Cluster Analysis Using the Ward Algorithm for Grouping Regency / City in Central Java Province Based on Poverty Indicators 2023

Agung Supriyono[(1)], Atika Nurani Ambarwati [(2)]

Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang.

Jl. Prof. Dr. Hamka Km 01 No. 17 Tambakaji Ngaliyan

e-mail: agungsupriyono2004@gmail.com[(1)], atika.nurani@gmail.com[(2)]

ABSTRAK

Kemiskinan masih menjadi isu utama di Provinsi Jawa Tengah, dengan tingkat kemiskinan yang tercatat sebesar 10,23% pada tahun 2023. Penanggulangan kemiskinan yang belum merata di berbagai wilayah turut menambah kompleksitas permasalahan ini. Penelitian ini bertujuan untuk mengklasifikasikan kabupaten/kota di Jawa Tengah berdasarkan indikator-indikator kemiskinan, dengan menggunakan pendekatan analisis klaster hierarki melalui algoritma Ward. Data yang digunakan merupakan data sekunder dari Badan Pusat Statistik (BPS), yang mencakup enam variabel yang memengaruhi kemiskinan. Salah satunya adalah persentase penduduk miskin (X1), yang merepresentasikan proporsi penduduk di bawah garis kemiskinan dan menjadi indikator penting dalam menilai kesejahteraan suatu daerah. Hasil penelitian menunjukkan bahwa penerapan analisis klaster hierarki melalui algoritma Ward menghasilkan empat klaster kabupaten/kota berdasarkan indikator kemiskinan tahun 2023, masing-masing dengan karakteristik yang berbeda. Klaster pertama terdiri dari daerah-daerah dengan tingkat kemiskinan tinggi serta keterbatasan dalam aspek pendidikan, kesehatan, dan infrastruktur ekonomi. Wilayah dalam klaster ini memerlukan perhatian khusus dalam penyusunan kebijakan. Berdasarkan karakteristik yang diperoleh dari masing-masing klaster, temuan ini dapat dimanfaatkan untuk merancang kebijakan pembangunan yang lebih terarah dan berbasis data, seperti pengalokasian anggaran, program penanggulangan kemiskinan, serta peningkatan akses terhadap layanan dasar sesuai dengan kondisi tiap klaster.

Kata kunci : Kemiskinan, Analisis Hierarki, Pendekatan Ward, Jawa Tengah

ABSTRACT

Poverty is still a major issue in Central Java Province, with the poverty rate recorded at 10.23% in 2023. Uneven poverty reduction in various regions adds to the complexity of this problem. This study aims to classify districts/cities in Central Java based on poverty indicators, using a hierarchical cluster analysis approach through Ward's algorithm. The data used is secondary data from the Central Bureau of Statistics (BPS), which includes six variables that affect poverty. One them is the percentage poor people (X1), which represents the proportion of the population below the poverty line and is an important indicator in assessing the welfare of a region. The results show that the application of hierarchical cluster analysis through Ward's algorithm produces four clusters of districts/cities based on poverty indicators in 2023, each with different characteristics. The first cluster consists regions with high poverty rates and limitations in education, health, and economic infrastructure. Regions in this cluster require special attention in policy formulation. Based on the characteristics obtained from each cluster, these findings can be used to design more targeted and data-based development policies, such as budget allocation, poverty reduction programs, and improving access to basic services according to the conditions each cluster.

Keywords: *Poverty, hierarchical analysis, Ward method, Central Java*

INTRODUCTION

Poverty is one of the challenges faced by various countries, especially in developing countries, namely Indonesia [1]. The problem of poverty has become a priority that must be addressed immediately and needs more attention. Because it jeopardizes the economic development and social stability of a country [2]. Poverty is a major challenge faced by developing countries such as Indonesia because it impacts economic development and social stability. According to BPS data in 2023, 9.36% or around 25.9 million Indonesians still live in poverty, struggling to access basic needs such as food, clean water, sanitation, housing, education, and health services. Java, as the most populous region, accounts for the largest number of poor people, at 13.94 million or more than half of the national total. Among the provinces in Java, Central Java recorded the highest poverty percentage at 10.23%, exceeding the national average. Several regions in the province, such as Kebumen, Brebes, and Wonosobo, are still classified as areas of extreme poverty, reflecting the welfare gap between urban and rural areas and the unevenness of poverty reduction efforts.

Some previous studies show that the percentage of poor people, population, open unemployment rate, expected years of schooling, per capita expenditure, and life expectancy have a significant effect on poverty [3],[4]. Therefore, this study uses these variables as variables to group districts/cities based on the factors that influence poverty. This grouping aims to make it easier for the government to design a more targeted poverty alleviation strategy. To facilitate decision making for the Central Java Provincial government in overcoming this poverty problem, of course, the right approach is needed to map the regions according to their characteristics.

Ward's approach was chosen in this study because it can reduce the number of error squares in the clustering process, thus creating clusters with high internal homogeneity [5]. Research conducted by [6] with the title "Analisis Cluster dengan Average Linkage Method dan Ward's Method pada Pengelompokan Kabupaten/Kota Di Provinsi Sumatera Utara Berdasarkan Indikator Indeks pembangunan Manusia Tahun 2022" stated that the ward's method algorithm proved superior because it has a lower Sum of Squared Errors (SSE) value, indicating that this method is able to form denser and more homogeneous clusters. In addition, evaluation of the dendrogram shows that the ward's method produces more uniform and consistent clusters. This approach aims to minimize the variance within clusters, Ward's method produces more homogeneous and compact groups, allowing the identification of areas with high poverty rates that also have limitations in education, health, and infrastructure. The results of this clustering are expected to provide a more structured picture of the condition of the region so that it can be used as a basis for formulating more targeted development policies [7].

ALGORITHMS

This research uses a quantitative Algorithm with a hierarchical cluster analysis Algorithm using the Ward Algorithm. The steps taken in this research include:

Data Source

The type of data used in this study are secondary data, which comes from the Central Java Provincial Statistics Agency Publication in 2023.

Research Variables

The variables used in this study only use predictor variables (X). Predictive variables are factors that have the potential to have an impact on Poverty in Central Java Province. The following are the variables:

Table 1. Risearch Variables

Variables	Describtions	Scale
X1	Percentage of Poor Population	Ratio
X2	Total Population	Ratio
X3	Open Unemployment Rate	Ratio
X4	Expected Years of Schooling	Ratio
X5	Per capita expenditure	Ratio
X6	Life Expectancy Rate	Ratio

Data Processing

The Algorithm used in this research uses the literature Algorithm or literature study with the following steps [8]:

1. Conduct a Descriptive Analysis of the Poverty Condition in Central Java Province in 2023.
2. Test assumptions in Cluster, In clustering, there are two assumptions that need to be met, namely:
   a. The sample represents the population, the sample used in cluster analysis must be able to represent the population to be explained, because this analysis is said to be good if the sample is representative. To find out that the sample can represent the population can be seen from the Kaiser Mayer Olkin (KMO) value, if the KMO value obtained is> 0.5 then the sample is sufficient to be analyzed [9].
   b. Multicollinearity, used to determine the existence of a linear relationship between independent variables. The VIF (Variance Inflation Factor) value can be used to determine the presence or absence of multicollinearity. If the VIF value < 10.00, then there are no symptoms of multicollinearity and vice versa [10], [11].

   $$VIF_i = \frac{1}{1-R_i^2} \quad (1)$$

   Description:
   R^2 = Koefisien determinan
   i = The i-th independent variable
3. Measuring the similarity between objects with Euclidean Distance, Euclid distance is used to measure the distance from the data object to the cluster center [12].

   $$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2)$$

   Description:
   d_{ij} = Euclid distance of i-th data object and j-th data object
   p = Number of parameters used
   x_{ik} = The i-th data object on the k-th variable
   x_{jk} = j-th data object on the k-th variable

4. The clustering process in the ward method uses Sum Square Error (SSE) to measure the homogeneity between two objects based on the least amount of squared error and also to measure the quality of the cluster [13].

   $$SSE = \sum_{j=1}^{p}(\sum_{i=1}^{n} x_{ij}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_{ij})^2) \quad (3)$$

   Description:

   x_{ij} = The value of the j-th variable for the i-th object

   p = Number of variables measured

   n = Number of objects in the formed cluster
5. Determining the number of clusters, aims to group data into internally homogeneous clusters by minimizing variation or squared error within clusters [9].
6. Interpretation of cluster results, at the interpretation stage the average (centroid) is used. Through this centroid value, to explain the purpose or characteristics of the cluster.
7. Conclusion

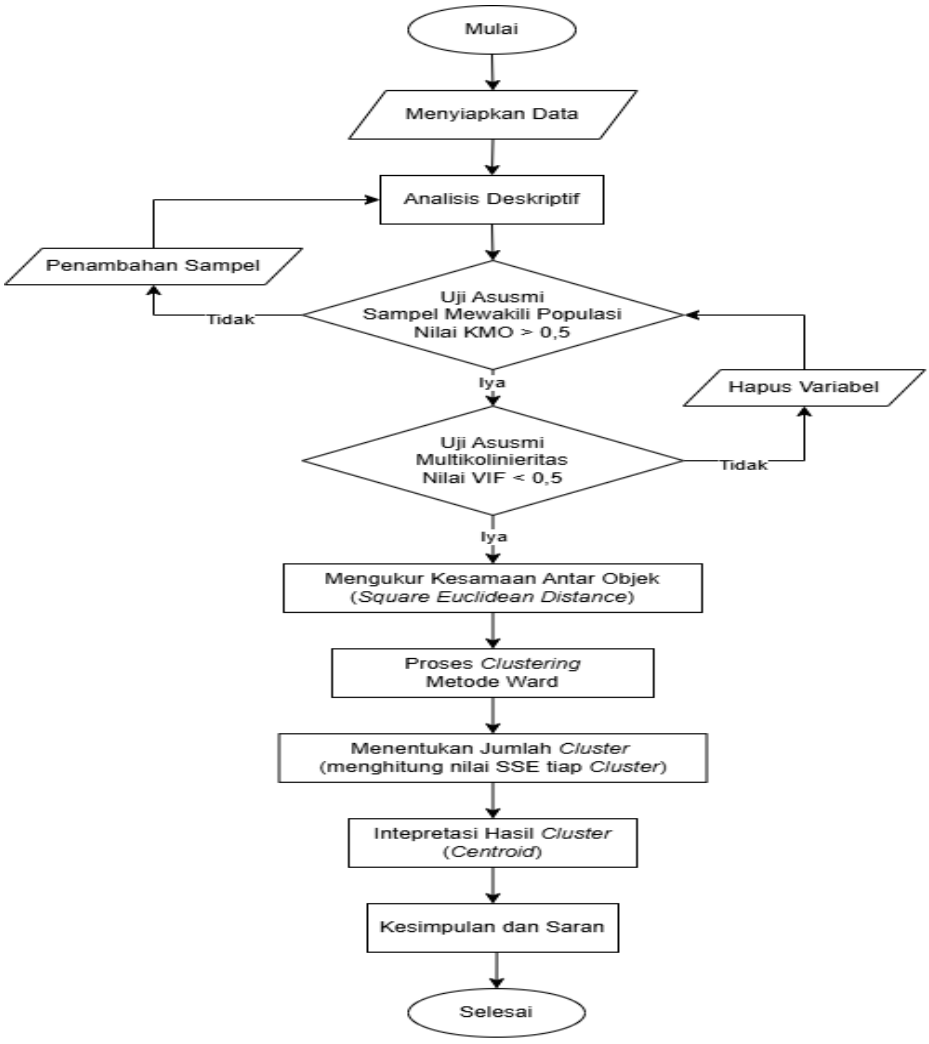

Figure 1. Flow chart

RESULTS AND DISCUSSION

Descriptive Analysis

Descriptive analysis was conducted to determine the general description of Poverty conditions in Central Java Province and the characteristics of each indicator or variable used.

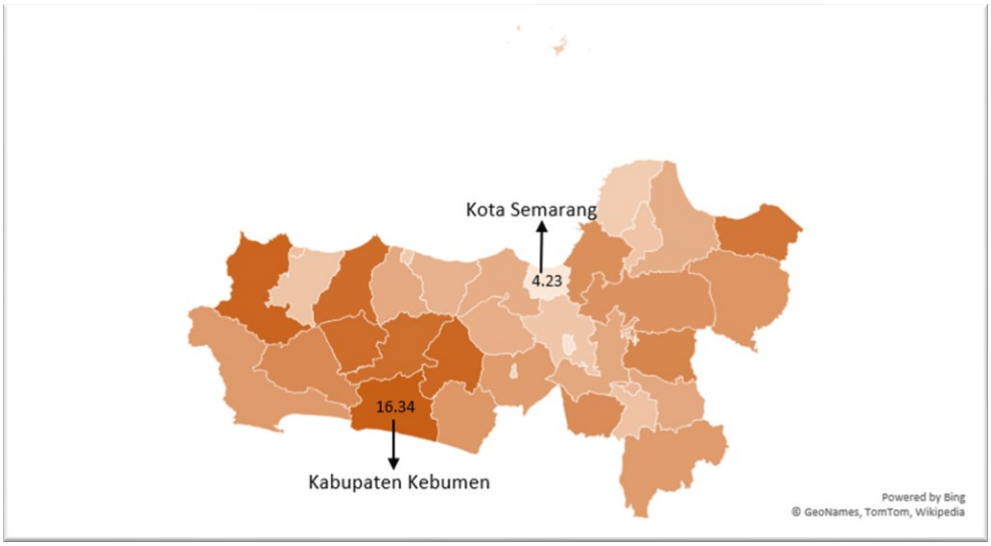

Figure 2. District/City Poverty Rates in Central Java 2023

It is also known that the lowest poverty condition is in Semarang City with a value of 4.23%, while the highest poverty condition is in Kebumen district with a value of 16.34%, which shows that the poverty rate in Central Java experiences significant inequality. This shows that poverty reduction has not been well-targeted in areas with high poverty rates.

Assumptions in Cluster Analysis

1. The sample is representative of the population
   To find out that the sample can represent the population can be seen from the Kaiser Mayer Olkin (KMO) value, if the KMO value obtained is > 0.5 then the assumption is fulfilled.

Table 2. Kaiser Mayer Olkin value

KMO and Bartlett's Test	
Kaiser-Meyer-Olkin	,698

   Based on table 2, the KMO value is 0.69 where the KMO value of 0.69 > 0.5, it can be concluded that the sample can represent the population and can be used for further analysis.
2. Multicollinearity
   The coefficient of determination of each variable is obtained by making the variable that you want to know the coefficient of determination as the dependent variable and the remaining variables as independent variables. If the Variance Inflation Factor (VIF) value is < 10.00, then the assumption is fulfilled.

Table 3. Varians Inflation Factor value

Variable	VIF value
Percentage of Poor Population	1,000
Total Population	1,608
Open Unemployment Rate	3,713
Expected Years of Schooling	2,424
Per capita expenditure	2,868
Life Expectancy Rate	3,678

Since the VIF value of each variable is less than 10.00 in the previous table, it can be said that there are no variables that show signs of multicollinearity.

Measuring similarity between objects with Euclidean Distance

In calculating the similarity of objects (districts), the Algorithm used is the Euclidean distance, there are 35 districts that will be calculated for similarity. Calculation of the Euclidean distance matrix using Equation (2). An example of calculating the distance between X1 and X2 is as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{35} (x_{ik} - x_{2k})^2} = 3{,}112$$

Using the same calculation, the distance between City 1 and City 3 and so on is also obtained. The smaller the distance value between two observations, the more similar the two observations are.

Ward Algorithm Cluster Process

The clustering process in this study groups regencies/cities in Central Java Province based on their characteristics using the ward Algorithm. The results of clustering using the ward Algorithm can be seen with a dendrogram diagram, as follows:

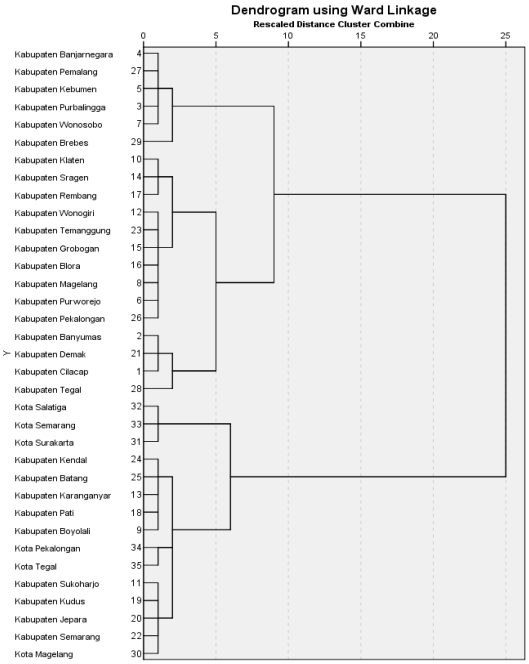

Figure 3. Dendrogram Diagram Results

Based on the dendrogram analysis, there are three possible numbers of clusters that can be formed, namely four, three, and two clusters. In the formation of four clusters, we get a first cluster consisting of 6 districts, a second cluster with 14 districts, a third cluster that includes 3 districts, and a fourth cluster consisting of 12 districts. If we group into three clusters, the first cluster will still consist of 6 districts, the second cluster of 14 districts, while the third cluster will include 15 districts. Finally, if only two clusters were formed, the first cluster would include 20 districts and the second cluster would consist of 15 districts.

Determining the Number of Clusters

In this study, the number of clusters is determined based on the Sum of Squared Errors (SSE) value. The SSE value shows how far each data in the group is from the center of the group. The smaller the SSE value in a cluster, the more homogeneous the objects in the cluster are, so the quality of the clustering is getting better.

Table 4. Sum Squared Error value

Cluster	SSE Value
1	3112610.36
2	13170017.30
3	321280.27
4	13936095.40
1	3112610.36
2	13170017.30
3	48419657.81
1	18952231.33
2	48419657.81

From table 4, it can be seen that the lowest Sum squared error cluster value of each cluster in the number of clusters formed is in the grouping with 4 clusters. So it can be concluded that in this study the object is classified with 4 groups. The members of each cluster formed are as follows:

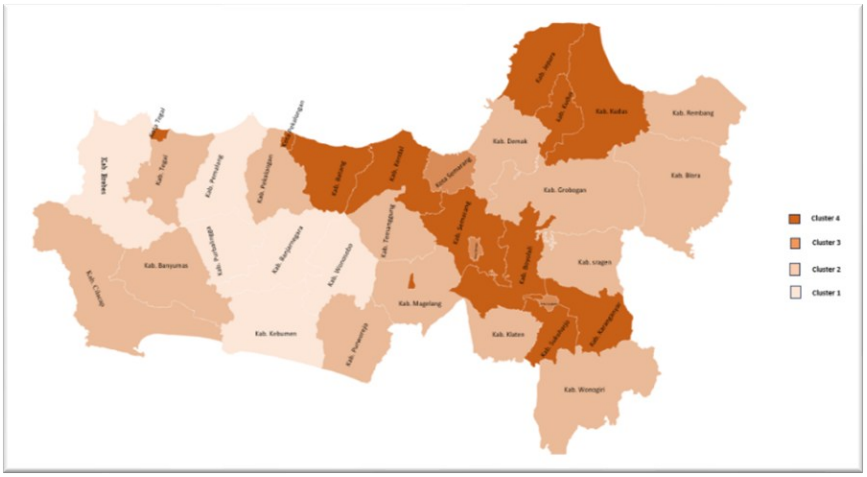

Figure 4. Cluster Interpretation Results

Interpretation of Cluster Visualization Results

Interpreting cluster results involves the centroid value, which is the average value of each variable on the objects in the cluster.

Table 5. Cluster Interpretation Result

Variable	Cluster 1	Cluster 2	Cluster 3	Clusster 4
X1	15,43	11,25	4,77	8,03
X2	1,31	1,21	0,8	0,83
X3	12,24	12,81	15,29	13,16
X4	6,24	4,48	5,04	4,57
X5	10,51	11,31	16,31	12,54
X6	72,94	74,92	77,82	76,08

Based on table 5, the characteristics (average) of each group formed on the Poverty indicator in Central Java in 2023 are obtained.

- Cluster 1 consists of areas with high poverty rates, high open unemployment, and relatively low expected years of schooling. These areas are generally rural areas with limited access to basic services and formal employment. The Central Java BPS report (2023) confirms that the development imbalance between coastal and mountainous areas is still the main cause of the high poverty rate in this region.
- Cluster 2 includes areas with moderate poverty, large populations, but lower unemployment than Cluster 1. This may indicate that informal economic activity is quite developed, but has not been able to lift per capita expenditure significantly. The study [14] shows that limitations in the quality of education and access to MSME capital are still obstacles to poverty alleviation in these areas.
- Cluster 3, which consists only of large cities, shows the lowest poverty indicators and the highest life expectancy, reflecting better access to health services, education, and formal sector employment opportunities. However, the high unemployment in this cluster suggests that rapid urbanization is not always followed by adequate employment. This is consistent with the World Bank's (2022) finding that Indonesia's major cities face the challenge of "jobless growth".
- Cluster 4 consists of areas with low poverty rates and high expected years of schooling, but significant unemployment. This may indicate a mismatch between education and local labor market needs. According to the Ministry of National Development Planning/Bappenas (2021), vocational education and job training reforms are urgently needed in these areas to improve labor productivity.

CONCLUSIONS

Based on the analysis and discussion presented in the previous chapter, the following conclusions can be drawn:

1. Based on the results of the hierarchical cluster analysis research that applies Ward's Algorithm, the grouping of districts/cities in Central Java Province based on poverty indicators in 2023 resulted in 4 (four) clusters formed according to their respective characteristics.
2. Based on the dendrogram analysis, four distinct clusters were identified. The first cluster includes six districts/cities, while the second cluster consists of fourteen districts/cities. The

third cluster consists of three kabupaten/kota, and the fourth cluster consists of twelve kabupaten/kota. The following are the details of each of these clusters.

a. Cluster 1 includes Banjarnegara Regency, Pemalang Regency, Kenumen Regency, Purbalingga Regency, Wonosobo Regency, Brebes Regency. Cluster 1 requires special attention from the government in the form of direct interventions, such as increasing access to education and job creation. Poverty alleviation programs based on economic empowerment can be implemented in this region.
b. Cluster 2 includes Klaten Regency, Sragen Regency, Rembang Regency, Wonogiri Regency, Temanggung Regency, Grobogan Regency, Blora Regency, Magelang Regency, Purworejo, Pekalongan Regency, Banyumas Regency, Demak Regency, Cilacap Regency, Tegal Regency. Cluster 2 shows that poverty in some regions is still high despite lower unemployment rates. This indicates the need to improve access to business capital and job skills training.
c. Cluster 3 includes Surakarta City, Salatiga City, Semarang City. Cluster 3 reflects better socioeconomic conditions than the other clusters, but still faces challenges in unemployment. Therefore, strengthening the industrial sector and investing in productive labor can be the main solutions.
d. Cluster 4 includes Kendal Regency, Batang Regency, Karanganyar Regency, Pati Regency, Boyolali Regency, Pekalongan City, Tegal City, Sukoharjo Regency, Kudus Regency, Jepara Regency, Semarang Regency, Tegal City. Cluster 4 has a low poverty rate, but still faces challenges in terms of education. Therefore, policies that focus more on improving the quality of education and access to scholarships are needed.

Based on the characteristics obtained from each cluster, these findings can be used to design more targeted and data-driven development policies, such as budget allocation, poverty reduction programs, and improving access to basic services in accordance with the conditions of each cluster.

SUGGEST

Suggestions for future research suggest that clustering be done by considering the spatial aspect or location of the region. For example, geographically close regions may have similar poverty conditions. This can be analyzed using the spatial cluster method, so that it can be used to see whether poverty is evenly distributed or concentrated in certain areas. By adding a spatial approach, the research results are expected to be more accurate and can help the government in making more targeted policies, especially in areas that are close to each other and have similar poverty problems.

REFERENCES

[1] A. Salsabila, A. Fitrianto, and M. A. Aliu, "Association of Poverty Categories , Educational Characteristics , and Area of Residence in Indonesia Using a Three-Way Log-Linear Model," vol. 17, no. 1, pp. 624–634, 2024.
[2] M. Irfan, A. Samsir, M. Jamli, M. Syafri, and S. Astuty, "Faktor-Faktor Yang Mempengaruhi Tingkat Kemiskinan di Provinsi Sulawesi Selatan," *Ekonodinamika Jurnal Ekonomi Dinamis*, vol. 6, no. 2, pp. 182–197, 2024.
[3] A. Valiant Kevin, A. Bhinadi, and A. Syari'udin, "Pengaruh Pdrb, Angka Harapan Hidup, Dan Rata Rata Lama Sekolah Terhadap Kemiskinan Di Kabupaten/Kota Provinsi Jawa

Tengah Tahun 2013-2021," *SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan*, vol. 1, no. 12, pp. 2959–2968, 2022, doi: 10.54443/sibatik.v1i12.482.

[4] S. Yulianto, A. Z. Utami, and A. N. Ambarwati, "Perbandingan Model Spasial dalam Permasalahan Kemiskinan di Provinsi Jawa Timur," vol. 24, no. 2, pp. 143–150, 2024.

[5] I. Insiyah, M. Khasanah, and T. P. Hendarsyah, "Penerapan Metode Ward Clustering Untuk Pengelompokkan Daerah Rawan Kriminalitas Di Jawa Timur Tahun 2021," *Jurnal Statistika dan Komputasi*, vol. 2, no. 1, pp. 44–54, 2023, doi: 10.32665/statkom.v2i1.1664.

[6] putri nazwa Maharani, "Jurnal Pendidikan Inklusif Analisis Cluster Dengan Average Linkage Method," vol. 8, no. 12, pp. 48–67, 2024.

[7] M. J. Budiman and Fanny Jouke Doringin, "Jurnal Ilmu Komputer," *Biomaterials*, vol. 07, no. 12, pp. 85–90, 2023.

[8] S. Bao, Arman, La Gubu, W. Somayasa, Bahriddin, and Agusrawati, "Analisis Cluster Terhadap Tingkat Pencemaran Udara Pada Sektor Industri Di Sulawesi Tenggara," *Jurnal Matematika Komputasi dan Statistika*, vol. 4, no. 1, pp. 547–557, 2024, doi: 10.33772/jmks.v4i1.79.

[9] Y. I. Harnanto, A. Rusgiyono, and T. Wuryandari, "Penerapan Analisis Klaster Metode Ward Terhadap Kabupaten/Kota Di Jawa Tengah Berdasarkan Pengguna Alat Kontrasepsi," *Jurnal Gaussian*, vol. 6, no. 4, pp. 528–537, 2017.

[10] M. A. Nahdliyah, T. Widiharih, and A. Prahutama, "Metode K-Medoids Clustering Dengan Validasi Silhouette Index Dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018)," *Jurnal Gaussian*, vol. 8, no. 2, pp. 161–170, 2019, doi: 10.14710/j.gauss.v8i2.26640.

[11] R. K. Putri, M. Athoillah, and A. Haqiqiyah, "Analisis Faktor Yang Mempengaruhi Ketepatan Kelulusan Mahasiswa Dengan Algoritma Regresi Linear," *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 671–680, 2024.

[12] I. Imasdiani, I. Purnamasari, and F. D. T. Amijaya, "Perbandingan Hasil Analisis Cluster Dengan Menggunakan Metode Average Linkage Dan Metode Ward," *Eksponensial*, vol. 13, no. 1, p. 9, 2022, doi: 10.30872/eksponensial.v13i1.875.

[13] E. Saraswi, H. Perdana, and A. Fakhrunnisa, "Distribusi Tenaga Kesehatan di Kalimantan Barat Menggunakan Metode Ward," *Jurnal Forum Analisis Statistik (FORMASI)*, vol. 4, no. 1, pp. 39–48, 2024, doi: 10.57059/formasi.v4i1.73.

[14] T. Angriani *et al.*, "Peran UMKM Dalam Menanggulangi Kemiskinan Di Provinsi Kalimantan Tengah," vol. 2, no. 1, pp. 12–21, 2024.