Data Science : Journal Of Computing And Applied Informatics

Journal homepage: https://jocai.usu.ac.id

Spatial Clustering Analysis of Stunting in North Sumatra Based on Environmental Factors Using K-Means Algorithm

Fanny Ramadhani*[1], Dian Septiana[1], Sisti Nadia Amalia[1], Putri Maulidina Fadilah[1] , Andy Satria[2]

[1] *Universitas Negeri Medan, Medan, 20221, Indonesia*

[2] *Universitas Dharmawangsa, Medan, 20115, Indonesia*

*Corresponding Author: fannyr@unimed.ac.id

ARTICLE INFO

Article history:
Received 4 July 2024
Revised 20 July 2024
Accepted 29 July 2024
Available online 15 July 2025

E-ISSN: 2580-829X
P-ISSN: 2580-6769

How to cite:
Ramadhani F., D. Septiana, S. S. Amalia, P. M. Fadilah, A. Satria," Spatial Clustering Analysis of Stunting in North Sumatra Based on Environmental Factors Using K-Means Algorithm", Data Science: Journal of Computing and Applied Informatics, vol. V9, no. 2 , Jul. 2025, doi: 10.32734/jocai.v9.i2-17179

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International.
http://doi.org/10.32734/jocai.v9.i2-17179

ABSTRACT

This research aims to analyze the spatial grouping of stunting events in North Sumatra based on environmental factors using the K-Means algorithm. The data used in this research includes the incidence of stunting, environmental factors (such as access to health services, living environment conditions, water use and sanitation), and spatial data (geographical coordinates). The data comes from Basic Health Research (RISKESDAS 2018, then processed and normalized. The elbow method and silhouette analysis are used to determine the optimal number of clusters, resulting in four different clusters. The application of the K-Means algorithm produces the following cluster characteristics: Cluster 1, with good environmental conditions and access to health services, shows low levels of stunting; Cluster 2, with moderate environmental conditions, shows moderate levels of stunting; Cluster 3, which is characterized by poor living conditions and limited access to health services, has levels high stunting; and Cluster 4, with varied environmental conditions but very limited access to health and sanitation services, also shows a high stunting rate. Validation using the Silhouette Coefficient produces an average score of 0.65 which indicates good clustering quality shows that environmental factors, access to health services, and sanitation conditions have a significant impact on the incidence of stunting. Based on these findings, policy and intervention recommendations are focused on Clusters 3 and 4, which have high stunting rates. The interventions carried out include increasing access and quality of nutrition, health services, sanitation conditions, economic empowerment, and health education.

Keyword: Stunting, Spatial Clustering, K-Means, Environmental Factors, Silhouette Coefficient

ABSTRAK

Penelitian ini bertujuan untuk menganalisis pengelompokan kejadian spasial stunting di Sumatera Utara berdasarkan faktor lingkungan dengan menggunakan algoritma K-Means. Data yang digunakan dalam penelitian ini meliputi kejadian stunting, faktor lingkungan (seperti Akses Terhadap Pelayanan Kesehatan, Kondisi Lingkungan Tempat Tinggal, Pemakaian air dan sanitasi), dan data spasial (koordinat geografis). Data tersebut bersumber dari Riset Kesehatan Dasar (RISKESDAS 2018, selanjutnya diolah dan dinormalisasi. Metode elbow dan Silhouette coefficient digunakan untuk menentukan jumlah cluster yang optimal, sehingga menghasilkan empat cluster yang berbeda. Penerapan algoritma K-Means menghasilkan karakteristik cluster sebagai berikut: Cluster 1, dengan kondisi lingkungan dan akses layanan kesehatan yang baik, menunjukkan tingkat stunting yang rendah; Cluster 2, dengan kondisi lingkungan moderat, menunjukkan tingkat stunting yang moderat; Cluster 3, yang ditandai dengan kondisi tempat tinggal yang buruk dan terbatasnya akses terhadap layanan kesehatan, memiliki tingkat stunting yang tinggi; dan Klaster 4, dengan kondisi lingkungan bervariasi namun akses layanan kesehatan dan sanitasi yang sangat terbatas, juga menunjukkan angka stunting yang tinggi. Validasi menggunakan Silhouette Coefficient menghasilkan skor rata-rata sebesar 0,65 yang menunjukkan kualitas clustering yang baik. Analisis menunjukkan bahwa faktor lingkungan, akses layanan kesehatan, dan kondisi sanitasi berdampak signifikan terhadap kejadian stunting. Berdasarkan temuan ini, rekomendasi kebijakan dan intervensi difokuskan pada Klaster 3 dan 4, yang memiliki angka stunting yang tinggi. Intervensi yang dilakukan meliputi peningkatan akses dan kualitas gizi, layanan kesehatan, kondisi sanitasi, pemberdayaan ekonomi, dan pendidikan kesehatan.

Keyword: Stunting, Spatial Clustering, K-Means, Faktor Lingkungan, Silhouette

1. Introduction

Stunting is a critical health indicator that affects the quality of life of children. It is caused by chronic malnutrition during the growth period and has serious impacts on physical and cognitive development[1]. According to WHO data (2014), around 162 million children under the age of five experience stunting [2]. In Indonesia, stunting is a major concern due to its significant effects on child health and development [3]. According to the Ministry of Health of the Republic of Indonesia (2016), stunting can occur while the fetus is still in the mother's womb and becomes apparent when the child is two years old[3] [4]. In anthropometric standards, the measurement results are at a threshold (Z-Score) of < -2 SD to -3 SD (short/stunting) and < -3 SD (very short/severe stunting). The nutritional status of stunting is based on the PB/U or TB/U index [5].

The use of machine learning algorithms, such as clustering, can provide a deeper understanding of spatial patterns and the relationship between family economic factors and stunting incidents [6] [7]. In the era of information technology, the machine learning offers advantages in processing and analyzing complex data [8] [9]. K-Means clustering is an unsupervised machine learning algorithm used for data clustering and pattern recognition [1]. It works by randomly selecting a few initial data points (k), then moving them around until the most ideal clustering is found [10]. By K-Means clustering algorithm, this research is expected to harness the power of machine learning, to identify spatial patterns and clusters of stunting incidents that may be difficult to observe using conventional methods.

Through a deep understanding of the spatial relationship between environmental factors and stunting incidents in North Sumatra, this study aims to understand stunting incidents in North Sumatra, focusing on spatial clustering analysis based on environmental factors using K-Means algorithm. This research is expected to provide more accurate and relevant information for policy planning and also assist policymakers and health practitioners in designing more effective intervention programs that align with the local economic conditions, thereby reducing stunting rates and improving the well-being of children in North Sumatra.

2. Method

This research is a descriptive and an analytical quantitative research. The descriptive quantitative approach is used to depict the spatial distribution of environmental factors and stunting occurrences in North Sumatra, while the analytical approach is employed to analyze the relationship between environmental factors and stunting occurrences using machine learning methods: K-means algorithm.

The data used in this study is secondary data, which is environmental conditions such as access to health services, residential environment conditions, water usage, and sanitation from RISKESDAS 2018. The research was conducted in North Sumatra by accessing various journals and relevant books on the research topic, as well as performing data processing and simulations.

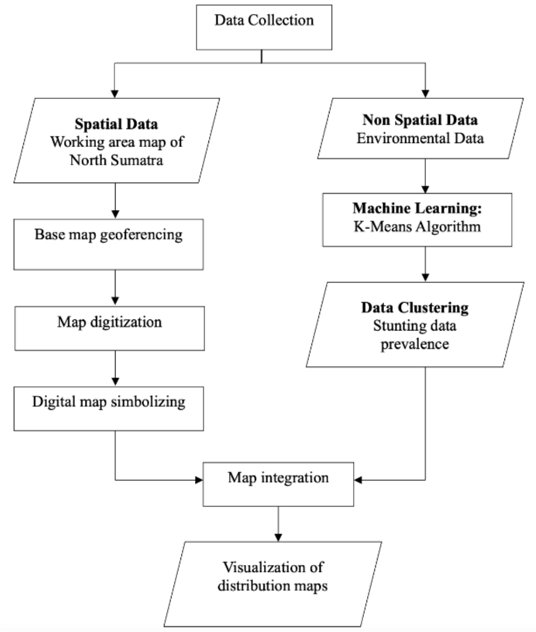


Figure 1. Research flowchart

2.1. Data Collection

In the data collection process, the researcher conducts a literature study to obtain data and information used in the problem discussion [11]. The data used is data with environmental factors (Access to Health Services, Living Environmental Conditions, Air Consumption and Sanitation), and spatial data (geographical coordinates) from the 2018 RISKESDAS survey conducted by the Ministry of Health of the Republic of Indonesia.

2.2. Data Processing

Data preprocessing starts from the data acquisition, acquisition and preparation stages, followed by cleaning to remove missing or invalid values. Feature selection is then carried out to reduce the dimensions of the data and improve the efficiency of the clustering algorithm. Only relevant data is used for analysis. Finally, data normalization ensures consistent scale of economic data, then proceed with the Spatial Data Integration and Mapping stage. At this stage, data on stunting incidents and environmental factors are combined with the geographical coordinates of each region or data points on the map of North Sumatra Province. Next, the data is visualized geographically to understand the initial distribution of stunting incidents and environmental conditions. then proceed to the Feature Selection stage. At this stage, environmental variables are selected which will be used as features in the clustering model. The variables used include access to health services, environmental conditions, water use and sanitation

2.3. Implementation Of K-Means Algorithm

After the data is obtained and its feasibility tested, data processing is carried out to classify the data (Spatial Data and Non-Spatial Data) so that the researcher can use it for clustering using machine learning methods K-Means Algorithm [8]. Run the K-Means clustering algorithm on the prepared data set to identify clusters based on environmental factors. then proceed with the resulting cluster analysis to understand the characteristics and distribution of stunting incidents. In this research, the elbow method was used to determine the optimal number of clusters. and silhouette coefficients are used to assess the quality of clusters formed by clustering algorithms, including K-Means :

2.3.1. Initialization

Choose the number of clusters *k* and randomly select *k* initial centroids from the dataset.

2.3.2. Assignment Step

For each data point, calculate the distance between the data point and each centroid. Assign each data point to the nearest centroid based on the Euclidean distance. This forms *k* clusters.

2.3.3. Update Step

For each cluster, calculate the new centroid by computing the mean of all data points assigned to that cluster and update the centroid to this new mean position.

2.3.4. *Repeat*

Repeat steps 2 and 3 until the centroids do not change significantly between iterations (i.e., the changes are below a small threshold) and a maximum number of iterations is reached.

2.3.5. *Convergence*

The algorithm stops when the centroids have stabilized and the clusters do not change significantly with further iterations.

By following these steps, the K-means algorithm partitions the data into *k* clusters with the goal of minimizing the variance within each cluster. To determine how well the clustering has performed in grouping the data, the silhouette procedure is used. The silhouette procedure provides a graphical representation of how well each object lies within its cluster, often visualized using a silhouette plot [12]. This plot helps in identifying clusters that are well separated and those that might be overlapping or poorly defined. The silhouette value *s(i)* for each point *i* is given by:

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \tag{1}$$

where *a(i)* is the average separation (cohesion) between data point *i* and every other point in the same cluster. The lowest average distance (separation) between data point *i* and all other points in any other cluster is denoted by *b(i)*.

Table 2 presents the categories of silhouette coefficient based on Kaufman and Rousseeuw [13].

Table 2. The categories of silhouette coefficient

Silhouette Coefficient (SC)	Proposed interpretation
0.71 – 1.00	Strong structure
0.51 – 0.70	Medium Structure
0.26 – 0.50	Weak Structure
≤ 0.25	No Structure

2.4. *System Testing*

Complex testing involves a series of rigorous assessments designed to identify system weaknesses that basic testing may miss. The objective is to identify weaknesses, errors or problems that could weaken the system's performance, thus ensuring better stability and reliability. By implementing corrective measures, complex testing aims to minimize or eliminate these errors, ensuring that the system functions according to design specifications and provides accurate results across a wide range of operational scenarios. The main objective of this system testing approach is to produce accurate and reliable results.

2.5. *Visualization of Stunting Distribution Map in North Sumatra*

Map visualization is used to analyze and display geographic data, presenting it in different types of maps. Information is more likely to be revealed through visualization, making it clearer and more intuitive. It allows for the observation of data distribution or proportions in each area, thus making it easier for people to engage with the information and make better decisions [14].

3. Result and Discussion

The research process starts from reading data to visualizing distribution maps using the Python programming language. The spatial grouping procedure based on environmental factors on stunting incidents in North Sumatra was carried out using the K-Means algorithm. In this research, the elbow method was used to determine the optimal number of clusters. The results of the clustering process produce 4 clusters or groups of data. The elbow indicates that the optimal number of clusters is four. the distortion score decreases significantly when the k value (number of clusters) increases from 2 to 4. In clustering analysis, the distortion score, also known as inertia or within-cluster sum of squares, is an important statistic for determining cluster coherence. It calculates the sum of squared distances from each data point to the nearest centroid, which indicates the cluster center. A lower distortion score indicates clusters that are more densely packed and internally consistent, indicating a greater clustering of data points around their respective centroids. The "elbow" point on the curve indicating the optimal k value is seen at k = 4. The distortion score is 27434.547, quite low compared to the k value other, as shown in the following diagram:

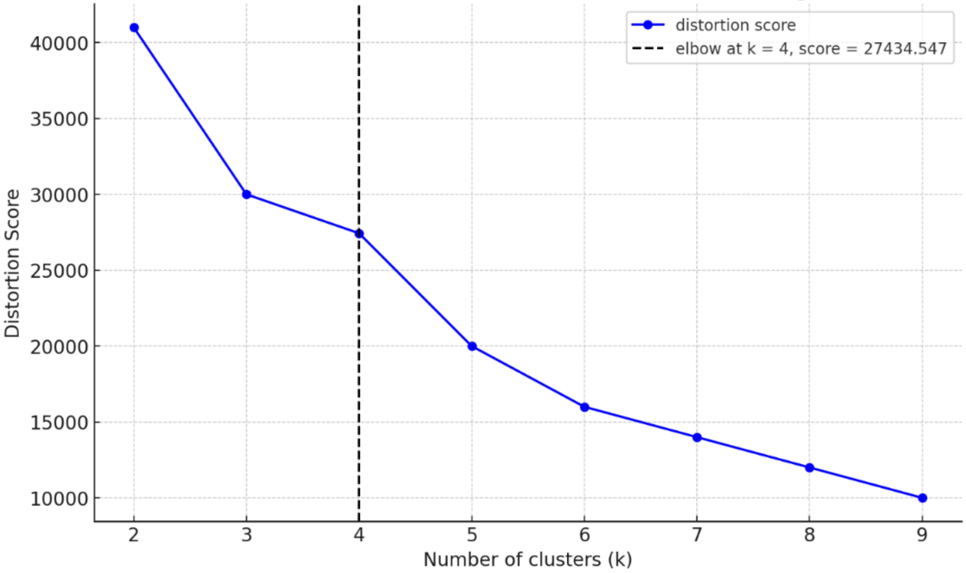
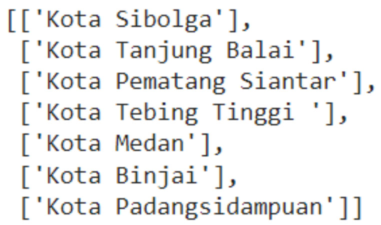


Figure 2. Distortion Score Elbow for KMeans Clustering

The blue line represents the distortion score for different values of k. It shows how the distortion decreases as the number of clusters increases. The plot highlights an "elbow" point at $k = 4$ with a distortion score of 27434.547. The elbow point is where the rate of decrease in distortion slows down significantly, indicating a suitable number of clusters.

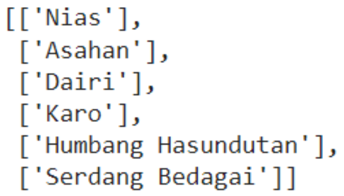
After the optimal number of clusters or k value has been determined as 4, the next step is to carry out the clustering process using the k-means algorithm. This analysis will involve examining data from various variables including longitude, latitude, access to health services, living environmental conditions, water use and sanitation. Findings from the clustering process categorized areas into four distinct groups. The clusters are identified as cluster 1 (highlighted in green), cluster 2 (highlighted in blue), cluster 3 (highlighted in orange), and cluster 4 (highlighted in red). The silhouette coefficient score is 0.65 which is a moderate or medium structure. This indicates a good quality of clustering, with clear separation between clusters and meaningful grouping within clusters.This means that every good object is placed in its cluster using the K-Means algorithm [13].



```
[['Kota Sibolga'],
 ['Kota Tanjung Balai'],
 ['Kota Pematang Siantar'],
 ['Kota Tebing Tinggi '],
 ['Kota Medan'],
 ['Kota Binjai'],
 ['Kota Padangsidampuan']]
```

Figure 3. Cluster 1

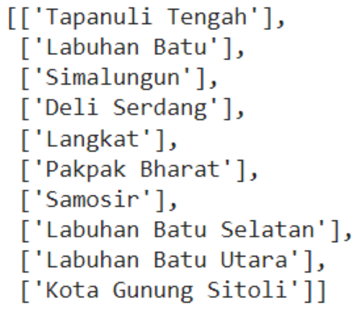
Cluster 1 is the region where stunting prevalence is below 20% statistically, several factors contribute to this lower rate. These include good access to healthcare services, where residents can easily reach health facilities such as clinics and hospitals, and benefit from comprehensive maternal and child health programs. The living conditions in these areas are adequate, with proper housing, good ventilation, and cleanliness, which reduces the risk of infections and diseases. Additionally, access to clean and sufficient water supplies ensures that residents have enough safe water for drinking, cooking, and hygiene, which is crucial for preventing illnesses that could hinder children's growth. Furthermore, good sanitation practices, such as having proper toilets, effective waste management systems, and maintaining personal hygiene, also play a significant role in reducing stunting rates. Examples of cities in North Sumatra that have low stunting prevalence due to these factors include Sibolga, Tanjung Balai, Pematang Siantar, Tebing Tinggi, Medan, Binjai, and Padang Sidempuan [15].



```
[['Nias'],
 ['Asahan'],
 ['Dairi'],
 ['Karo'],
 ['Humbang Hasundutan'],
 ['Serdang Bedagai']]
```

Figure 4. Cluster 2

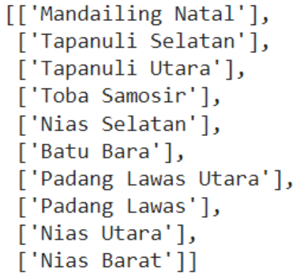
Cluster 2 is the areas with a stunting prevalence of 20% to 30% statistically, access to healthcare services is moderate, which means that residents have some but not complete access to necessary medical facilities and services. The living conditions in these locations are moderately adequate, indicating that, while housing and environmental conditions are not ideal, they are also not the worst. Water consumption is adequate, ensuring that the population has a sufficient supply of clean water for drinking, cooking, and hygiene, although there may be problems with water quality or availability. Sanitation in these places is adequate, which means that basic sanitation needs are satisfied; however, further changes are needed to totally prevent health risks caused by poor hygiene and waste management. Examples of similar regions in North Sumatra include Asahan Regency, Asahan, Dairi, Karo, Humbang hasundutan and Serdang bedagai.



```
[['Tapanuli Tengah'],
 ['Labuhan Batu'],
 ['Simalungun'],
 ['Deli Serdang'],
 ['Langkat'],
 ['Pakpak Bharat'],
 ['Samosir'],
 ['Labuhan Batu Selatan'],
 ['Labuhan Batu Utara'],
 ['Kota Gunung Sitoli']]
```

Figure 5. Cluster 3

Cluster 3 identifies the access to healthcare is difficult in locations where stunting prevalence is between 30% and 40%, indicating that individuals face major barriers to receiving medical facilities and services. These places have poor living circumstances, including inadequate housing and environmental norms, which contribute to health problems. Water usage is insufficient, which means that the population has limited access to clean and safe water for drinking, cooking, and hygiene. Sanitation is also poor, with insufficient waste management and hygiene standards, increasing the risk of sickness and infection. Pakpak Bharat and Tapanuli Tengah Regency are two such places in North Sumatra where these negative variables contribute to a high frequency of stunting.



```
[['Mandailing Natal'],
 ['Tapanuli Selatan'],
 ['Tapanuli Utara'],
 ['Toba Samosir'],
 ['Nias Selatan'],
 ['Batu Bara'],
 ['Padang Lawas Utara'],
 ['Padang Lawas'],
 ['Nias Utara'],
 ['Nias Barat']]
```

Figure 6. Cluster 4

Cluster 4 is the zone where access to healthcare is extremely difficult in places where the prevalence of stunting exceeds 40%, posing significant hurdles for inhabitants in obtaining any medical facilities or services. Living conditions are deplorable, with substandard housing and bad environmental conditions that have a negative influence on health. Water consumption is extremely low, indicating a significant shortage of access to clean and safe water for drinking, cooking, and hygiene. Sanitation is also exceedingly poor, with little waste management and hygiene standards, posing a high risk of sickness and infection. Nias Regency, in North Sumatra, is one such place where severe inadequacies lead to the very high frequency of stunting.

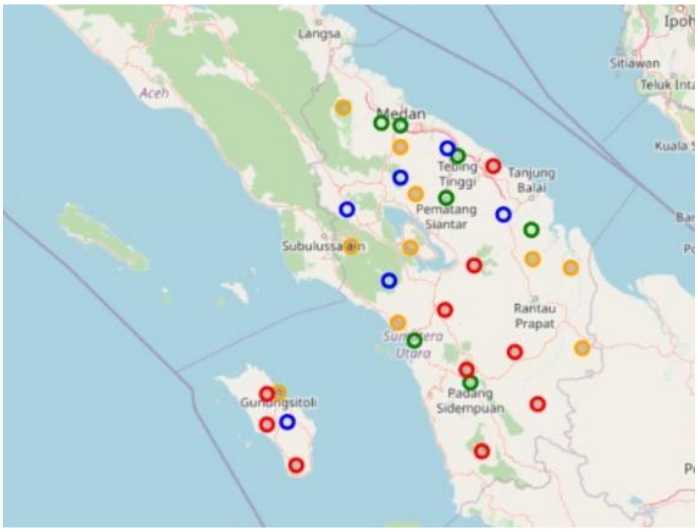


Figure 7 Stunting distribution map in North Sumatra

Figure 6 shows the results of the integrated map combined with cluster data, where each cluster is given a color. The area in red is cluster 4 which has the highest cases of stunting. The stunting cases in cluster 4 mean that access to health and sanitation services is very limited. Residential environmental conditions and water use in the area also need attention. The orange area is cluster 3 which has high cases of stunting. Apart from that, there is also limited access to health services, sanitation, water use, and residential environmental conditions that require attention. The blue area is cluster 2 which has moderate stunting cases. Although residential environmental conditions vary, cluster 2 has good access to health services, sanitation and water use. Meanwhile, areas colored green or cluster 1 have good residential environmental conditions, good access to health services, good sanitation and good water use. With good environmental factors, stunting cases in cluster 1 are relatively low [15].

4. Conclusion

Clustering analysis using the K-Means algorithm succeeded in identifying four main clusters based on environmental factors and the incidence of stunting. These clusters show significant variations in access to health services, environmental conditions, water use and sanitation. Clusters with high rates of stunting are mainly located in rural and less developed areas, where the level of environmental cleanliness is low, and access to health and sanitation services is limited. In contrast, clusters with a low incidence of stunting are more often found in urban and more developed areas. Cluster validation using the Silhouette Coefficient with an average score of 0.65 shows that the quality of clustering is quite good, with clear visibility between clusters and meaningful grouping within clusters. Factors such as access to health services, environmental conditions, air usage conditions and sanitation conditions are proven to have a significant influence on the incidence of stunting. Areas with higher incomes and better access to education and health services tend to have lower stunting rates.

Acknowledgements

A deep gratitude to LPPM Universitas Negeri Medan for the financial support provided through research funds. These funds is very important in supporting this research.

References

[1] W. Hadikurniawati, K. D. Hartomo, And I. Sembiring, "Spatial Clustering of Child Malnutrition in Central Java: A Comparative Analysis Using K-Means And DBSCAN," In *Proceedings: ICMERALDA 2023 - International Conference on Modeling And E-Information Research, Artificial Learning And Digital Applications*, Institute of Electrical and Electronics Engineers Inc., 2023, Pp. 242–247. Doi: 10.1109/ICMERALDA60125.2023.10458202.

[2] T. Vaivada, N. Akseer, S. Akseer, A. Somaskandan, M. Stefopulos, And Z. A. Bhutta, "Stunting in Childhood: An Overview of Global Burden, Trends, Determinants, And Drivers of Decline," *Am J Clin Nutr*, Vol. 112, Pp. 777S-791S, Sep. 2020, Doi: 10.1093/AJCN/NQAA159.

[3] A. A. Azis and A. Aswi, "Spatial Clustering of Stunting Cases in Indonesia: A Bayesian Approach," *Communications in Mathematical Biology and Neuroscience*, Vol. 2023, 2023, Doi: 10.28919/Cmbn/7898.

[4] S. Supadmi *Et Al.*, "Factor Related to Stunting of Children Under Two Years with Working Mothers in Indonesia," *Clin Epidemiol Glob Health*, Vol. 26, P. 101538, Mar. 2024, Doi: 10.1016/J.CEGH.2024.101538.

[5] A. N. W. S. Saimu, "Penanganan Resiko Stunting Berbasis Data Tingkat Kecamatan Mawasangka Tengah Kabupaten Buton Tengah," *Jurnal Inovasi Penelitian*, Vol. 4, No. 1, Pp. 75–87, 2023.

[6] A. Yusuf, "K-Means Clustering Based on Distance Measures: Stunting Prevalence Clustering in South Kalimantan," In *2022 5th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, Pp. 706–710. Doi: 10.1109/ISRITI56927.2022.10052925.

[7] H. Sulastri, H. Mubarok, And S. S. Iasha, "Implementasi Algoritma Machine Learning Untuk Penentuan Cluster Status Gizi Balita," *Jurnal Rekayasa Teknologi Informasi (JURTI)*, Vol. 5, No. 2, P. 184, Dec. 2021, Doi: 10.30872/Jurti.V5i2.6779.

[8] I. P. Sari, Al-Khowarizmi, F. Ramadhani, A. Satria, And O. K. Sulaiman, "Leukocoria Identification: A 5-Fold Cross Validation CNN And Adaboost Hybrid Approach," *6th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2023 - Proceeding*, Pp. 486–491, 2023, Doi: 10.1109/ISRITI60336.2023.10467242.

[9] F. Ramadhani, M. Zarlis, And S. Suwilo, "Improve BIRCH Algorithm for Big Data Clustering," In *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jan. 2020. Doi: 10.1088/1757-899X/725/1/012090.

[10] K. P. Sinaga and M. S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, Vol. 8, Pp. 80716–80727, 2020, Doi: 10.1109/ACCESS.2020.2988796.

[11] A. Satria, O. S. Sitompul, And H. Mawengkang, "5-Fold Cross Validation on Supporting K-Nearest Neighbour Accuration of Making Consimilar Symptoms Disease Classification," In *Proceedings - 2nd International Conference on Computer Science and Engineering: The Effects of The Digital World After Pandemic (EDWAP), IC2SE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. Doi: 10.1109/IC2SE52832.2021.9792094.

[12] P. Thongnim, E. Charoenwanit, And T. Phukseng, "Cluster Quality in Agriculture: Assessing GDP And Harvest Patterns in Asia and Europe with K-Means and Silhouette Scores," *2023 7th International Conference on Electronics, Materials Engineering and Nano-Technology, Iementech 2023*, 2023, Doi: 10.1109/IEMENTECH60402.2023.10423469.

[13] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. Wiley, 1990. Doi: 10.1002/9780470316801.

[14] D. Danila, I. D. Pawa, A. Choiruni, A. Wijayanti, "Geospatial Analysis Pada Prevalensi Stunting di Kabupaten Manggarai," *Berita Kedokteran Masyarakat*, Vol. 34, No. 11, 2018.

[15] M. De Onis, E. Borghi, M. Arimond, P. Webb, T. Croft, K. Saha, L. M. De-Regil, F. Thuita, R. Heidkamp, J. Krasevec, C. Hayashi, "Prevalence Thresholds for Wasting, Overweight and Stunting In Children Under 5 Years," *Public Health Nutr*, Vol. 22, No. 1, Pp. 175–179, Jan. 2019, Doi: 10.1017/S1368980018002434.