of values ($\boldsymbol{v}'$) based on the range of values present in the previous value ($\boldsymbol{v}$). The values generated through Z-score normalization are based on the difference between the mean and the standard deviation ($\boldsymbol{\sigma}$).

$$v' = \frac{v - \bar{x}}{\sigma} \tag{1}$$

3.3 Modified DBSCAN Algorithm

The DBSCAN algorithm works by examining each data point and identifying other points that are within an ε distance. Then, the points are clustered if they meet the *Minpts* criterion. The DBSCAN algorithm initiates the clustering process by visiting data that has not yet been assigned to a cluster. Choosing the right parameters ε and *MinPts can be challenging and significantly impact* the clustering results. To produce the best cluster, this study modified the DBSCAN algorithm by substituting various distance functions in the algorithm. In addition, this study also analyzes ε (with a value in the range of 0 -15) and Minpts (with a value in the range of 1-20) obtained using the modified algorithm.

Pseudocode Modified DBSCAN Algorithm

Input: Database *D*
Radius ε
Density threshold *MinPts*
Distance function *dist* ← Improved by five distance functions Euclidean, Manhattan, Minkowski, Chebychev, Cosine
Initially undefined label on each data in *D*
foreach data (p) in database D do
  if label (p) ≠ undefined then continue
  Neighbors N ← RANGEQUERY (D, dist, p, ε)
  if |*N*| < M*inPts* then
    label (p) ← Noise
    continue
  c ← next cluster label
  Seed set *S* ← *N* \{p}
  Foreach q in S do
    if label (q) = Noise then label (q) ← c
    if label (q) ≠ undefined then continue
    Neighbors N ← RANGEQUERY (D, dist, p, ε)
    label (q) ← c
    if |*N*| < M*inPts* then continue
    S ← S ∪ N

3.4 Evaluation

This study used Silhouette Index (*SI*) to evaluate the clustering results. *SI* evaluates the extent to which an object is correctly placed in a cluster that has been formed [29]. The *SI* calculation process is obtained from the average value of Silhouette score $s(i)$ for each data, mathematically can use Equation (2).

$$SI = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{2}$$

Meanwhile, *s(i)* for each data was calculated by using Equation (3). Value of $a(i)$ the average distance between i and all other data points in the same cluster (internal cohesion) and can be mathematically written as Equation (4). Then $b(i)$ is the average distance between i and all data points in the nearest distinct cluster (external separation), and then takes the smallest value. (Shown in Equation (5)).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)} \tag{3}$$

$$a(i) = \frac{1}{|A| - 1} \sum j \in A, j \neq i \; d(i,j) \tag{4}$$

$$b(i) = \min dist(i, C) \tag{5}$$

The calculation of $a(i)$ and $b(i)$ In s(i), previously used the Euclidean distance. This study modifies the calculation of *s(i)* by adjusting the distance function as in modified DBSCAN. The classical DBSCAN algorithm will be evaluated with the classical silhouette index. If modified DBSCAN substitutes the distance function with the Manhattan distance, then the Manhattan distance is also used to calculate $a(i)$ and $b(i)$ in *s(i).* The treatment is intended to be fair in the experiment.

The value of *SI* is in the range of -1 to +1. High positive *SI* values (close to +1) indicate that the cluster members are tightly clustered, compact, and well separated from other clusters. Negative *SI* values (close to -1) indicate that the data is not right in the existing cluster; it could be closer to other clusters. Therefore, the resulting cluster needs to be reviewed or reconfigured[30]. In detail, the interpretation of *SI* scores can be seen in Table 4.

Table 4 *SI interpretation*

SI value	Interpretation
$0.7 < SI \leq 1$	Clustering configuration is very strong, with well-separated and well-defined clusters
$0.5 < SI \leq 0.7$	Clusters are reasonably well-formed
$0.25 < SI \leq 0.5$	Clusters are likely to overlap, and the cluster boundaries are not very clear
$SI \leq 0.25$	Clustering is not meaningful

4. Results and Discussion

This section presents the dataset demonstration, statistical analysis of the data, discusses the clustering results of each variant of the DBSCAN algorithm and distance functions, and analyzes the comparison of the quality of the resulting clusters.

4.1 Dataset

Medan City has 21 districts. Each district has a *Puskesmas* that functions to record cases of infant malnutrition and is at the forefront of handling and alleviating it. Figure 3 is a demonstration of the distribution of *Puskesmas* locations in each district. The red pin is the location point (longitude, latitude) of the puskesmas in the district. Medan Labuhan is the largest district and Medan Maimum is the smallest district in Medan City. Medan Perjuangan district is the most densely populated area with 25,533 people/km^2. While the sparsest area is Medan Labuhan distict with 3,698 people/km^2. Medan Deli district is the area with the largest population. Meanwhile, the Medan Baru district is the area with the fewest number of people, at 36,545.

J. of Soft Computing and Data Mining Vol. 6 No. 1 (2025) p. 262-277

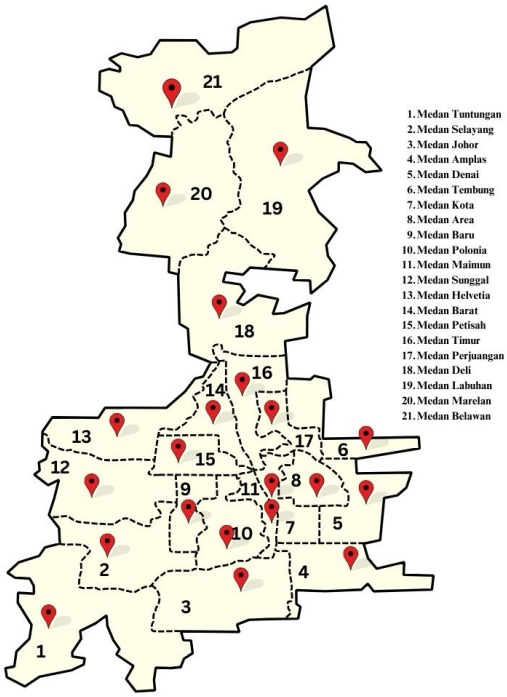

Fig. 3 *Distribution of Puskesmas and covering districts in Medan City*

The distribution of cases of malnourished toddlers in each district is presented in Figure 4. Medan Labuhan district is the area with the highest number of malnourished toddlers in Medan City. Meanwhile, the Medan Baru district is free from cases of malnourished toddlers. On average, there are 39 malnourished toddlers in each district, which is a substantial number. The condition is worsened because the recovery rate in each district is smaller than the number of toddlers who died from malnutrition. In other words, malnourished toddlers are more likely to die than recover. There are even 10 areas with a lack of recovery of malnourished toddlers. Malnourished toddlers in the district have died.

Next, the dataset is normalized using the Z-score method. Normalization is applied to standardize the range of values in each variable. As the ranges of spatial and non-spatial variables differ significantly, spatial variables fall within the value range of 3-10, while the non-spatial variable range spans 0-105. After normalization, the dataset will be modeled with the DBSCAN algorithm. There are five experiments with the DBSCAN algorithm. These are the classical DBSCAN algorithm, DBSCAN + Manhattan distance function, DBSCAN + Minkowski distance function, DBSCAN + Cosine distance function, and DBSCAN + Chebyshev distance function. Classical DBSCAN is the baseline experiment, where Euclidean distance is used to calculate similarity and ε.

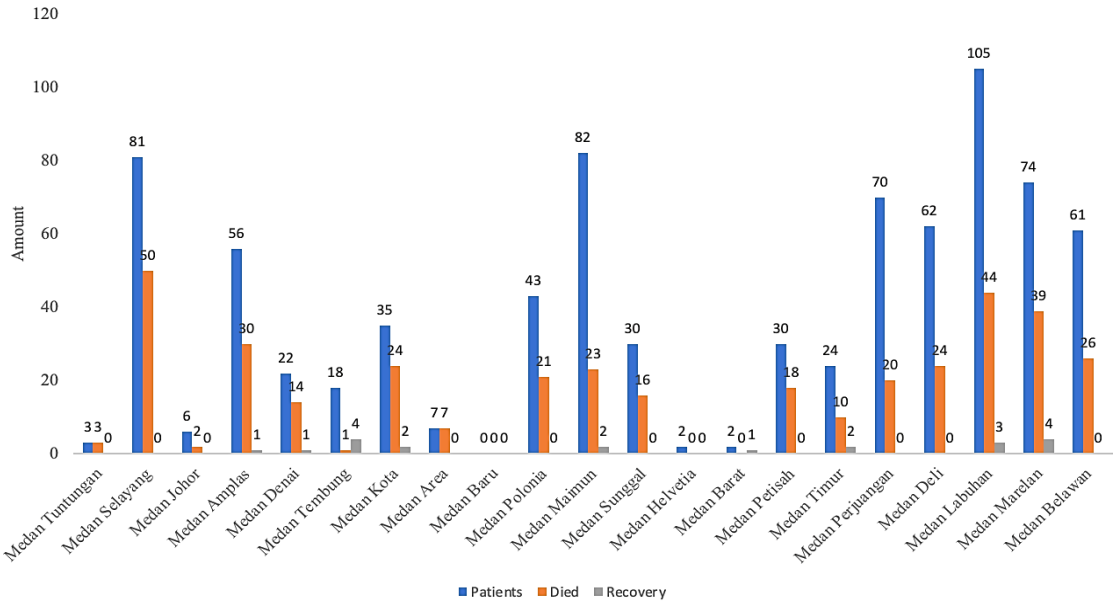

Fig. 4 *Statistical data on cases of malnourished toddlers in Medan City*

4.2 Baseline Experiment

The first experiment generalizes the dataset of malnourished toddler cases with the classical DBSCAN algorithm. Classical DBSCAN uses Euclidean distance function in determining the closeness and parameter ε. Classical DBSCAN successfully maps the malnourished infant case data into two clusters (shown in Figure 5). The first cluster consists of three districts (Medan Labuhan, Marelan and Selayang) while the other districts are included in cluster two. This study noted, the best parameter of *Minpts* is 11 - 14, and ε = 3.87 - 3.94, resulting in *SI* = 0.3529.

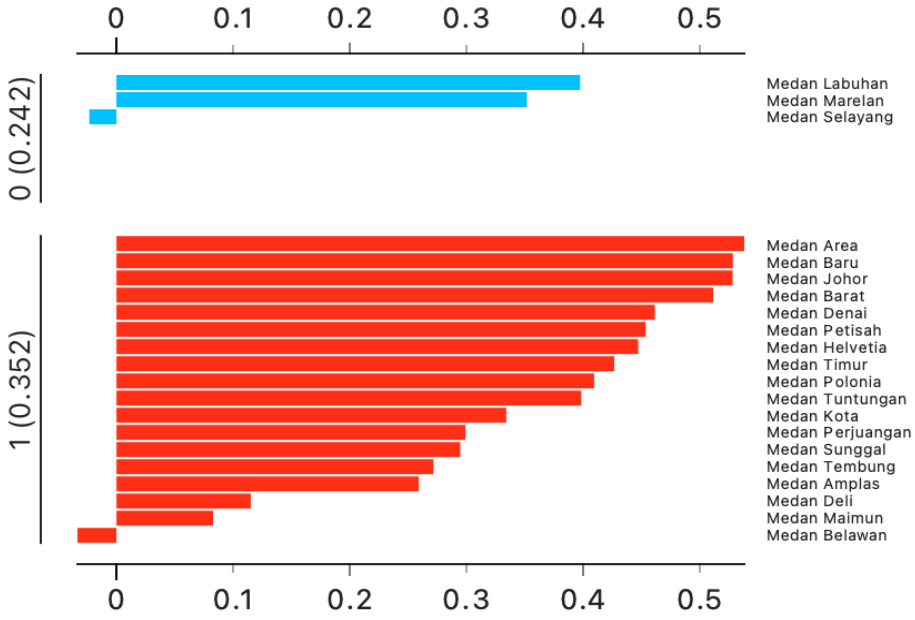

Fig. 3 *Classical DBSCAN result*

The resulting cluster quality has a weak structure and is not very clear, because 0.25 < *SI* ≤ 0.5. There are even four sub-districts with *SI* value < 0.25. There are the Medan Selayang district in the first cluster; Medan Maimun, Deli, and Belawan in the second cluster. In other words, these four districts may be in the wrong cluster. However, there are three districts with reasonable clusters: Medan area, Medan Baru, and Medan Johor. These districts have a well-formed structure.

4.3 Analysis of Manhattan Distance on DBSCAN Algorithm

The second experiment generalizes the dataset of malnourished toddler cases by substituting Euclidean distance with the Manhattan distance function on the DBSCAN algorithm. The objective is still the same as the previous distance function. The DBSCAN algorithm with the Manhattan distance function successfully maps the data of malnourished toddler cases into two clusters (shown in Figure 6). The first cluster consists of two districts (Medan Labuhan and Marelan), while the other districts are included in cluster two. This study noted that the optimal parameters for Minpts are 6-9 and ε = 6.98-7.8, resulting in an SI of 0.51045.

J. of Soft Computing and Data Mining Vol. 6 No. 1 (2025) p. 262-277

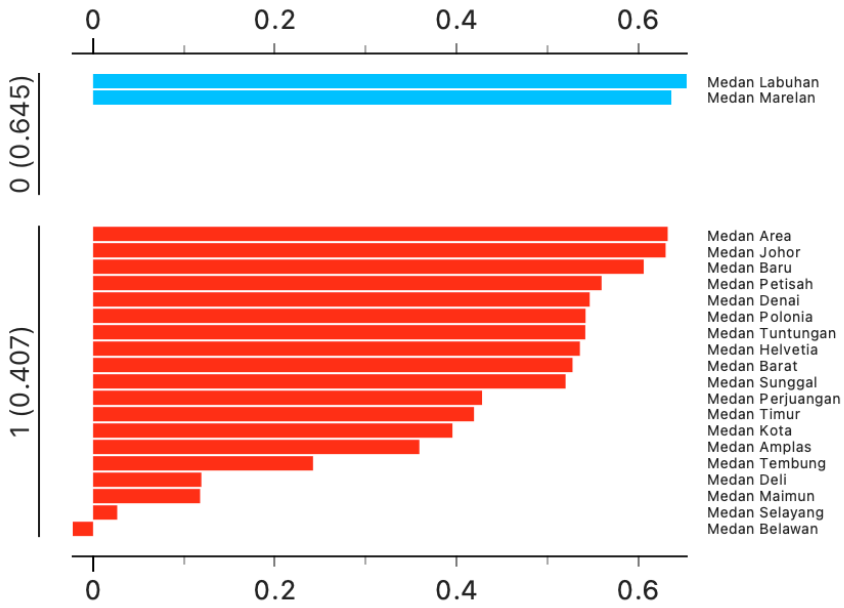

Fig. 6 *Result of DBSCAN algorithm + Manhattan distance*

The resulting cluster quality has a well-formed structure, as indicated by $0.5 < SI \leq 0.7$. Thus, the resulting cluster is reasonable. There are even 12 districts with *SI* value > 0.5, these are Medan Labuhan, Marelan, Area, Johor, Medan Baru, Petisah, Denai, Polonia, Tuntungan, Helvetia, West Medan, and Sunggal. These districts have a very strong and well-separated structure. However, there are four sub-districts with $SI < 0.25$: Medan Maimun, Deli, Selayang, and Belawan. These districts are in the second cluster

4.4 Analysis of Minkowski Distance on DBSCAN Algorithm

The third experiment is mapping the malnourished toddler case dataset by substituting Euclidean distance with the Minkowski distance function on the DBSCAN algorithm. The DBSCAN algorithm, combined with the Minkowski distance function, successfully clusters the malnourished infant case data into two distinct groups (as shown in Figure 7). The first cluster comprises four districts (Medan Labuhan, Marelan, Belawan, and Selayang), while the other districts are included in Cluster 2. This study noted that the optimal parameters for Minpts are 2-4 and ε = 3.84-3.98, resulting in an SI of 0.289.

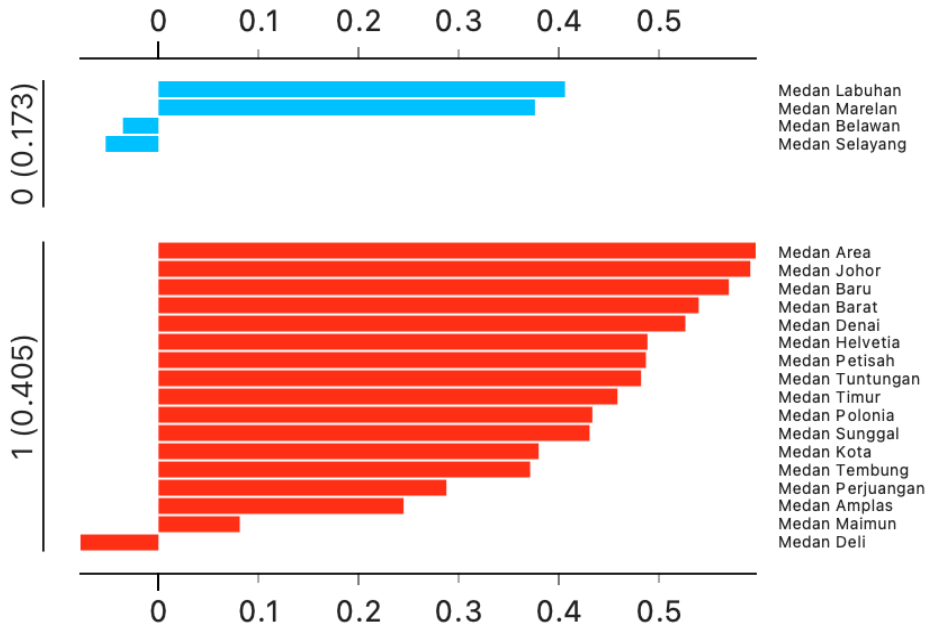

Fig. 7 *Result of DBSCAN algorithm + Mikowski distance*

The cluster quality produced in this experiment has a weak structure and is not very clear, because 0.25 < SI ≤ 0.5. Two sub-districts out of four in the first cluster have SI < 0.25; these are Medan Belawan and Selayang districts. In other words, half the members in cluster 2 are not meaningful. However, there are five districts with reasonable clusters: Medan area, Medan Baru, Medan Johor, Medan Denai, and Medan Helvetia. These districts have a well-formed structure.

4.5 Analysis of Cosine Distance on DBSCAN Algorithm

The fourth experiment is mapping the malnourished toddler case dataset by substituting Euclidean distance with the cosine distance function on the DBSCAN algorithm. The DBSCAN algorithm, combined with the cosine distance function, successfully clustered the malnourished toddler case data into two distinct groups (as shown in Figure 8). The first cluster comprises two districts (Medan Marelan and Medan Deli), while the other sub-districts are included in Cluster 2. This study noted that the best parameters of Minpts are 13 - 15, and ε = 1.4 - 1.74, resulting in *SI* = 0.259. The cluster quality produced in this experiment has a weak structure and is not very clear, because 0.25 < *SI* ≤ 0.5. There are no district areas with well-formed or strong structures. Even 15 districts of 21 are not a meaningful cluster.

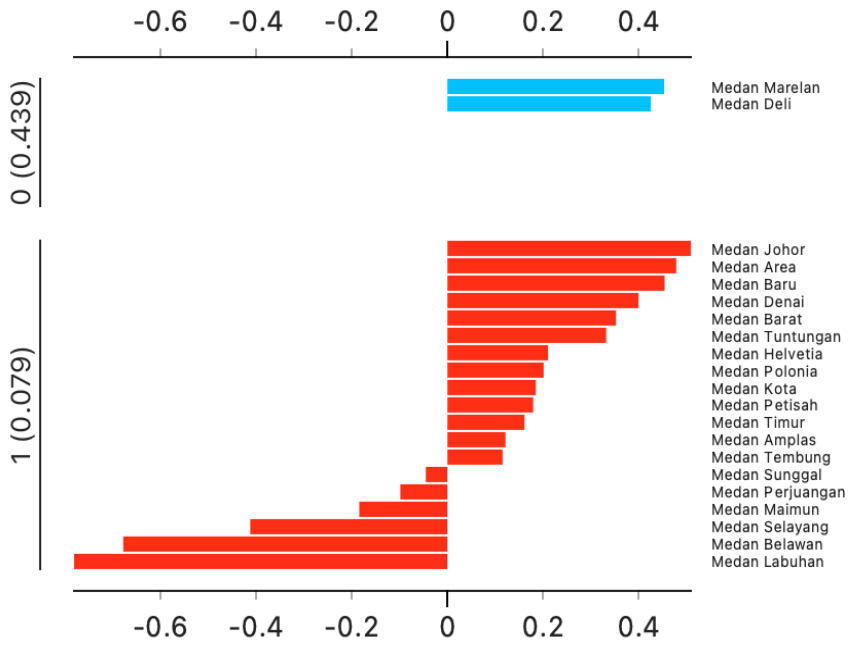

Fig. 8 *Result of DBSCAN algorithm + Cosine distance*

4.6 Analysis of Chebyshev Distance on DBSCAN Algorithm

The fifth experiment is mapping the malnourished toddler case dataset by substituting Euclidean distance with the cosine distance function on the DBSCAN algorithm. The DBSCAN algorithm, combined with the cosine distance function, successfully clustered the malnourished toddler case data into two distinct groups (as shown in Figure 8). The first cluster comprises three districts (Medan Marelan, Selayang, and Medan Perjuangan), while the other districts are included in Cluster 2. This study noted that the optimal parameters for Minpts are 3-5 and ε = 0.54-0.74, resulting in an SI of 0.1025. The cluster quality produced in this experiment is not meaningful because *SI* < 0.25. There are no sub-district areas with well-formed or strong structures. The majority of sub-districts are not meaningful and have weak structures.

J. of Soft Computing and Data Mining Vol. 6 No. 1 (2025) p. 262-277

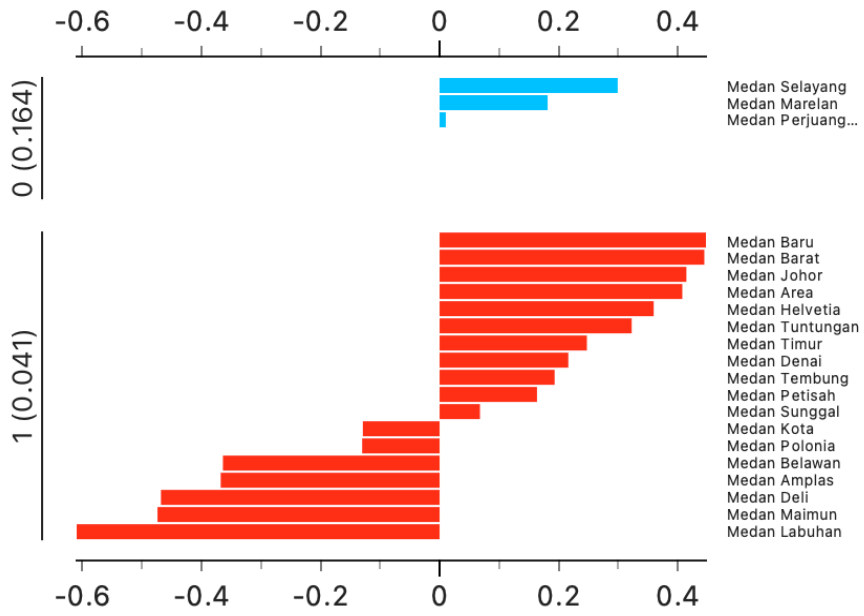

Fig. 9 *Result of DBSCAN algorithm + Chebyshev distance*

4.7 Comparative Analysis of Distance Functions on the DBSCAN Algorithm

This section presents a comparison of the cluster quality generated from five experiments. All experiments mapped the region into two clusters, with different cluster member compositions. Based on the results of the five experiments (shown in Fig 10), DBSCAN and Chebyshev distance function are not able to map the problem of malnourished toddlers based on sub-districts in Medan City, the *SI* value of the clustering results is below 0.25. In other words, Chebyshev distance function is not better than classical DBSCAN. The mapping results by Euclidean, Minkowski and Cosine distance functions on DBSCAN have a weak structure and are not very clear, thus they need to be reviewed. Unfortunately, Minkowski and Cosine distance functions on DBSCAN are not better than classical DBSCAN, because their *SI* is still lower than the *SI* of classical DBSCAN. The best mapping is obtained from the experiment DBSCAN algorithm with Manhattan distance; the *SI* value of this experiment is higher than the other four experiments. In other words, the Manhattan distance function has a positive and better effect than Euclidean on DBSCAN.

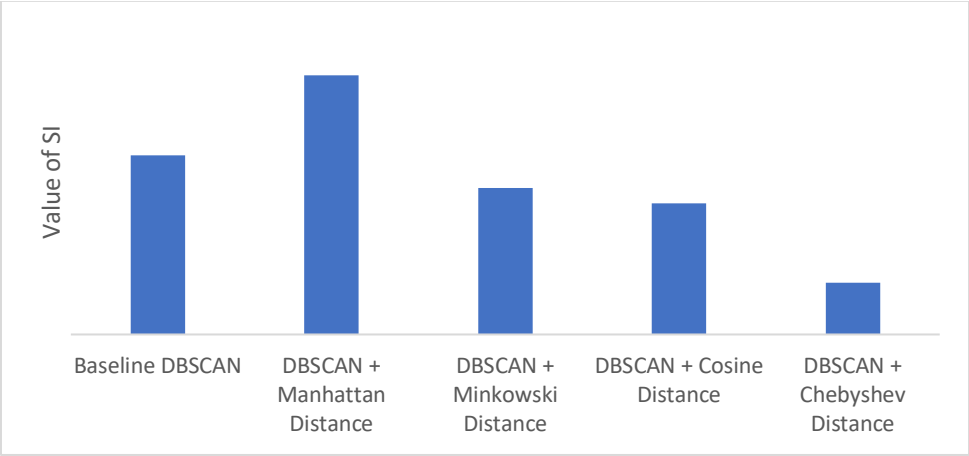

Fig. 10 *Result of experiments*

Although the mapping results of all experiments indicate that the number of clusters is 2, the best mapping for the case of malnourished infants in Medan City is achieved by the DBSCAN algorithm and the Manhattan distance function, resulting in two clusters. The composition of the first cluster consists of two sub-districts (Medan Labuhan and Marelan), while the other sub-districts are included in cluster two. Furthermore, when viewed from the spatial aspect, the regions with different clusters are also well separated, while the members

within one cluster are strongly bonded. This demonstrates that the DBSCAN algorithm can effectively map the spatial data of sub-districts in Medan city.

Multiple distance functions were analyzed in conjunction with the DBSCAN algorithm; the study only considered five commonly used metrics (Euclidean, Manhattan, Minkowski, Cosine, and Chebyshev). Other advanced or domain-specific distance functions might yield better clustering performance, but were not explored in this research. Although SI is a widely used metric, it may not capture all aspects of cluster validity, especially in the context of complex spatial data. Combining SI with other evaluation metrics, such as the Davies-Bouldin Index or external validation techniques, could provide a more comprehensive assessment.

5. Conclusion

This study focuses on mapping malnutrition cases among toddlers in Medan City, Indonesia, using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm. By utilizing spatial data (health center locations) and non-spatial data (malnutrition conditions), the research aims to identify areas with high malnutrition rates and provide recommendations for addressing them. The DBSCAN algorithm was chosen for its ability to group data based on density without needing to predefine the number of clusters, as well as its capability to identify outliers as noise. The study also analyzes various distance functions (Euclidean, Manhattan, Minkowski, Cosine, and Chebyshev) for the DBSCAN algorithm to determine the optimal clustering. Cluster evaluation was conducted using the silhouette index method. This research demonstrates the distribution of health center locations in each district of Medan City, Indonesia. Furthermore, the DBSCAN algorithm and Chebyshev distance function were unable to map the malnutrition problem among toddlers by district in Medan City, with clustering results having a silhouette index (SI) value below 0.25. Mapping results using the Euclidean, Minkowski, and Cosine distance functions on DBSCAN showed weak and unclear structures. Nevertheless, the cluster quality with the Minkowski and Cosine distance functions on DBSCAN was not better than that of the classical DBSCAN, as their SI values were still lower than those of the classical DBSCAN. The best mapping was obtained from the DBSCAN algorithm experiment using the Manhattan distance, yielding an SI value of 0.51045. The optimal parameters were MinPts = 6-9 and ε = 6.98–7.8. This SI was higher compared to the other four experiments. In other words, the Manhattan distance function had a positive and better impact compared to the Euclidean distance on DBSCAN. The best mapping for malnourished toddler cases in Medan City resulted in two clusters. The first cluster consisted of two districts (Medan Labuhan and Marelan), while the remaining districts were in the second cluster. Districts within the same cluster are expected to serve as a basis for decision-making in addressing malnutrition issues among toddlers in Medan City, with a focus on areas that require special attention. This study also notes the influence of incorporating spatial attributes and distance functions in the clustering process. Therefore, for future work, it is suggested that spatial similarity functions be incorporated into the clustering process.

Acknowledgment

The authors would like to thank the Faculty of Mathematics and Natural Science, Big data and Artificial Intelligent Research Center Medan State University, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM) and School of Information Technology Deakin University for supporting this work.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

Draft and revised manuscript, design, running and analyze experiment: *Ichwanul Muslim Karo Karo;* ***validation of methodology and results interpretation:*** *Mohd Farhan Bin Md. Fudzee;* ***paper conception:*** *Shahreen Binti Kasim, Azizul Azhar Ramli, Jemal H. Abawajy, Mohammad Syafwan Arshad. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Ponirin, T., Rambe, T., & Khairani, L. (2021). Mapping tourism potential based on urban heritage tourism in Medan City. CSR International Journal, 1(1). https://doi.org/10.35307/csrij.v1i1.16

[2] Sitorus, R., Munthe, D. S., Sinuhaji, L. N. B., Situmorang, T. S., & Sitorus, S. (2023). Predisposing, supporting and reinforcing factors of stunting risk: A case-control study. Jurnal Gizi dan Dietetik Indonesia (Indonesian Journal of Nutrition and Dietetics), 11(1). https://doi.org/10.21927/ijnd.2023.11(1).11-21

[3] Absori, A., Hartotok, H., Dimyati, K., Nugroho, H. S. W., Budiono, A., & Rizka, R. (2022). Public health-based policy on stunting prevention in Pati Regency, Central Java, Indonesia. Open Access Macedonian Journal of Medical Sciences, 10. https://doi.org/10.3889/oamjms.2022.8392

[4] Rahmadiyah, D., Sahar, J., & Widyatuti, W. (2022). Public health interventions to reduce stunting in toddlers: A systematic review. Open Access Macedonian Journal of Medical Sciences, 10(F). https://doi.org/10.3889/oamjms.2022.8610

[5] Darques, R., Trottier, J., Gaudin, R., & Ait-Mouheb, N. (2022). Clustering and mapping the first COVID-19 outbreak in France. BMC Public Health, 22(1). https://doi.org/10.1186/s12889-022-13537-7

[6] Nurhaliza, M. N. (2021). Clustering of data COVID-19 cases in the world using DBSCAN algorithms: Pengelompokan data kasus COVID-19 di dunia menggunakan algoritma. Indonesian Journal of Informatic Research and Software Engineering, 1(1).

[7] Herawati, N., Nisa, K., & Saidi, S. (2022). Implementation of the trimmed K-means clustering method in mapping the distribution of COVID-19 in Indonesia. AIP Conference Proceedings. https://doi.org/10.1063/5.0103175

[8] Valerian, F., & Yulianto, S. (2022). Identification of the COVID-19 distribution area on the island of Kalimantan using the K-means spatial clustering method. Jurnal Teknik Informatika (JUTIF), 3(4). https://doi.org/10.20884/1.jutif.2022.3.4.314

[9] Slaam, M. A., Gouda, K., & Naguib, A. (2022). Data clustering mapping of COVID-19 pandemic based on geo-location and machine learning. International Journal of Computer Science and Network Security, 22(4).

[10] Indra, I., Nur, N., Iqram, M., & Inayah, N. (2023). Perbandingan K-means dan hierarchical clustering dalam pengelompokan daerah beresiko stunting. INOVTEK Polbeng – Seri Informatika, 8(2). https://doi.org/10.35314/isi.v8i2.3612

[11] Christyanti, R. D., Sulaiman, D., Utomo, A. P., & Ayyub, M. (2022). Clustering wilayah kerawanan stunting menggunakan metode fuzzy subtractive clustering. Jurnal Ilmiah Teknologi Informasi Asia, 17(1). https://doi.org/10.32815/jitika.v17i1.877

[12] Anggraeni, M. R., Yudatama, U., & Maimunah. (2023). Clustering prevalensi stunting balita menggunakan agglomerative hierarchical clustering. Jurnal Media Informatika Budidarma, 7(1).

[13]

[14] Siallagan, S. A., & Safii, M. (2021). Grouping of toddlers with malnutrition based on provinces in Indonesia using K-medoids algorithm. Journal of Artificial Intelligence and Engineering Applications (JAIEA), 1(1). https://doi.org/10.59934/jaiea.v1i1.53

[15] Ghosh, K., Chakraborty, A. S., Haloi, B., & Zakir, S. (2024). Spatial clustering of malnutrition and anemia among reproductive women and its associated risk factors in India: Evidence from National Family Health Survey-5. Food and Nutrition Bulletin, 45(1). https://doi.org/10.1177/03795721241234086

[16] Narulita, S., Prihati, P., Oktaga, A. T., & Widyantoro, A. E. (2023). Performansi algoritma clustering K-means untuk penentuan status malnutrisi pada balita. Jurnal Informasi, Sains dan Teknologi, 6(1). https://doi.org/10.55606/isaintek.v6i02.128

[17] Nagari, S. S., & Inayati, L. (2020). Implementation of clustering using K-means method to determine nutritional status. Jurnal Biometrika dan Kependudukan, 9(1). https://doi.org/10.20473/jbk.v9i1.2020.62-68

[18] Saputra, M. A. W., & Harini, S. (2022). Java Island health profile clustering using K-means data mining. International Journal on Information and Communication Technology (IJoICT), 8(1). https://doi.org/10.21108/ijoict.v8i1.606

[19] Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager K-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. Information Systems, 101. https://doi.org/10.1016/j.is.2021.101804

[20] Husna, F. A., Purwitasari, D., Sidharta, B. A., Sihombing, D. A., Fahmi, A., & Purnomo, M. H. (2022). A clustering approach for mapping dengue contingency plan. Scientific Journal of Informatics, 9(2). https://doi.org/10.15294/sji.v9i2.36885

[21] Li, S. S. (2020). An improved DBSCAN algorithm based on the neighbor similarity and fast nearest neighbor query. IEEE Access, 8. https://doi.org/10.1109/ACCESS.2020.2972034

[22] Regilan, S., & Hema, L. K. (2024). Optimizing environmental monitoring in IoT: Integrating DBSCAN with genetic algorithms for enhanced clustering. International Journal of Computers and Applications, 46(1). https://doi.org/10.1080/1206212X.2023.2277966

[23] Ahmed, M. A., Baharin, H., & Nohuddin, P. N. E. (2020). Analysis of K-means, DBSCAN and OPTICS cluster algorithms on Al-Quran verses. International Journal of Advanced Computer Science and Applications, 11(8). https://doi.org/10.14569/IJACSA.2020.0110832

[24] Ferreira, J. M., et al. (2023). Spatiotemporal analysis of rabies outbreaks in herbivores in Tocantins as a subsidy for the detection of epizootics. Contribuciones a las Ciencias Sociales, 16(12). https://doi.org/10.55905/revconv.16n.12-057

[25] Moayedi, A., Abbaspour, R. A., & Chehreghan, A. (2019). An evaluation of the efficiency of similarity functions in density-based clustering of spatial trajectories. Annals of GIS, 25(4). https://doi.org/10.1080/19475683.2019.1679254

[26] Taghva, K., & Veni, R. (2010). Effects of similarity metrics on document clustering. In ITNG2010 – 7th International Conference on Information Technology: New Generations. https://doi.org/10.1109/ITNG.2010.65

[27] Karo, I. M. K., Khosuri, A., & Setiawan, R. (2021). Effects of distance measurement methods in K-nearest neighbor algorithm to select Indonesia smart card recipient. In 2021 International Conference on Data Science and Its Applications (ICoDSA). https://doi.org/10.1109/ICoDSA53588.2021.9617476

[28] Henderi, H. (2021). Comparison of Min-Max normalization and Z-score normalization in the K-nearest neighbor (kNN) algorithm to test the accuracy of types of breast cancer. International Journal of Informatics and Information Systems, 4(1). https://doi.org/10.47738/ijiis.v4i1.73

[29] Karo, I. M. K., & Hendriyana, H. (2022). Klasifikasi penderita diabetes menggunakan algoritma machine learning dan Z-score. Jurnal Teknologi Terpadu, 8(2), 94–99.

[30] Karo, I. M. K., Huda, A. F., & MaulanaAdhinugraha, K. (2018). A cluster validity for spatial clustering based on Davies-Bouldin index and polygon dissimilarity function. In Proceedings of the 2nd International Conference on Informatics and Computing (ICIC). https://doi.org/10.1109/IAC.2017.8280572

[31] Akbar, T., Tinungki, G. M., & Siswanto, S. (2023). Performance comparison of K-medoids and density-based spatial clustering of application with noise using silhouette coefficient test. BAREKENG: Jurnal Ilmu Matematika dan Terapan, 17(3), 1605–1616. https://doi.org/10.30598/barekengvol17iss3pp1605-1616