International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 03, No. 03, page 86 - 101
Submitted 17/72025; Accepted 14/11/2025; Published 26/11/2025

Comparative Analysis of K-Means and Hierarchical Clustering for Regional Welfare Disparity Identification in West Java Province

[1]Muhamad Dani Yusuf and [2]Tb Ai Munandar [3]Khairunnisa Fadhilla Ramdhania
[1]Informatics Department, Universitas Bhayangkara Jakarta Raya, Jakarta, INDONESIA
[2]Informatics Department, Universitas Bhayangkara Jakarta Raya, Jakarta, INDONESIA
[3]Informatics Department, Universitas Bhayangkara Jakarta Raya, Jakarta, INDONESIA
e-mail : [1]muhamad.dani.yusuf19 @mhs.ubharajaya.ac.id, [2]tbaimunandar@gmail.com, [3]khairunnisa.fadhilla@dsn.ubharajaya.ac.id

Corresponding Autor: Muhamad Dani Yusuf

Abstract
This study aims to cluster regencies/cities in West Java Province based on public welfare indicators using the K-Means Clustering and Hierarchical Clustering methods. The data used includes health, economic, population density, and average length of schooling indicators in 2023. Cluster quality evaluation was performed using the silhouette score. The results show that K-Means Clustering with five clusters yields the highest silhouette score of 0.219. For comparison, Hierarchical Clustering with the Ward Linkage method and eight clusters was chosen, having a silhouette score of 0.202, which is the largest among other Hierarchical Clustering methods. The identification of each cluster's characteristics in K-Means reveals areas with multidimensional challenges (Cluster 1), industrial areas with unemployment issues (Cluster 2), areas with high stunting prevalence despite good access to basic facilities (Cluster 3), densely populated urban areas with good welfare but high unemployment (Cluster 4), and areas with very high health complaints and low welfare (Cluster 5). K-Means clusters (except Cluster 4) tend to have a low average length of schooling, below 12 years. Consistency in cluster patterns was found between K-Means and Ward Linkage, especially in advanced urban areas and areas with multidimensional welfare challenges in southern West Java. These findings are expected to serve as a reference for the government and policymakers in formulating more targeted and effective development strategies.

Keywords—K-Means Clustering, Hierarchical Clustering, welfare indicators

1 Introduction

Public welfare constitutes one of the primary objectives of national development, serving as a focal point in various policies at both national and regional levels. This commitment is reflected in the 2015-2030 Sustainable Development Goals (SDGs) agenda, which encompasses diverse aspects of welfare. These include poverty eradication (SDG 1), improved health (SDG 3), quality education (SDG 4), and inclusive and sustainable economic growth (SDG 8). Indonesia actively participates in implementing this development agenda, which is stipulated in its National Medium-Term Development Plan (RPJMN). The RPJMN serves as a crucial guideline for formulating national policies and regional development strategies [1].

To measure the level of public welfare, Statistics Indonesia (BPS) regularly publishes its 'Indikator Kesejahteraan Rakyat Indonesia' (Indonesian Public Welfare Indicators) publication. In 2024, these indicators encompass eight main components: demography, health, education, employment, consumption levels and patterns, housing, poverty,

and other social aspects [2]. In the context of this research, the data and indicators that will be utilized are health indicators, economic indicators, education indicators, and demographic indicators.

West Java Province is the most populous province in Indonesia, with its population reaching 50,345,200 people in 2024 [3]. With such a large population, there are significant challenges in maintaining a balance of development and welfare levels across its regions. Disparities among regencies and cities in health, education, and economic indicators reveal considerable variations in welfare levels across these areas. Therefore, a data-driven approach is required to understand the distribution of welfare more comprehensively and to support the formulation of more targeted regional development policies.

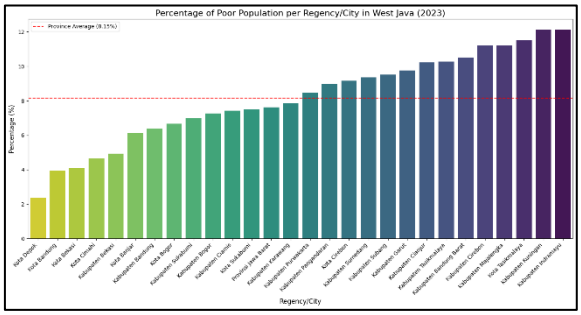

Figure 1 Percentage of Poor Population per Regency/City in West Java (2023)

Figure 1 illustrates the disparity in poverty levels across regions in West Java Province, with an average provincial rate of 7.62%. Indramayu Regency, Kuningan Regency, and Tasikmalaya City recorded poverty percentages above 11%, reflecting serious challenges in socioeconomic aspects and limited access to basic services. Conversely, Depok City, Bandung City, and Bekasi City showed significantly lower poverty rates, below 4%. These low figures indicate a concentration of economic activity, better infrastructure, and a relatively high availability of employment opportunities in these areas. This pattern suggests that areas closer to government and economic centers tend to have lower poverty rates compared to peripheral regions.

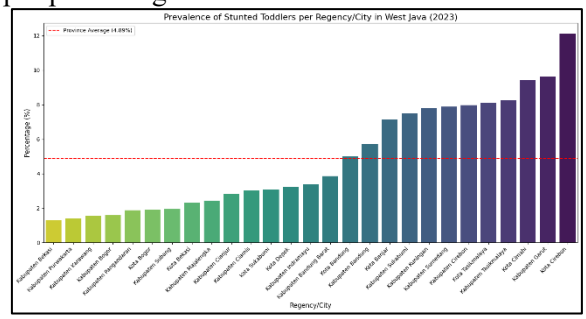

Figure 2 Prevalence of Stunted Toddlers per Regency/City in West Java (2023)

In addition to poverty, issues of malnutrition and stunting also pose serious challenges, indicating developmental disparities between regions in West Java. Figure 2 reveals significant variations in stunting prevalence across various regencies and cities in West Java in 2023, with the provincial average at 4.90%. Cirebon City recorded the highest figure, exceeding 12%, followed by Garut Regency, Cimahi City, and Tasikmalaya Regency, which were also well above the provincial average. Conversely, areas such as Bekasi Regency, Karawang, Purwakarta, and Bogor recorded much lower stunting rates. This striking difference once again underscores the disparities in socioeconomic conditions, access to health services, sanitation, and nutritional consumption patterns.

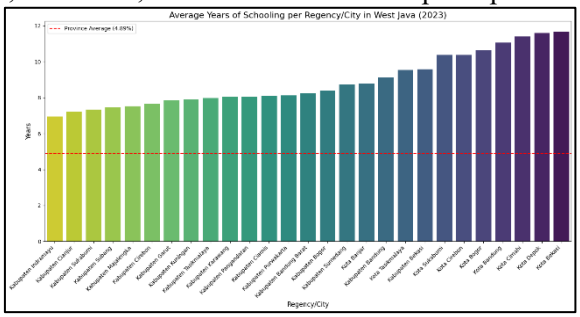

Figure 3 Average Years of Schooling per Regency/City in West Java (2023)

Figure 3 illustrates that urban areas in West Java tend to have a higher average length of schooling compared to rural areas. Bekasi City recorded the highest figure, with an average length of schooling approaching 12 years, followed by Depok City, Cimahi City, and Bandung City, all of which were above 10 years. This figure is close to the national compulsory education standard of 12 years, which covers basic to upper secondary education levels. Conversely, several regencies such as Indramayu Regency, Cianjur Regency, and Sukabumi Regency have an average length of schooling below 9 years, meaning most of the population in these areas only complete education up to junior high school level. This difference reflects the inequality in access to education between urban and rural areas, in terms of the availability of educational facilities, quality of teaching, and the socioeconomic conditions of the community.

Previous studies have extensively demonstrated the efficacy of clustering analysis methods in grouping regions based on welfare levels. Both hierarchical clustering [4], [5] and K-Means clustering [6], [7], [8] have been widely applied. Hierarchical methods are often favored for their ability to present hierarchical structures and dendrogram visualizations [9], while K-Means is recognized for its computational efficiency, particularly with large datasets [6], [9]. However, K-Means can be sensitive to the initial centroid placement and the pre-determined number of clusters. Research has explored various hierarchical linkage methods, including single, complete, average, and Ward, to determine the most effective fit for specific datasets [10], [11].

For non-hierarchical approaches like K-Means, determining the optimal number of clusters is crucial. The Elbow Method is a common technique used for this purpose [12], [13], identifying a point where further increases in cluster count yield diminishing returns in reducing the Within-Cluster Sum of Squares (WCSS). The quality of the formed clusters is subsequently evaluated using various metrics. The silhouette method [14] assesses how well each data point fits its own cluster compared to others [15], while the Dunn index [16], [17] aims to identify compact and well-separated clusters. Previous research on welfare indicators has varied in the number and type of variables used, ranging from 8 variables including economic, health, and education [4] to 10 socioeconomic indicators [6], and more specific sets of 4 variables [7].

Despite these valuable contributions, a comprehensive analysis specifically directed at unveiling the patterns of welfare disparity between regions remains a notable research gap [18]. Building upon existing methodologies and identified welfare disparities, this study aims to provide a more in-depth understanding of welfare patterns in West Java Province. This will be achieved through a comparative analysis of K-Means Clustering and Hierarchical Agglomerative Clustering (HAC). The research systematically applies both methods to group regencies and cities based on a refined set of 12 multidimensional welfare indicators, encompassing health, economy, population density, and average years of schooling. By comparing the outcomes and characteristics of clusters generated by both K-Means and HAC, this study seeks not only to objectively quantify welfare distribution and disparities but also to identify which clustering approach provides more robust and interpretable insights for regional development strategies. The findings are expected to serve as an important reference for local governments, academics, and other stakeholders in designing more inclusive, adaptive, and sustainable development strategies.

2 Research methods

This research employs the Cross Industry Standard Process for Data Mining (CRISP-DM) [19] methodology to cluster West Java's regencies and cities based on public welfare indicators. The objective was to identify regional welfare disparity patterns to support targeted policymaking.

Data was collected from Badan Pusat Statistik Jawa Barat and the Open Data Jabar platform, encompassing 12 indicators across health, economy, education, and demography (detailed in Table 1). The data preparation phase involved cleaning the dataset to include only the 27 regencies/cities, checking for missing values (none were found), merging the datasets, and normalizing all 12 variables using min-max scaling (Equation 1) to ensure equal weighting in the models.

$$x'_{i,j} = \frac{x_{i,j} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (1)$$

For the modeling phase, K-Means and Hierarchical Agglomerative Clustering (HAC) were applied using Euclidean distance. The K-Means algorithm was run using the k-means++ initialization method to ensure stable convergence, with the optimal cluster count (k) explored using the Elbow Method. For HAC, a comprehensive comparison was conducted using four linkage methods: single, complete, average, and Ward.

Finally, in the evaluation phase, the resulting clusters were assessed quantitatively to find the optimal solution. Cluster quality was measured using the Silhouette Method. This quantitative assessment was followed by a

qualitative interpretation of each cluster's unique characteristics to provide actionable insights into regional welfare challenges.

2.1 *K-Means Clustering*

K-Means clustering is an unsupervised learning method used to group data points into a predefined number of clusters (K) based on similarity. This iterative, centroid-based algorithm aims to minimize the variation or distance within each cluster, where a cluster's centroid is the conceptual center representing the average value of its objects [20].

The K-Means process begins by randomly initializing K centroids from the dataset. Each data point is then assigned to its nearest or most similar centroid. Subsequently, the centroid for each cluster is updated by calculating the mean of all data points assigned to it. This assignment and centroid update process is repeated iteratively until cluster membership no longer changes. The clustering process is illustrated in Figure 4 below.

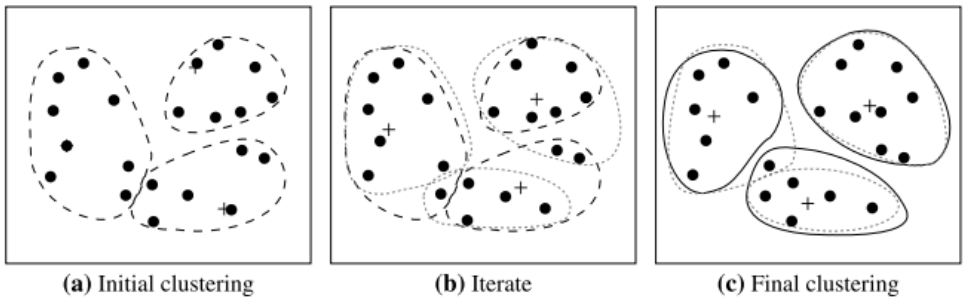

Figure 4 The k-means partitioning algorithm.

The dissimilarity between an object and its centroid (representing the cluster) is measured using a distance function. The within-cluster variation is quantified using the Sum of Squared Errors (SSE) between all objects within a cluster and its respective centroid, defined as follows [9]:

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - m_i\|^2 \quad (1)$$

2.2 *Hierarchical Agglomerative Clustering (HAC)*

Hierarchical clustering methods form a nested decomposition of data objects, utilizing two main approaches: agglomerative (bottom-up merging) and divisive (top-down splitting). Agglomerative methods start with individual objects and gradually merge them, while divisive methods begin with one large group that is progressively split into smaller ones [9]. This research employs the following linkage methods:

1. Single Linkage
   Single linkage measures the distance between two clusters based on the shortest (minimum) distance between any pair of points from each cluster [9]. This method tends to form elongated or non-spherical clusters due to merging based on the closest points.
   $$dist(C_i, C_j) = \min_{x \in C_i, x \in C_j} \{\|x - x'\|\} \quad (2)$$
2. Complete Linkage
   Complete linkage measures the distance between two clusters based on the farthest (maximum) distance between any pair of points from each cluster [9]. This method is more conservative than single linkage and tends to produce more compact and spherical clusters.
   $$dist(C_i, C_j) = \max_{x \in C_i, x \in C_j} \{\|x - x'\|\} \quad (3)$$
3. Average Linkage
   Average linkage offers a compromise between single linkage's minimum distance and complete linkage's maximum distance [9]. By using the average distance, this method helps mitigate the outlier sensitivity often seen in the other two.
   $$dist(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i,\ x' \in C_j} \|x - x'\| \quad (4)$$
4. Ward Linkage

Ward linkage merges two clusters based on minimizing the increase in the sum of squared errors (SSE) upon combination [9]. This method aims to maintain data closeness and similarity within clusters, typically yielding relatively balanced and compact clusters. The inter-cluster distance, as shown in Equation 5, increases with larger average differences and cluster sizes.

$$dist(C_i, C_j) = \frac{n_i n_j}{n_i + n_j} \|\mathrm{m_i} - m_j\|^2 \quad (5)$$

2.3 *Sillhouette Method*

The Silhouette Method is an evaluation technique used to assess the quality of clustering results by measuring how similar a data point is to its own cluster compared to other clusters. It is also employed to determine the optimal number of clusters by calculating and averaging the silhouette coefficient for each data point [14]. Silhouette values range from -1 to 1: values near 1 indicate a good fit within its cluster, values near 0 suggest the data is on a cluster boundary, and negative values imply the data might be misclassified. The silhouette value is calculated using the following formula:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), (bi)\}} \quad (6)$$

3 Results and Discussion

3.1 *Data Understanding*

The data utilized in this research was collected from the West Java Central Statistics Agency (Badan Pusat Statistik Jawa Barat) and the Open Data Jabar platform, referencing public welfare indicators consistent with BPS's publication titled "Indonesia Welfare Indicators 2024" [2]. The welfare indicators leveraged primarily focused on health, socioeconomic, education, and demographic aspects.

Table 1 Research Variables (Data for 2023)

Variable	Description	Unit	Data type
X1	Percentage of Population with Health Complaints in the Last Month	Percentage	float
X2	Percentage of Stunted Toddlers Prevalence	Percentage	float
X3	Percentage of Exclusive Breastfeeding for Infants < 6 months	Percentage	float
X4	Percentage of Households with Decent Sanitation	Percentage	float
X5	Percentage of Households with Access to Decent Drinking Water	Percentage	float
X6	Average Monthly Per Capita Expenditure on Food and Non-Food	IDR	float
X7	Open Unemployment Rate	Percentage	float
X8	Percentage of Poor Population	Percentage	float
X9	Average Years of Schooling	Year	float
X10	Human Development Index (HDI)	Index	float
X11	Population Density Rate	People/Km^2	Integer
X12	Population Growth Rate	Percentage	float

To gain a deeper understanding of the dataset's characteristics and the distribution of values for each variable, a descriptive statistical analysis was performed. The results of this analysis, which include the mean, standard deviation, quartiles (25%, 50%, and 75%), minimum, and maximum values for each research variable, are presented in Table 2. This statistical summary provides crucial insights into the central tendency, dispersion, and range of the data points, which are essential for identifying potential outliers and informing subsequent data preparation and modeling stages.

Table 2 Descriptive Statistics of Research Variables

Variabel	Mean	Std	25%	50%	75%	Min	Max
X1(%)	26,91	6,14	22,13	26,55	30,86	16,51	41,48
X2(%)	4,89	3,10	2,14	3,40	7,84	1,29	12,12
X3(%)	74,33	16,19	64,11	73,86	80,23	41,59	133,60
X4(%)	75,70	16,60	59,60	77,32	89,72	45,88	99,08
X5(%)	94,13	5,43	92,08	95,84	98,31	81,13	99,61
X6(million)	1,58	0,53	1,23	1,38	1,79	1,03	2,73
X7(%)	7,185	1,98	6,53	7,65	8,50	1,52	10,52
X8(%)	8,17	2,62	6,53	8,46	10,25	2,38	12,13
X9(year)	8,87	1,42	7,86	8,23	9,97	6,94	11,66
X10	74,11	4,26	70,66	73,25	76,64	68,18	83,29
X11(thousand)	3,85	4,48	0,84	1,43	5,83	0,39	15,42
X12(%)	1,15	0,24	0,96	1,18	1,38	0,66	1,55

Table 2 presents the descriptive statistics for each variable utilized in this study, offering a comprehensive overview of the data's distribution and characteristics. This summary includes the mean, standard deviation, quartiles (25th, 50th or median, and 75th percentiles), along with minimum and maximum values for each variable. A thorough understanding of these statistics is essential for discerning the value range, data dispersion, and the potential presence of outliers within each variable. For health indicators, X1 (Percentage of Population with Health Complaints in the Last Month) had a mean of 26.91%, ranging from 16.51% to 41.48%. X2 (Percentage of Stunted Toddlers Prevalence) showed a mean of 4.89% with a maximum of 12.12%, indicating significant regional variation. X3 (Percentage of Exclusive Breastfeeding for Infants < 6 months) averaged 74.33%, suggesting generally good coverage, though some areas recorded a low of 41.59%. Both X4 (Percentage of Households with Decent Sanitation) and X5 (Percentage of Households with Access to Decent Drinking Water) showed high averages (75.70% and 94.13% respectively), reflecting relatively good access to these basic facilities in West Java. Regarding economic indicators, X6 (Average Monthly Per Capita Expenditure on Food and Non-Food) averaged IDR 1.58 million, with a wide range from IDR 1.03 million to IDR 2.73 million, highlighting inter-regional economic disparities. X7 (Open Unemployment Rate) and X8 (Percentage of Poor Population) had means of 7.185% and 8.17% respectively, with some areas experiencing poverty rates as high as 12.13%. In terms of education, X9 (Average Years of Schooling) showed a mean of 8.87 years, still below the 12-year compulsory education target, with values ranging from 6.94 to 11.66 years. Finally, X10 (Human Development Index) averaged 74.11, demonstrating variability in human development levels.

3.2 *Data Preparation*

This phase involved processing the collected data to ensure its readiness for the modelling stage, which focused on the clustering analysis of regencies/cities in West Java Province based on public welfare indicators. The steps undertaken included data cleaning and data normalization.

Initially, the dataset underwent a thorough cleaning process. This involved removing instances that represented provincial averages, as these were not relevant for the regency/city-level analysis. Specifically, data entries for the "West Java Province average" were excluded, ensuring that the dataset comprised precisely 27 instances, corresponding to the total number of regencies/cities in West Java. Additionally, any supplementary textual

information within the datasets was removed, retaining only the names of the regencies/cities and their respective variable values.

A check for missing values was then conducted across all 12 variables for the 27 regencies/cities. No missing data was found.

Upon completion of the data cleaning, the 12 individual datasets were merged into a single, unified dataset. This consolidated dataset served as the comprehensive feature matrix for the subsequent clustering algorithms. A critical step in this preparation was addressing outliers, which were identified during the preceding data understanding phase. According to Barus & Sutarman in [21] Normalization was applied to mitigate the potential impact of these outliers, as their presence can lead to non-normal data distribution, introduce bias, and potentially decrease the performance of machine learning algorithms.

To ensure that no single variable's scale disproportionately influenced the clustering results, Min-Max normalization was implemented. This method scales all feature values to a uniform range between 0 and 1. The Min-Max normalization was calculated using the Formula (1) defined before.

3.3 *Modeling*

Following the business understanding, data understanding, and data preparation phases of data analysis, the research findings will be presented in two main sections: K-Means Clustering and Hierarchical Clustering. The chosen clustering results for discussion are those exhibiting the highest silhouette score, as well as those that most effectively reveal cluster characteristics and deeper insights into the welfare challenges, aligning with the study's objectives.

3.3.1 K-Means Clustering

The modeling process using the K-Means method begins by determining the optimal number of clusters so that the resulting clusters are of good quality and their members do not overlap with other cluster members. One way to determine the optimal number of clusters is by using the Elbow Method.

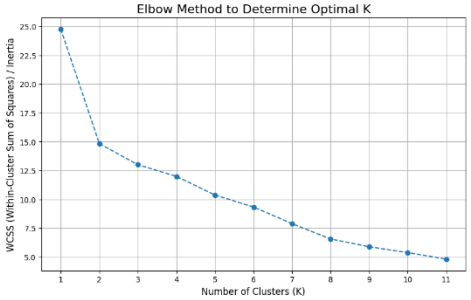

Figure 5 Elbow method to determine the optimal number of k for K-Means Clustering

Looking at Figure 5, the curve initially shows a significant decrease in WCSS as the number of clusters increases. However, this reduction in WCSS begins to slow down around k=3 and k=4, forming an "elbow" shape on the graph. This indicates that adding more clusters beyond this point does not lead to a significant further decrease in WCSS, suggesting that most of the data's variance is explained by 3 or 4 clusters. To ensure the quality of the resulting clusters and to aid in determining the optimal number of clusters for presentation, the silhouette score is also utilized as a metric. This indicator measures how well each data instance fits its own cluster and how distinctly separated it is from other clusters, providing insight into cluster compactness and separation. The silhouette values obtained for each variation in the number of clusters are presented in Table 3.

Table 3 Silhouette score based on the number of K-Means clusters

	Number of Clusters								
	3	4	5	6	7	8	9	10	11
Silhouette Score	0,199	0,204	**0,219**	0,188	0,181	0,178	0,184	0,159	0,169

Based on the silhouette scores presented in Table 3, the chosen number of clusters is 5, yielding the highest silhouette value of 0.219. This selection is further supported by the relatively low WCSS value observed in the elbow method graph (Figure 5). Following the determination of the optimal number of clusters, the centroids for each cluster are shown in Table 34

Table 4 Centroid of Each K-Means Clustering Cluster

Variabel	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
X1	0,54	0,179	0,47	0,21	0,85
X2	0,47	0,02	0,57	0,26	0,10
X3	0,35	0,43	0,31	0,34	0,33
X4	0,16	0,70	0,67	0,53	0,71
X5	0,30	0,71	0,84	0,94	0,49
X6	0,03	0,23	0,23	0,86	0,11
X7	0,59	0,75	0,61	0,79	0,17
X8	0,73	0,53	0,75	0,25	0,69
X9	0,16	0,29	0,35	0,88	0,19
X10	0,07	0,34	0,35	0,82	0,21
X11	0,03	0,07	0,15	0,72	0,02
X12	0,55	0,75	0,49	0,66	0,11

3.3.2 Hierarchical Agglomerative Clustering

Hierarchical clustering analysis was performed using four methods: single linkage, complete linkage, Ward linkage, and average linkage, on data previously normalized with the min-max method. The clustering result selected for further analysis will be the one with the highest silhouette score.

Single linkage was the first method utilized in the hierarchical clustering analysis. This method merges clusters based on the minimum distance between any pair of data instances. Due to its reliance on the closest points, the resulting clusters often exhibit a long and thin, chain-like structure, allowing clusters to extend and connect data points that may otherwise be quite distant, provided a single close point bridges them.

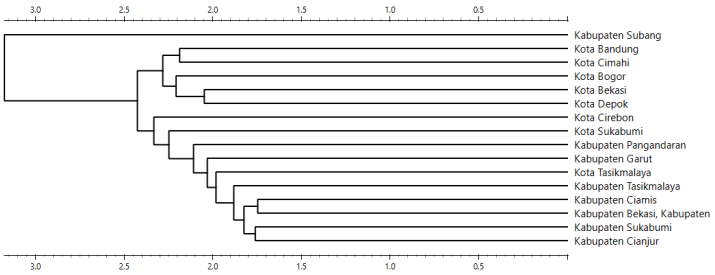

Figure 6 Single linkage clustering results.

From the dendrogram shown in Figure 6, it can be observed how initial clusters are formed from very close data points, which then gradually merge to create larger structures.

Following the analysis with single linkage, the next method to be utilized is complete linkage. In contrast to single linkage's focus on minimum distance, complete linkage merges clusters based on the maximum distance between any two points from different clusters. This method is more conservative and tends to produce more compact and spherical clusters. The subsequent analyses in this research will also present results from average linkage and Ward linkage, offering a comprehensive comparison of different hierarchical clustering approaches.

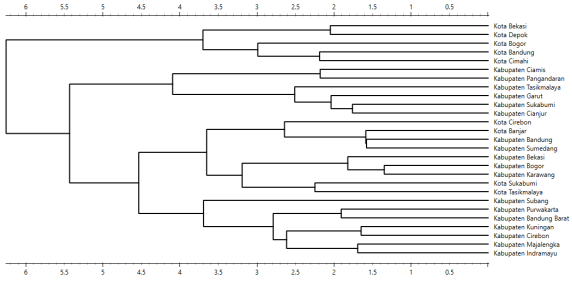

Figure 7 Complete linkage clustering results

The clusters formed in Figure 7 tend to be more compact and have closer internal distances compared to single linkage, as clusters will only merge if all their members are relatively close to one another. However, the distances between clusters are larger compared to single linkage at the same merging level.

The next grouping utilized the Average Linkage method. Average Linkage serves as a compromise between the Single Linkage and Complete Linkage methods. In this approach, clusters are merged based on the average distance between all pairs of points from two different clusters. This method is less sensitive to outliers compared to single linkage and less compact than complete linkage, consequently yielding more balanced and representative clusters.

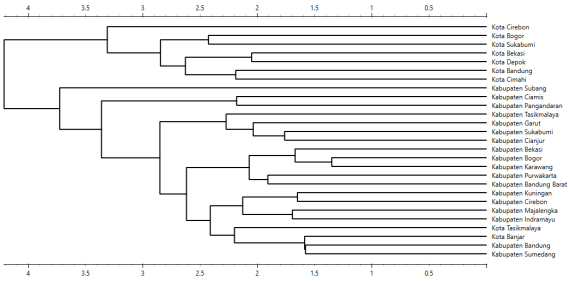

Figure 8 Average linkage clustering results

The dendrogram presented in Figure 8 illustrates the clustering results obtained using the average linkage method. The cluster merging pattern appears balanced and measured, indicating that data instances form groups at various levels of similarity without excessively dense initial clusters or isolated extreme data points.

The last method for grouping utilized was Ward Linkage. Unlike the three preceding methods, which rely on distance calculations, Ward linkage focuses on minimizing the increase in variance within clusters after a merger. Specifically, Ward linkage selects the two clusters that, when combined, will result in the smallest sum of squared errors (SSE).

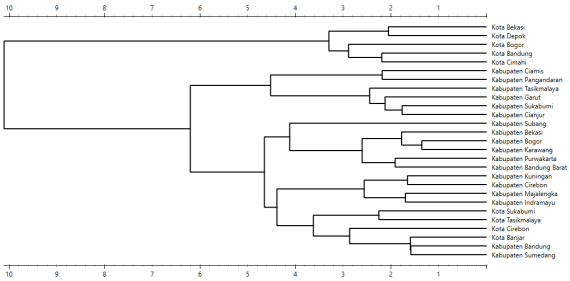

Figure 9 Ward linkage clustering results

Figure 9 presents the clustering results using the Ward Linkage method. This dendrogram reveals a cluster structure that is generally more homogeneous in terms of internal variance. The horizontal lines, labeled from 1 to 10, indicate the level of dissimilarity or variance of the clusters intended for merging. Observing the dendrogram, large urban areas such as Bekasi City, Depok City, Bogor City, Bandung City, and Cimahi City show significant differences compared to other regions.

3.4 *Evaluation*

This section elaborates on the clustering analysis results for regencies/cities in West Java Province, utilizing both K-Means and hierarchical clustering methods, with the data gathered in the preceding section. The performance of these clustering methods is then compared based on the silhouette scores obtained from the generated clusters, as detailed in Table 5. The method yielding the highest silhouette score, and capable of explaining the welfare characteristics of regions in West Java Province, will be selected for further in-depth analysis.

Table 5 Comparison of Silhouette Scores Across Clustering Methods

Methods	Number of Clusters							
	3	4	5	6	7	8	9	10
K-means	0,199	0,204	**0,219**	0,188	0,181	0,178	0,184	0,159
Single Linkage	0,129	0,094	0,069	0,079	0,084	0,079	0,043	0,024
Complete Linkage	0,180	0,129	0,151	0,122	0,108	0,127	0,157	0,141
Average Linkage	0,166	0,149	0,106	0,138	0,119	0,132	0,162	0,139
Ward Linkage	0,180	0,165	0,181	0,188	0,189	**0,202**	0,189	0,172

Based on the comparative silhouette analysis presented in Table 5, the K-Means method with five clusters was identified as yielding the highest silhouette score of 0.219. For comparative purposes, the Ward Linkage clustering method with eight clusters was also selected to compare their respective grouping results. This selection aligns with the research objective of understanding welfare indicator disparity patterns across West Java Province through identifying the characteristics of the formed clusters. These selected clustering results will subsequently be the focus for identifying and discussing the characteristics of each formed cluster. Cluster characteristics will be analyzed based on the centroid values obtained from the K-Means clustering results, with data found in Table 4. To facilitate the identification of high and low centroid patterns for each variable within every cluster, a heatmap will be utilized as an interpretative tool.

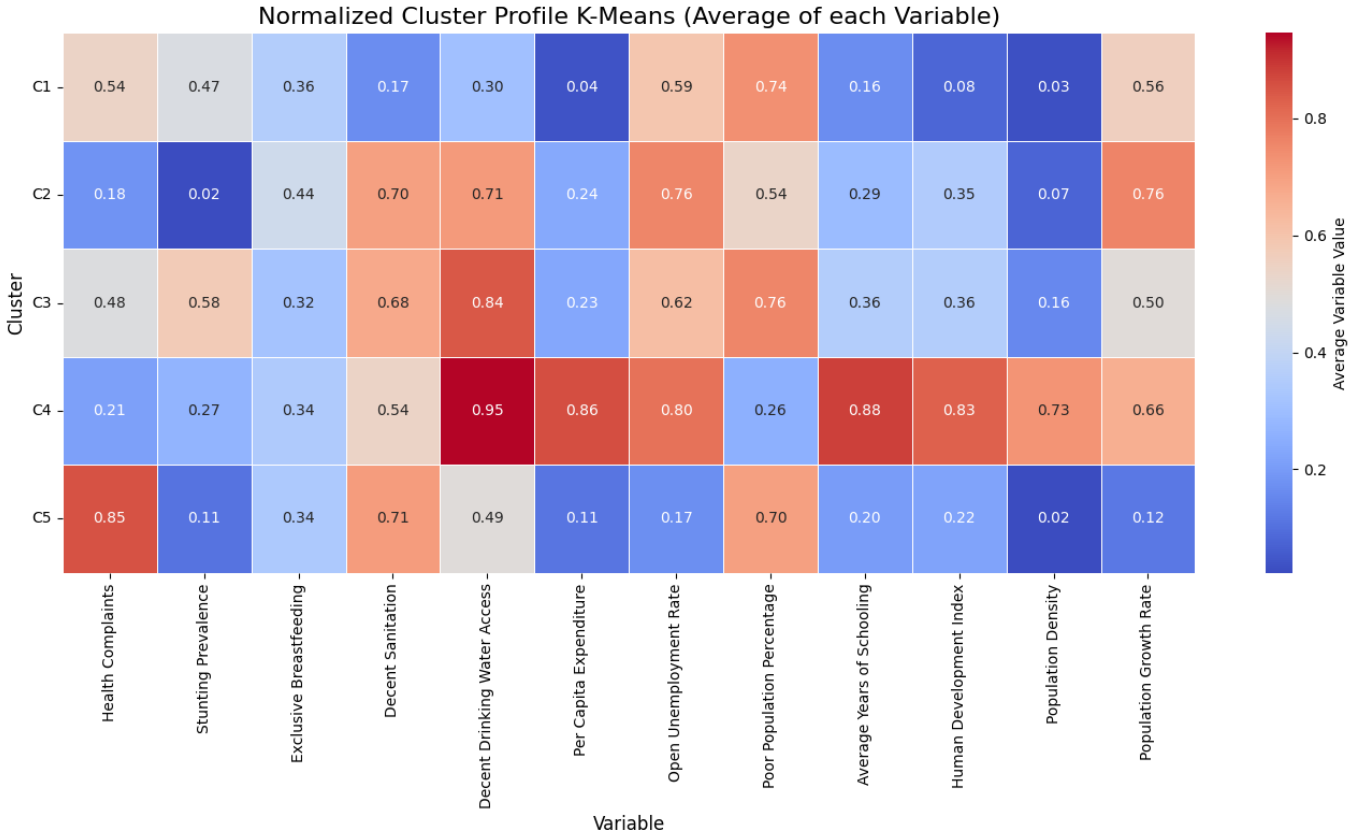

Figure 10 Visualizing Normalized K-Means Cluster Centroid Heatmap

Based on the heatmap visualization in Figure 4.6, the characteristics of each cluster can be interpreted through color intensity, where red/orange indicates high variable values and blue indicates low values. From this analysis, five diverse clusters were identified.

Cluster 1 exhibits multidimensional welfare challenges, characterized by high rates of health complaints, a significant number of poor residents, and a relatively high population growth rate. However, it also shows very low per capita expenditure, indicating a low level of economic development, accompanied by low open unemployment rates (TPT), Human Development Index (HDI), population density, and stunting rates. This combination points to regions facing substantial basic welfare challenges amidst underdeveloped economic conditions.

Cluster 2, on the other hand, demonstrates relatively good health conditions with low health complaints and stunting prevalence. While it boasts good access to sanitation and drinking water, it grapples with a high TPT and a moderate percentage of poor residents, alongside not-so-large per capita expenditure. Education indicators are moderate, and despite low population density, the population growth rate is very high. In contrast, Cluster 3 has the highest stunting prevalence and moderate health complaints, even with excellent access to decent sanitation and drinking water. This condition is likely attributed to the high poverty rate in this cluster, which may also correlate with inadequate basic health facilities.

Cluster 4 generally presents very strong welfare indicators across various aspects. Regions within this cluster tend to have excellent basic facilities, a highly advanced economy (marked by very high per capita expenditure), and very high population density. However, a prominent issue is the high TPT, potentially due to intense competition in these areas' labor markets. Finally, Cluster 5 faces severe health problems, evidenced by very high health complaints. This cluster consistently shows very low values across almost all other welfare indicators, including stunting, per

capita expenditure, TPT, poor population, average years of schooling, HDI, population density, and population growth rate. This suggests areas with broad and diverse socioeconomic challenges, coupled with low economic activity.

After identifying the characteristics of each cluster based on their centroid values, the next step involves identifying the specific regencies/cities belonging to each cluster. This determination of cluster members will facilitate assigning representative labels to each group.

Table 6 Members of Clusters Resulting from K-Means Clustering Method

Clusters	Members	Count
Cluster 1	Kabupaten Sukabumi, Kabupaten Cianjur, Kabupaten Garut, Kabupaten Tasikmalaya, Kabupaten Bandung Barat	5
Cluster 2	Kabupaten Bogor, Kabupaten Subang, Kabupaten Purwakarta, Kabupaten Karawang, Kabupaten Bekasi	5
Cluster 3	Kabupaten Bandung, Kabupaten Kuningan, Kabupaten Cirebon, Kabupaten Sumedang, Kabupaten Indramayu, Kota Cirebon, Kota Tasikmalaya, Kota Banjar	8
Cluster 4	Kota Bogor, Kota Sukabumi, Kota Bandung, Kota Bekasi, Kota Depok, Kota Cimahi	6
Cluster 5	Kabupaten Ciamis, Kabupaten Majalengka, Kabupaten Pangandaran	3
	Jumlah	27

After identifying the list of regencies/cities belonging to each cluster, as presented in Table 6, it is important to visualize the geographical distribution of these clusters. Figure 11 displays a map illustrating the spatial distribution of each regency/city in West Java Province based on its cluster membership, which can then provide a clearer understanding of the regional patterns of this grouping.

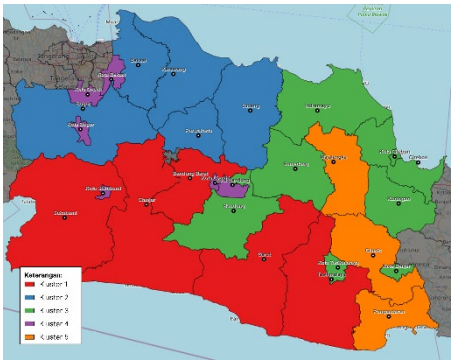

Figure 11 Visualizing K-Means Cluster Distribution with a Choropleth Map.

Based on the centroid characteristics and geographical visualization, the clusters formed can be summarized as follows. Cluster 1 generally identifies regencies in the southern and southwestern parts of West Java, such as Sukabumi, Cianjur, and Garut Regencies. This cluster typically exhibits multidimensional welfare issues and relatively underdeveloped economic conditions, where health, poverty, and educational challenges are intertwined. Similarly, Cluster 5 comprises regencies in southeastern West Java, namely Ciamis, Majalengka, and Pangandaran Regencies, indicating areas with generally very low basic welfare, marked by serious health problems and low economic activity.

Cluster 2 identifies rapidly developing industrial areas surrounding Jakarta, including Bekasi, Karawang, and Bogor Regencies. Despite showing good development and access to basic facilities, these areas still face significant employment issues, particularly high unemployment rates. In contrast, Cluster 3, spread across central to eastern West Java, presents a contradictory situation. Although it has good access to decent sanitation and drinking water, this region exhibits a high prevalence of stunting coupled with high poverty rates.

Lastly, Cluster 4 specifically represents the main metropolitan centers in West Java. This cluster is characterized by significant economic advancement and very high population density, with overall excellent welfare indicators. Nevertheless, these areas face substantial challenges due to high competition in the labor market.

Next, the results of clustering regencies/cities using the Ward Linkage method will be presented. The aim is to compare these results with the previously performed K-Means clustering. The following is the list of cluster members formed by Ward Linkage.

Table 7 Ward Linkage Clustering Results

Clusters	Members	Count
Cluster 1	Kota Bogor, Kota Bandung, Kota Bekasi, Kota Depok, Kota Cimahi	5
Cluster 2	Kabupaten Ciamis, Kabupaten Pangandaran	2
Cluster 3	Kabupaten Sukabumi, Kabupaten Cianjur, Kabupaten Garut, Kabupaten Tasikmalaya	4
Cluster 4	Kabupaten Subang	1
Cluster 5	Kabupaten Bogor, Kabupaten Purwakarta, Kabupaten Karawang, Kabupaten Bekasi, Kabupaten Bandung Barat	5
Cluster 6	Kabupaten Kuningan, Kabupaten Cirebon, Kabupaten Majalengka, Kabupaten Indramayu	4
Cluster 7	Kota Sukabumi, Kota Tasikmalaya	2
Cluster 8	Kabupaten Bandung, Kabupaten Sumedang, Kota Cirebon, Kota Banjar	4
	Jumlah	27

Clustering using the Ward Linkage method resulted in 8 clusters, with a more concise number of members in each cluster compared to the K-Means grouping results. The distribution and members of each cluster from the Ward Linkage grouping can be further observed on the map presented in Figure 12.

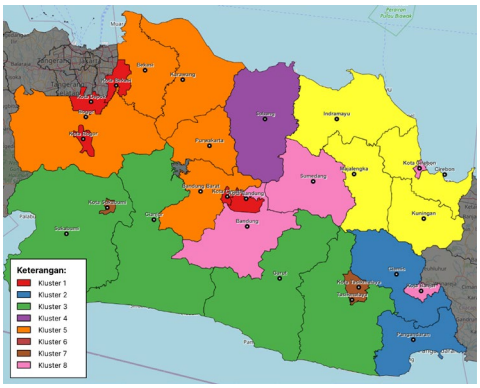

Figure 12 Visualizing Ward Linkage Cluster Distribution with a Choropleth Map.

Observing the cluster distribution map in Figure 12, a consistent grouping pattern with the previously obtained K-Means results is evident. This pattern indicates that regencies/cities belonging to the same cluster tend to be geographically proximate. This geographical closeness suggests a relative similarity in welfare characteristics among these areas. Such proximity can reflect shared social, economic, or environmental factors that influence welfare levels. To understand the specific characteristics of each formed cluster, particularly concerning the mean values of each welfare indicator feature, the cluster profile visualization in the heatmap in Figure 13 can be examined. This heatmap provides a clear visual representation of the intensity of each variable's value within each cluster.

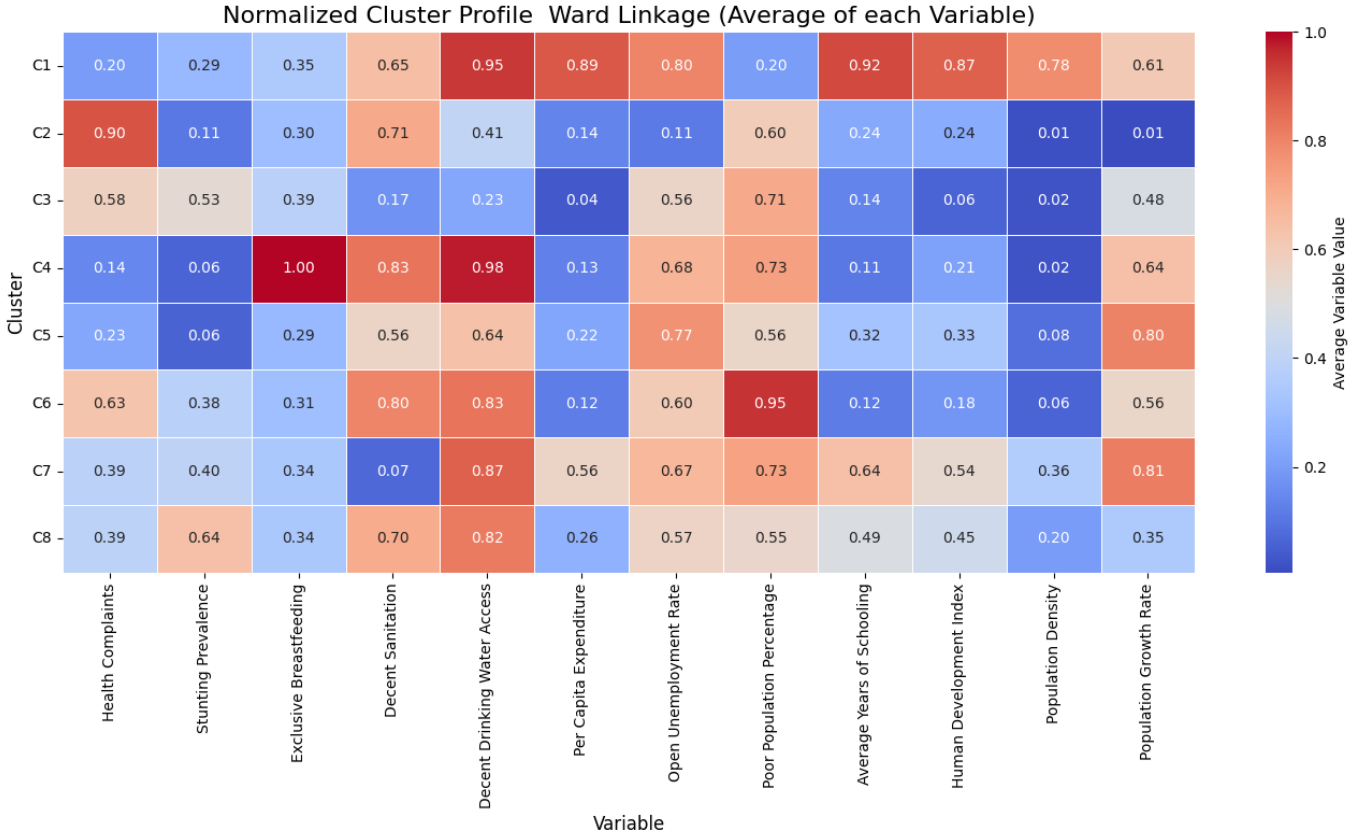

Figure 13 Visualizing Normalized Ward Linkage Cluster Centroid Heatmap.

Based on the heatmap profile (Figure 13), geographical distribution (Figure 12), and cluster member list (Table 7), the eight clusters derived from the Ward Linkage method reveal diverse welfare characteristics across West Java Province. Cluster 1, representing metropolitan and urban centers (e.g., Bogor, Bandung, Bekasi Cities), is characterized by excellent access to basic facilities, strong educational indicators (high Average Years of Schooling and HDI), high per capita expenditure, and low poverty, though high population density and Open Unemployment Rate (TPT) suggest intense labor market competition. Clusters 2 (Pangandaran and Ciamis Regencies) and 3 (Sukabumi, Cianjur, Garut, and Tasikmalaya Regencies) generally face significant welfare challenges, marked by high health complaints, low access to drinking water (Cluster 2), high poverty, and low educational indicators (Cluster 2 and 3). Cluster 3 also shows relatively high stunting rates, possibly due to inadequate sanitation and clean water access. These southern West Java clusters require focused attention on improving basic quality of life.

Cluster 5 (Bogor, Purwakarta, Karawang, Bekasi, and West Bandung Regencies) demonstrates good health indicators but struggles with high unemployment (TPT) and poverty, despite moderate education and high population growth, likely reflecting challenges in industrial buffer zones. Cluster 6 (Kuningan, Cirebon, and Majalengka Regencies) shows high health complaints and the highest poverty rates despite adequate basic facilities, alongside low educational indicators, indicating a disparity between infrastructure and actual community welfare. Cluster 4 (Subang Regency) presents a unique profile with very high exclusive breastfeeding rates and low health/stunting issues, supported by good sanitation and clean water access, yet still facing high TPT and poverty with a growing population. Cluster 7 (Sukabumi and Tasikmalaya Cities) has low sanitation access but high clean water access, moderately high poverty and TPT, good education (though not as high as Cluster 1), and the highest population growth rate, with sanitation access remaining a potential health risk despite relatively low health complaints and stunting. Lastly, Cluster 8 (Bandung, Sumedang Regencies, Cirebon, and Banjar Cities) stands out with the highest stunting rates despite good basic sanitation and clean water access. Economically, this cluster shows moderate per capita expenditure, TPT, and poverty, with decent educational indicators. The high stunting here might point to unexamined factors like nutritional practices or specific health issues.

Overall, the Ward Linkage results offer a more detailed portrayal due to the greater number of clusters, allowing for better identification of specific characteristics and challenges facing each regional group. A notable consistency was observed between some K-Means and Ward Linkage clusters. For example, K-Means Cluster 4 (high-density urban areas with excellent welfare) closely resembles Ward Linkage Cluster 1. Similarly, K-Means Cluster 1, representing areas in southern West Java with low multidimensional welfare, is analogous to Ward Linkage Cluster

3. These similarities across both methods underscore a high degree of consistency in the clustering outcomes, indicating that the identified welfare patterns in West Java Province reflect an inherent structure or typology rather than being an artifact of a specific algorithm. This cross-method consistency provides strong validation for the formed clusters, confirming that distinct segments of regions with similar welfare characteristics exist, irrespective of the algorithmic approach used.

4 Conclusion

This study aimed to cluster regencies/cities in West Java Province based on public welfare indicators and to understand the characteristics and welfare challenges faced by the resulting regional groups. To achieve this objective, both K-Means and Hierarchical Clustering methods were employed. Based on silhouette score evaluations, the K-Means method with five clusters was selected as the optimal result due to its highest silhouette score of 0.219. For comparative analysis, the Ward Linkage method with eight clusters was also utilized, yielding a silhouette score of 0.202, which was the highest among other Hierarchical Clustering methods.

The application of K-Means clustering in this research identified five distinct welfare typologies within West Java Province. Generally, areas within the same cluster tended to be geographically contiguous, with the exception of Cluster 4, which predominantly comprised advanced urban regions. Cluster 1 characterized areas facing multidimensional welfare challenges. Cluster 2 indicated regions with good welfare levels but significant issues of unemployment and poverty. Cluster 3 was distinguished by the highest stunting prevalence despite good access to basic facilities, alongside the highest percentage of poor residents. Cluster 4 encompassed densely populated urban areas with generally good welfare but a primary concern of high unemployment. Finally, Cluster 5 exhibited very high health complaints, coupled with substantial poverty and low population density. Notably, most K-Means clusters, excluding Cluster 4, showed a low Average Years of Schooling, falling below 12 years.

Concurrently, clustering with Ward Linkage yielded eight distinct clusters, each presenting specific characteristic profiles for regional groups. Analysis revealed a high degree of consistency between the K-Means and Ward Linkage results, particularly in identifying extreme clusters. For instance, K-Means Cluster 4 (advanced urban areas) exhibited similarities with Ward Linkage Cluster 1, while K-Means Cluster 1 (regions with multidimensional welfare challenges in southern West Java) was comparable to Ward Linkage Cluster 3. This consistency underscores that the identified welfare patterns are not merely products of a single algorithm but reflect the inherent structural typologies within the region's welfare conditions. The comprehensive findings from both clustering methods effectively address the research objective of classifying and understanding the disparities in public welfare indicators across West Java Province based on the multidimensional characteristics of these indicators.

5 Suggestion

Based on the research conclusions, the findings are expected to assist the West Java Provincial Government and other policymakers in formulating more targeted and specific policies. By leveraging these clustering results, policies can be tailored to address the unique challenges faced by each regional group, thereby enhancing the effectiveness and precision of public welfare improvement programs.

Furthermore, the following suggestions are put forth for future research:

1. Future studies could explore the use of more robust clustering methods, such as DBSCAN, to potentially enhance the quality of the groupings.
2. For analyses involving a large number of research variables (high dimensionality), future research may consider employing dimensionality reduction techniques like Principal Component Analysis (PCA).
3. Subsequent research could integrate other relevant public welfare indicators to enrich the analysis and foster a more comprehensive understanding of welfare conditions across regencies/cities in West Java Province.

BIBLIOGRAPHY

[1] M. Musrafiyan, "Potensi Pembangunan Kawasan Ekonomi Khusus (KEK) Halal Barsela sebagai Destinasi Pariwisata Prioritas di Era Society 5.0," *PROCEEDINGS ICIS 2021*, vol. 1, no. 1, 2021.

[2] Badan Pusat Statistik Indonesia, "Indikator Kesejahteraan Rakyat 2024," 2024. Accessed: Mar. 12, 2025. [Online]. Available: https://www.bps.go.id/id/publication/2024/11/06/3ef10d3d82ed93f616ba9113/indikator-kesejahteraan-rakyat-2024.html

[3] Badan Pusat Statistik Provinsi Jawa Barat, "Provinsi Jawa Barat Dalam Angka 2024," Feb. 2024. [Online]. Available: https://jabar.bps.go.id/id/publication/2024/02/28/35ffe2d35104b39feb577e8f/provinsi-jawa-barat-dalam-angka-2024.html

[4] I. Wahyuni and S. P. Wulandari, "Pemetaan Kabupaten/Kota di Jawa Timur Berdasarkan Indikator Kesejahteraan Rakyat Menggunakan Analisis Cluster Hierarki," *JURNAL SAINS DAN SENI ITS*, vol. 11, no. 1, 2022.

[5] M. H. Asnawi and P. Rahmah, "Analisis Klaster Hirarki untuk Mengelompokan Provinsi di Indonesia berdasarkan Indikator Kesejahteraan Rakyat," *SEMINAR NASIONAL STATISTIKA X*, 2021, doi: 10.1234/pns.v10i.84.

[6] D. Andiani, S. Dwi, R. Septiani, and A. Riana, "Analisis Teknik non-Hierarki untuk Pengelompokan Kabupaten/Kota di Provinsi Jawa Barat Berdasarkan Indikator Kesejahteraan Rakyat 2020," *Jurnal Riset Matematika dan Sains Terapan*, vol. 21, no. 1, pp. 21–28, 2022.

[7] N. Oktaviani, A. Fauzan, and G. Widyastuti, "Pengelompokan Kabupaten/Kota di Jawa Barat Berdasarkan Tingkat Kesejahteraan Masyarakat Menggunakan K-Means Cluster," *Emerging Statistics and Data Science Journal*, vol. 2, no. 2, 2024.

[8] A. Eka Putra Haryanto, M. Ulfa Yanuar, D. Statistika Bisnis, and F. Vokasi, "Metode K-Means Clustering untuk Pengelompokan Kabupaten/Kota dalam Upaya Pengendalian Tingkat Inflasi di Pulau Jawa dan Sumatera K-Means Clustering Method for District/City Grouping in Effort to Control Inflation Rates in Java and Sumatera," pp. 29–42, 2022, doi: 10.21787/govstat.1.1.2022.29-42.

[9] J. Han, J. Pei, and T. Hanghang, *Data Mining: Concepts and Techniques*, 4th ed. Morgan Kaufmann, 2022.

[10] U. Syafiyah, D. Puspitasari, I. Asrafi, B. Wicaksono, and F. M. Sirait, "Analisis Perbandingan Hierarchical dan Non-Hierarchical Clustering Pada Data Indikator Ketenagakerjaan di Jawa Barat Tahun 2020," *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 803–812, Nov. 2022, doi: 10.34123/semnasoffstat.v2022i1.1221.

[11] Daniel Wicaksono Nugroho, Farhan Bramhatchi, Sri Pingit Wulandari, and Albertus Eka, "Pengelompokan Indikator Kesejahteraan Masyarakat Berdasarkan Kabupaten/Kota di Jawa Tengah Tahun 2023 Menggunakan Analisis Cluster," *Switch : Jurnal Sains dan Teknologi Informasi*, vol. 2, no. 5, pp. 87–101, Nov. 2024, doi: 10.62951/switch.v2i5.285.

[12] A. N. Alifah, H. N. Fadhilah, and T. M. Sianipar, "Klasterisasi Kabupaten/Kota di Jawa Barat Berdasarkan Tingkat Kenyamanan dengan Metode K-Means Clustering," *Seminar Nasional Sains Data*, vol. 2022.

[13] M. Cui and others, "Introduction to the k-means clustering algorithm based on the elbow method," *Accounting, Auditing and Finance*, vol. 1, no. 1, pp. 5–8, 2020.

[14] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J Comput Appl Math*, vol. 20, pp. 53–65, 1987.

[15] L. Hakim and A. Saefuddin, *Introduction to machine learning using R: konsep, teori, dan praktik*. IPB Press, 2022.

[16] A. R. Damayanti and A. W. Wijayanto, "Comparison of hierarchical and non-hierarchical methods in clustering cities in Java Island using the human development index indicators year 2018," *Eigen Mathematics Journal*, pp. 8–17, 2021.

[17] N. Thamrin and A. W. Wijayanto, "Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 141–160, Mar. 2021, doi: 10.29244/ijsa.v5i1p141-160.

[18] E. S. Barry, J. Merkebu, and L. Varpio, "State-of-the-art literature review methodology: A six-step approach for knowledge synthesis," *Perspect Med Educ*, vol. 11, no. 5, pp. 281–288, Oct. 2022, doi: 10.1007/s40037-022-00725-9.

[19] T. A. Munandar, *Data Mining Menggunakan R Teori dan Praktik*, 1st ed. Serang: PT Bale Damar Publishing, 2023.
[20] E. Suprihadi, *Machine Learning: Dasar dan Praktis*. Yogyakarta: Deepublish, 2022.
[21] P. Palinggik Allorerung, A. Erna, M. Bagussahrir, and S. Alam, "Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," 2024.