BAREKENG: Journal of Mathematics and Its Applications
September 2025 Volume 19 Issue 3 Page 2039-2056
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss3pp2039-2056

# A COMPARATIVE ANALYSIS OF DBSCAN AND GAUSSIAN MIXTURE MODEL FOR CLUSTERING INDONESIAN PROVINCES BASED ON SOCIOECONOMIC WELFARE INDICATORS

**Sri Andayani[1], Namita Retnani[2], Thesa Adi Saputra Yusri[3*], Bambang Sumarno Hadi Marwoto[4]**

[1,2,3,4]*Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta*
*Jln. Colombo No.1, Karang Malang, Caturtunggal, Sleman, Yogyakarta, 55281, Indonesia*

*Corresponding author's e-mail: * **thesaadisaputrayusri@uny.ac.id***

***Article History:***

*Received: 3rd February 2024*
*Revised: 11th March 2025*
*Accepted: 9th April 2025*
*Published: 1st July 2025*

***Keywords:***

*Clustering Analysis;*
*DBSCAN;*
*Gaussian Mixture Model (GMM);*
*Socioeconomic Welfare.*

*ABSTRACT*

*Public welfare refers to a condition in which people experience happiness, comfort, prosperity, and can adequately fulfill their basic needs. Indonesia consists of several provinces, each with varying levels of welfare. One crucial aspect in promoting equitable development is ensuring that all regions in Indonesia achieve similar welfare standards. This study aims to classify Indonesian provinces based on socioeconomic welfare indicators, with the results serving as a basis for policy-making that considers regional potential and challenges. The data used in this study are secondary data obtained from the official website of BPS-Statistics Indonesia on provincial welfare indicators from 2020 to 2023. The research methodology includes data collection, descriptive statistical analysis, determining the optimal number of clusters, and comparing the clustering performance of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and the Gaussian Mixture Model (GMM) using Silhouette Index, Davies-Bouldin Index, and Calinski-Harabasz Index as evaluation metrics. The DBSCAN-based clustering resulted in two clusters: high-welfare and low-welfare regions. Meanwhile, GMM clustering produced five clusters: moderate, fairly low, low, high, and fairly high welfare regions. Based on cluster validity measures, GMM outperformed DBSCAN, achieving a Silhouette score of 0.28, a Davies-Bouldin Index of 1.12, and a Calinski-Harabasz Index of 10.9.*



***How to cite this article:***

S. Andayani, N. Retnani, T. A. S. Yusri and B. S. H. Marwoto, "A COMPARATIVE ANALYSIS OF DBSCAN AND GAUSSIAN MIXTURE MODEL FOR CLUSTERING INDONESIAN PROVINCES BASED ON SOCIOECONOMIC WELFARE INDICATORS," *BAREKENG: J. Math. & App.*, vol. 19, no. 3, pp. 2039-2056, September, 2025.


Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com*; *barekeng.journal@mail.unpatti.ac.id*
**Research Article· Open Access**

## 1. INTRODUCTION

People's welfare refers to a condition in which individuals within a country experience prosperity and can adequately fulfill their basic needs. A high level of welfare signifies that both material and spiritual needs of the population are well met [1]. This welfare level can be assessed through seven key indicators: population, health, education, employment, consumption levels and patterns, housing and environment, and other social factors **[2]**.

The 2023 provincial welfare level can be evaluated based on factual data from each province **[3]**. Java Island remains the most populous region, housing 55.84% of Indonesia's total population. In terms of access to basic health services, Papua Province reports the lowest percentage (36.52%), whereas Bali Province has the highest (90.49%) **[4]**. Education levels, measured by the School Participation Rate (APS), are also highly varied, with Papua Province recording the lowest rate (65.53%) and Yogyakarta Province the highest (85.03%) **[5]**. Similarly, the employment indicator, represented by the unemployment rate, highlights disparities, with West Java Province having the highest unemployment rate (9.01%) and West Sulawesi the lowest (2.76%) **[6]**. Consumption patterns, as indicated by per capita expenditure (PPK), reveal that Papua Province has the lowest PPK (7,154.25 thousand rupiahs), whereas DKI Jakarta exhibits the highest (18,761.75 thousand rupiahs). Given these disparities, ensuring equitable development is a priority for the government, requiring a precise and data-driven approach to policy formulation. Identifying regional welfare characteristics is essential to implementing targeted and effective development strategies **[7]**. A robust method for this identification is clustering, which enables the grouping of provinces based on similar welfare indicators, facilitating comparative analysis and informed decision-making.

Clustering is a widely used technique for categorizing regions into groups based on similarity in selected indicators **[7]-[10]**. Broadly, clustering techniques are classified into hard clustering and soft clustering **[8]**. Hard clustering assigns each data point exclusively to a single cluster based on proximity to a cluster center. Popular algorithms in this category include K-Means and DBSCAN **[9]**. Soft clustering, in contrast, allows for probabilistic membership, meaning a single data point may belong to multiple clusters with different degrees of certainty. Examples of soft clustering techniques include Fuzzy C-Means and Gaussian Mixture Model (GMM) **[10]**. Previous research on clustering people's welfare indicators has utilized various methods. For example, Belinda applied the CEBMDC algorithm **[11]**, while Ambarsari compared the performance of Fuzzy C-Means and GMM **[12]**. Saputri examined the effectiveness of K-Means, K-Medoids, and DBSCAN **[13]**, whereas Dwitiyanti explored Fuzzy C-Means **[14]** and MMSDR **[15]**. However, there has been no recent study on clustering Indonesian provinces based on 2023 welfare data, leaving a gap in updated analysis for policy-making and development planning.

This study fills a research gap by analyzing Indonesian provincial welfare clustering using recent data, providing a timely and accurate classification for policy-making and regional development. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is well-suited for identifying clusters of arbitrary shapes and is robust to outliers, making it ideal for datasets with varying density distributions. In contrast, GMM (Gaussian Mixture Model) is a probabilistic approach that models data as a mixture of multiple Gaussian distributions, allowing for a more flexible representation of cluster boundaries. By comparing these two methods, this research aims to determine which clustering approach more effectively captures the underlying structure of welfare disparities across Indonesian provinces.

## 2. RESEARCH METHODS

### 2.1 Dataset

This study utilizes secondary data related to people's welfare indicators across Indonesian provinces, covering 13 variables derived from official publications by BPS-Statistics Indonesia **[3]**. The variables are categorized into seven key welfare aspects: population, health, education, employment, consumption levels and patterns, housing and environment, and other social indicators. The dataset spans from 2020 to 2023, and the values are averaged over this period to ensure consistency before conducting clustering analysis. The variables used in this research are described in **Table 1**, which defines each welfare indicator in terms of its measurement and relevance to socio-economic conditions. Each variable represents a critical aspect

of welfare and is analyzed to identify patterns of regional disparity, providing insights into economic and social conditions across Indonesian provinces.

**Table 1. Operational Definition of Research Variables**

| No | Indicator | Variables | Description |
|---|---|---|---|
| 1 | Population | Population Density (KP) | Total population per $km^2$ (inhabitants/$km^2$) |
| 2 | Health | Access to Basic Health Facilities (AFKD) | Percentage of basic health facility utilization (%) |
| | | Life Expectancy Rate (AHH) | Average life expectancy (years) |
| 3 | Education | School Participation Rate (APS) | Percentage of school enrollment rate (%) |
| | | Average Years of Schooling (RLS) | Average years of schooling (years) |
| 4 | Employment | Open Unemployment Rate (TPT) | Percentage of unemployment (%) |
| | | Labor Force Participation Rate (TPAK) | Percentage of labor force participation rate (%) |
| | | Percentage of Employment (PLK) | Employment Availability Rate (%) |
| 5. | Other Social | Percentage of Poor Population (PPM) | Percentage of poor population (%) |
| | | Number of tourist trips (RPW). | Tourist Mobility Ratio (%) |
| 6 | Consumption levels and patterns | Per Capita Expenditure (PPK) | Per capita expenditure rate (Thousand Rupiah /capita/year) |
| | | Gross Regional Domestic Product GRDP) | GRDP figures (million rupiah) |
| 7 | Housing and Environment | Percentage of Affordable Housing (PPL) | Proportion of housing units in a region that meet the criteria for affordability, relative to the total number of housing units (%) |

## 2.2 Data Collection

The data used in this study is obtained from official publications by the Indonesian Central Bureau of Statistics (BPS), ensuring reliability and standardization across provinces. The dataset consists of people's welfare indicators recorded from 2022 to 2023, which are averaged to provide a more stable representation of each province's socio-economic condition. The data collection process involves gathering information on population density, health access, education levels, employment, consumption patterns, housing conditions, and other social factors, which are then compiled into a structured dataset. Before proceeding with clustering analysis, the data undergoes preprocessing, including handling missing values and normalizing different scales to ensure comparability. By utilizing secondary data from BPS, this research ensures consistency in measurement across all provinces, allowing for an objective and data-driven approach in analyzing welfare disparities in Indonesia.

## 2.3 Clustering Analysis

This research applies two clustering methods, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and Gaussian Mixture Model (GMM), and compares their performance in identifying clusters of Indonesian provinces based on welfare indicators.

### 2.3.1 Density-Based Spatial Clustering of Application with Noise (DBSCAN)

Density-Based Spatial Clustering of Application with Noise Algorithm (DBSCAN) is one of the density-based clustering methods. This method performs cluster formation by determining areas based on high density separated by low density according to cluster density. DBSCAN can be used to identify clusters with various forms including noise and outliers [16]. DBSCAN involves two crucial parameters [17] that significantly impact clustering performance:

a. Minimum Points (minPts). This parameter represents the minimum number of neighboring points required to form a cluster.

b. Epsilon (ε) This defines the radius of the neighborhood around a point. The epsilon value can be determined using a k-distance graph, where the optimal epsilon is identified at the point where the

graph exhibits a sharp bend. To determine the epsilon value using the k-distance graph, it is essential to first compute the Euclidean distance and establish the appropriate minPts value.

Once the minPts and epsilon (ε) parameters are determined, clustering analysis using DBSCAN can be performed. The following are the steps for conducting clustering analysis with DBSCAN [16]:

1. Determine the value of minimum points (minPts) using ln($n$) and epsilon ($\varepsilon$) using the k-distance graph. In determining the minPts parameter for DBSCAN, this study adopts the widely used heuristic of setting minPts = ln($n$), where n is the number of data points in the dataset. This approach is recommended in several clustering studies as a practical guideline, particularly when no prior domain-specific knowledge is available to set minPts manually [18].
2. Select an initial point ($p$) randomly to start the clustering process.
3. Calculate the distance between point ($p$) and all data observations. In this study, the distance calculation is performed using the Euclidean distance formula [19].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where $x$ and $y$ are data points, and n represents the number of dimensions. In this study, PCA was applied to reduce the 13 original variables to 2 principal components, so n = 2 in the distance calculation. This reduction aims to retain the most significant variance while improving clustering performance and interpretability.

4. All data points that satisfy the minPts and epsilon conditions relative to point ($p$), based on the distance calculation, will be grouped into a cluster with ($p$) as the cluster center.
5. If a cluster cannot be formed because the minPts and epsilon conditions are not met for ($p$), another point is selected as the new initial point for cluster formation. Then, steps 3 and 4 are repeated until all data observations have been processed.
6. Data points that do not satisfy the minPts and epsilon conditions for any analyzed cluster center will be classified as outliers. These outliers can be identified based on their differences from other data points that belong to a cluster.

### 2.3.2 Gaussian Mixture Model (GMM)

The Gaussian Mixture Model (GMM) is a probabilistic clustering algorithm that utilizes weighted combinations of multiple normal distributions, making it commonly referred to as a mixture model [20]. GMM is defined as a set of component functions representing density functions. When multiple variables are involved, it becomes a multivariate Gaussian density function. GMM consists of multiple Gaussian components, and its mathematical formulation can be expressed as follows [19]:

$$p(x_i|\pi,\mu,\Sigma) = \sum_{k=1}^{K}\pi_k N(x_i|\mu_k,\Sigma_k), i = 1,2,\dots,n \quad (2)$$

In the Gaussian Mixture Model (GMM), each data point $x_i$ in the dataset is represented as part of a probabilistic distribution composed of multiple Gaussian components. Each component $k$ is characterized by specific parameters: $\mu_k$, which denotes the mean vector of the $k$-th Gaussian component; $\Sigma_k$, the covariance matrix that defines the spread and orientation of the distribution; and $\pi$, the mixing coefficient, which represents the proportion of each Gaussian component in the overall mixture. The total number of Gaussian components in the model is denoted by $K$, determining the complexity and flexibility of the clustering representation.

In the context of Gaussian Mixture Models (GMM), the EM algorithm refines model parameters iteratively to improve clustering accuracy [21]. The process consists of the following steps.

1. Set the initial values for the mean vector $\mu_k$, covariance matrix $\Sigma_k$, and mixing coefficient $\pi_k$ for each Gaussian component $k$.
2. Perform the Expectation Step (E-Step)

   Compute the responsibility $\gamma_{nk}$ , which represents the probability that data point $x_n$ belongs to the Gaussian component $k$, using Bayes' theorem:

$$\gamma_{nk} = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\Sigma_j N(x_n|\mu_k, \Sigma_k)} \quad (3)$$

where $N(x_n|\mu_k, \Sigma_k)$ is the probability density function of a multivariate Gaussian distribution. $\gamma_{nk} \in [0,1]$, and $\sum_{k=1}^{K} \gamma_{nk} = 1$ for each data point $n$.

For example, if $\gamma_{n1} = 0.86$, $\gamma_{n2} = 0.09$, and $\gamma_{n3} = 0.05$, then observation $x_n$ is most strongly associated with cluster 1, but still has a small probability of belonging to clusters 2 or 3. This soft assignment framework allows GMM to handle overlapping clusters more flexibly than hard clustering methods like k-means.

3. Perform the Maximization Step (M-Step)

The Maximization Step (M-step) follows the Expectation Step (E-step) in the Expectation-Maximization (EM) Algorithm for Gaussian Mixture Model (GMM) clustering. In this step, the algorithm updates the parameters of the Gaussian components by maximizing the log-likelihood function based on the probabilities (responsibilities) computed in the E-step.

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x_n \,, \quad (4)$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x_n - \mu_k)(x_n \mu_k)^T, \quad (5)$$

$$\Pi_k = \frac{N_k}{N} \quad (6)$$

After the iterative Expectation-Maximization (EM) optimization of the GMM parameters $\mu_k$ (mean vector), $\Sigma_k$ (covariance matrix), and $\pi_k$ (mixing coefficient), each data point is assigned to the cluster with the highest posterior probability of belonging to one of the $k$ Gaussian components. The final cluster label for each observation is determined by the maximum a posteriori.

The mixing coefficients $\pi_k$ provide additional interpretive value by indicating the relative size or prevalence of each cluster in the population. Clusters with larger $\pi_k$ values represent denser regions in the data space, while those with smaller values represent rarer or more specific subgroups. Overall, the optimized parameters not only determine the probabilistic boundaries of clusters but also help quantify their spread (via $\Sigma_k$) and relative significance (via $\pi_k$) in the dataset.

The number of components $K$ plays a crucial role in defining the clustering structure. Selecting an optimal $K$ is essential to ensure the model adequately represents the data without overfitting or underfitting. One commonly used method for determining the optimal number of clusters is the Bayesian Information Criterion (BIC).

The BIC is a model selection criterion that balances model complexity and goodness of fit. The optimal number of clusters is determined by selecting the value of $K$ that results in the largest BIC value. The BIC formula is given as follows **[22]**:

$$BIC = -2\ln(L) + P\ln(n) \quad (7)$$

where $L$ represents the likelihood function of the GMM model, $p$ denotes the total number of model parameters, and $N$ is the total number of data points in the dataset. The first term (-2 ln $L$) evaluates how well the model fits the data, while the second term ($P$ ln $n$) introduces a complexity penalty, preventing overfitting by discouraging excessive components. The optimal $K$ is selected based on the highest BIC value across multiple model evaluations. This criterion ensures that the model achieves a balance between accuracy and simplicity, making it a widely used method for determining the appropriate number of Gaussian components in GMM-based clustering.

### 2.3.3 Model Performance Evaluation

To determine the best clustering approach, the following validation indices are used.

1. Silhouette Score (SI): Measures how well-separated clusters are.
2. Davies-Bouldin Index (DBI): Evaluates cluster compactness and separation:

$$DB = \frac{1}{C}\sum_{i=1}^{C} \max_{j \neq 1}(R_{ij}) \quad (8)$$

where $DB$ represents the overall DBI score for the clustering model, C is the total number of clusters, and $R_{ij}$ measures the similarity between cluster $c_i$ and cluster $c_j$. Lower DBI values indicate better clustering.

3. Calinski-Harabasz Index (CHI): Measures cluster separation and compactness:

$$\text{CHI} = \frac{\text{SSB}/(\text{k}-1)}{\text{SSW}/(\text{N}-\text{C})} \quad (9)$$

where SSB is the between-cluster sum of squares, SSW is the within-cluster sum of squares, C is the number of clusters, and N is the total number of data points. A higher CHI value indicates better clustering quality.

By comparing these indices across DBSCAN and GMM, the study determines which method better captures welfare disparities among Indonesian provinces.

### 2.3.4 Clustering Implementation and Interpretation

Once the optimal clustering model is identified, provinces are grouped based on welfare indicators, and each cluster is interpreted based on its characteristics. The final step involves visualizing the clusters on a geographical map to assess their spatial distribution, allowing policymakers to target development efforts more effectively. This methodological approach ensures that the clustering results are statistically valid, interpretable, and applicable for socio-economic analysis, providing valuable insights into regional welfare disparities in Indonesia.

## 3. RESULTS AND DISCUSSION

### 3.1 Descriptive Statistics

Descriptive statistics is a statistical analysis that presents an overview of the characteristics of each research variable seen from the minimum value, average value (*mean)*, and maximum value. The data used in this study are data on 34 provinces before the division into 38 provinces.

**Table 2. Descriptive Statistics**

| Variables | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| KP | 9.50 | 55.06 | 102.25 | 746.79 | 265.62 | 16,047.25 |
| AFKD | 36.52 | 74.93 | 79.50 | 77.95 | 84.69 | 90.49 |
| AHH | 65.53 | 68.98 | 70.37 | 70.37 | 71.85 | 75.11 |
| APS | 63.12 | 71.97 | 74.33 | 74.30 | 75.99 | 85.03 |
| RLS | 7.17 | 8.58 | 9.22 | 9.20 | 9.73 | 11.27 |
| TPT | 2.76 | 4.26 | 4.83 | 5.28 | 6.25 | 9.01 |
| TPAK | 63.19 | 65.44 | 68.92 | 68.54 | 70.27 | 76.35 |
| PLK | 35.15 | 54.91 | 60.95 | 60.13 | 65.02 | 82.23 |
| PPM | 4.28 | 6.35 | 8.65 | 10.38 | 12.57 | 26.52 |
| PPK | 7,154 | 9,636 | 10,863 | 10,999 | 11,802 | 18,762 |
| PDRB | $4.61\times 10^7$ | $1.24\times 10^8$ | $2.33\times 10^8$ | $5.29\times 10^8$ | $5.68\times 10^8$ | $3.08\times 10^9$ |
| PPL | 43.17 | 84.34 | 91.54 | 85.49 | 94.28 | 98.80 |
| RPW | 0.05 | 0.25 | 0.69 | 0.94 | 1.63 | 25.46 |

**Table 2** shows that each variable has a different format or unit, therefore it is necessary to standardize so that each variable has a uniform format so that the results of data analysis are more accurate **[23]**.

After the standardization process, a multicollinearity test will be conducted to see the correlation between variables. The following are the results of the multicollinearity test:

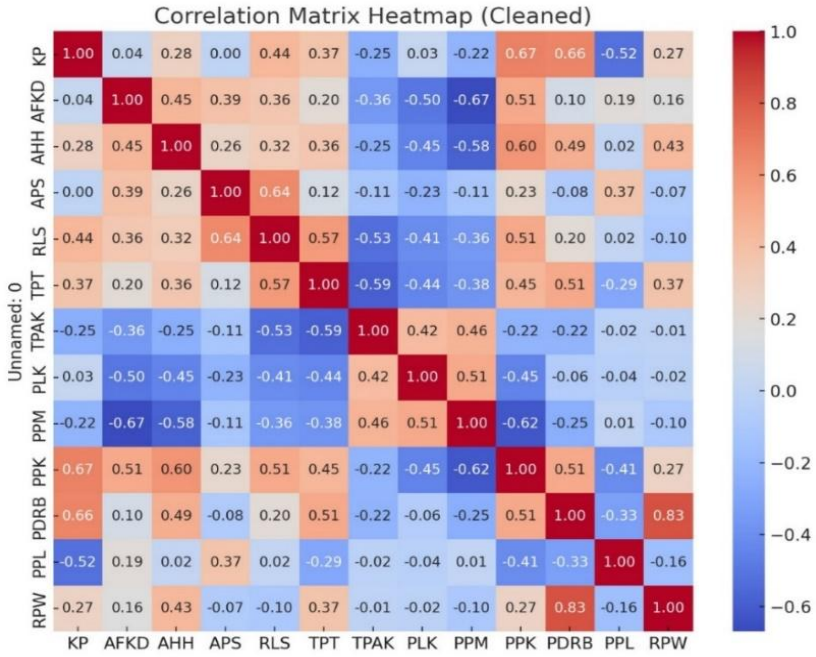



**Figure 1. Feature Correlation Matrix**

**Figure 1** shows that the correlation between AFKD and KP is 0.04, which is significantly smaller than the 0.85 threshold commonly used to indicate strong multicollinearity **[23]**, **[24]**, as well as for other variable pairs. This suggests that there is no strong correlation between variables in this study, meaning that multicollinearity is not a concern. Since multicollinearity occurs when two or more independent variables are highly correlated, leading to redundancy in regression models **[25]**, the correlation values presented in the heatmap indicate that each variable retains its unique contribution to the model.

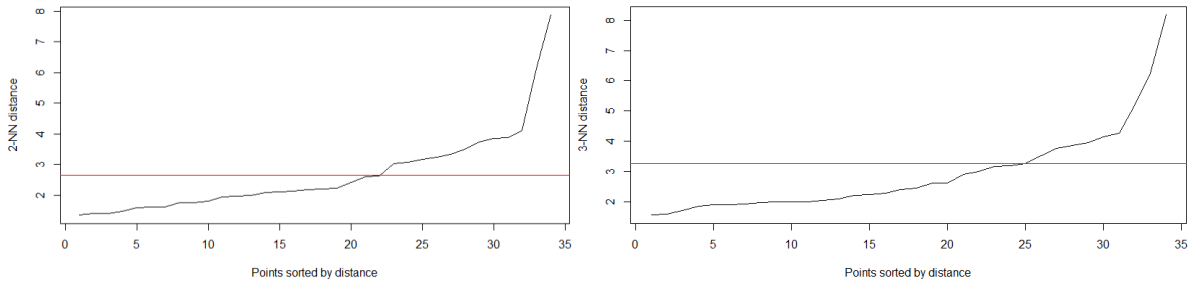
The assumption of multicollinearity has thus been met, demonstrating that the independent variables do not exhibit excessive interdependence. This validation is crucial for ensuring statistical reliability in regression and machine learning models, as high multicollinearity can distort coefficient estimates and reduce model interpretability **[17]**.

## 3.2 Cluster analysis with DBSCAN

### 3.2.1 Determine minPts and *Epsilon*

The epsilon value is used as a neighborhood radius that determines whether one data with other data can become a cluster or not. The epsilon value can be known after calculating the Euclidian distance and determining the minimum points (minPts).

Based on **Figure 2** (a), it can be observed that when the minPts value is set to 2, the optimal epsilon ($\varepsilon$) value that can be used is 2.65. This is determined by identifying the elbow point in the 2-NN distance plot, where the distance starts increasing significantly, indicating the transition between dense regions and sparser areas. Similarly, based on **Figure 2** (b), when the minPts value is set to 3, the optimal epsilon ($\varepsilon$) value that can be used is 3.25. The 3-NN distance plot follows a similar trend, where the elbow point marks a significant increase in the nearest neighbor distances, suggesting a suitable $\varepsilon$ threshold for DBSCAN clustering. These epsilon values are crucial in density-based clustering, as they determine the neighborhood radius used to form clusters while distinguishing core points from noise.




**(a)** **(b)**

**Figure 2. *k*-Distance Graphs for Determining Epsilon ($\varepsilon$) in DBSCAN, (a) 2-NN Distance Plot, (b) 3-NN Distance Plot**

### 3.2.2 Perform Clustering

Table 3 shows the clustering results obtained using the DBSCAN method under different parameter settings. When the epsilon value is set to 2.65 and minPts is set to 2, the algorithm identifies three clusters. However, when the epsilon value is increased to 3.25 and minPts is set to 3, the algorithm detects only two clusters. This outcome aligns with the nature of DBSCAN, where the choice of epsilon ($\varepsilon$) and minPts significantly affects the clustering structure. A lower epsilon value allows the formation of smaller, denser clusters, whereas increasing epsilon tends to merge clusters, reducing the overall number of detected groups. Similarly, increasing minPts requires more points to form a core cluster, making the algorithm more restrictive and potentially leading to fewer clusters.

**Table 3. DBSCAN Cluster Results**

| *Epsilon* | MinPts | Number of clusters |
|---|---|---|
| 2.65 | 2 | 3 |
| 3.25 | 3 | 2 |

These findings highlight the sensitivity of DBSCAN to parameter selection, reinforcing the importance of fine-tuning epsilon and minPts to achieve meaningful clustering results based on the dataset's density distribution.

### 3.2.3 Determining the Optimal Cluster

The best clustering configuration is determined by considering optimal conditions, characterized by the highest Silhouette value, the lowest Davies-Bouldin Index (DBI) value, and the highest Calinski-Harabasz Index (CHI) value.

**Table 4. DBSCAN Optimal Cluster**

| Eps | Min-PTS | Silhouette Value | Davies-Bouldin Index | Calinski-Harabasz Index | Number of clusters |
|---|---|---|---|---|---|
| 2.65 | 2 | 0.24 | **0.69** | 3.24 | 2 |
| 3.25 | 3 | **0.3** | 0.94 | **5.27** | 2 |

Table 4 presents the optimal cluster results for DBSCAN under different parameter settings. When the epsilon value is set to 2.65 and minPts is set to 2, the clustering model achieves a Silhouette score of 0.24, a Davies-Bouldin Index of 0.69, and a Calinski-Harabasz Index of 3.24, forming two clusters. Meanwhile, when the epsilon value is increased to 3.25 and minPts is set to 3, the clustering results improve, with a Silhouette score of 0.3, a Davies-Bouldin Index of 0.94, and a Calinski-Harabasz Index of 5.27, also forming two clusters.

From these results, the optimal clustering choice depends on the evaluation criteria used. The highest Silhouette value (0.3) and the highest Calinski-Harabasz Index (5.27) suggest that the best clustering configuration is obtained when epsilon = 3.25 and minPts = 3. However, based on the Davies-Bouldin Index, the lowest value (0.69) is achieved when epsilon = 2.65 and minPts = 2, which would suggest a better cluster separation in this case. Since both configurations result in two clusters, the final choice is

based on prioritizing Silhouette and Calinski-Harabasz Index, leading to the conclusion that the optimal clustering configuration is when epsilon = 3.25 and minPts = 3, with the final number of clusters being two.

### 3.2.4 Cluster Analysis Results with DBSCAN

**Figure 3** presents a Principal Component Analysis (PCA) cluster plot generated using the DBSCAN clustering method. The $x$-axis (Dim1: 37.6%) and $y$-axis (Dim2: 19.1%) represent the first two principal components, which capture the majority of the variance in the dataset. The first principal component (Dim1) is primarily influenced by variables such as (e.g., per capita expenditure, HDI, internet access), suggesting it captures economic well-being and infrastructure access. The second component (Dim2) is more strongly associated with (e.g., employment rate, affordable housing), indicating a focus on social inclusion and labor market conditions. These interpretations of Dim1 and Dim2 help contextualize the clusters observed in the DBSCAN output, where provinces group based on shared patterns of socio-economic characteristics.

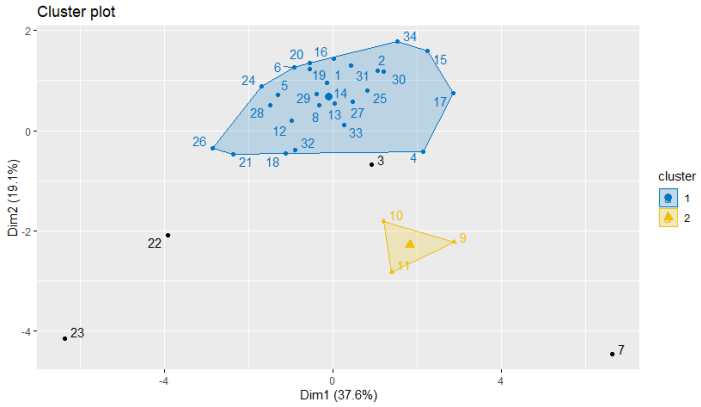



**Figure 3. PCA-Based Cluster Plot Using DBSCAN Method**

This visualization provides insights into the structure and separability of clusters formed by DBSCAN.

- Cluster 1 (Blue):

  This cluster contains the majority of data points and exhibits a wider spread across the PCA space. The convex hull surrounding the points indicates that Cluster 1 has greater variance, suggesting a more diverse distribution of data points within this group.

- Cluster 2 (Yellow):

  This smaller cluster consists of fewer data points, forming a compact grouping with less variance compared to Cluster 1. The triangular shape enclosing these points reflects the tight density of this cluster, implying that DBSCAN has identified this as a distinct subgroup with shared characteristics.

- Outliers (Black Dots):

  Several black points (e.g., data points 7, 22, and 23) are isolated and not enclosed within any cluster, indicating that DBSCAN has classified them as noise or anomalies. These points are located far from the main clusters, suggesting that they do not meet the minPts and epsilon criteria required to be part of any defined cluster.

**Table 5. Average of Each Variable in DBSCAN Cluster Results**

| Variables | Cluster 1 | Cluster 2 |
|---|---|---|
| KP | 96.5 | **1,105** |
| AFKD | 79.53 | **82.36** |
| AHH | 70.19 | **73.27** |
| APS | **75.05** | 72.84 |
| RLS | **9.32** | 8.61 |
| TPT | 5.15 | **6.75** |
| TPAK | 68.21 | **68.98** |
| PLK | 59.00 | **59.74** |
| PPM | 9.88 | 10.01 |
| PPK | 10,840 | 11,471 |

| Variables | Cluster 1 | Cluster 2 |
|---|---|---|
| PDRB | 302,010,633 | **2,149,799,822** |
| PPL | **90.56** | 82.56 |
| RPW | 1.11 | **20.39** |

Based on **Table 5**, the DBSCAN clustering results reveal two distinct clusters characterized by different levels of people's welfare. The labeling of these clusters is determined based on key socioeconomic indicators such as population density (KP), life expectancy (AHH), education levels (APS and RLS), employment (TPT and TPAK), income levels (PPK and PDRB), and social welfare metrics (PPL and RPW).

Cluster 1 represents regions where people's welfare tends to be lower compared to Cluster 2. This is evident from several key indicators, including lower population density (96.5 people/km²), lower life expectancy (70.19%), and lower per capita income (10,840 thousand rupiahs per year). Additionally, social protection levels (PPL) in this cluster are higher (90.56%), which may indicate a greater reliance on government assistance programs.

In contrast, Cluster 2 comprises areas where people's welfare tends to be higher. The population density (1,105 people/km²) is significantly greater than in Cluster 1, indicating more urbanized regions. Higher life expectancy (73.27%), a slightly lower school participation rate (APS 72.84%), and a higher per capita income (11,471 thousand rupiahs per year) suggest that economic conditions and access to resources are relatively better in this cluster. Moreover, the Gross Regional Domestic Product (PDRB) in Cluster 2 (2,149,799,822 million rupiahs) is considerably higher than in Cluster 1, reflecting stronger economic activity.

The welfare disparity between the two clusters is further highlighted by the percentage of poor households (RPW), where Cluster 1 has a lower poverty rate (1.11%), whereas Cluster 2 exhibits a higher poverty rate (20.39%), possibly due to higher living costs and economic disparities in more developed regions. These results demonstrate that DBSCAN successfully differentiates regions based on socioeconomic characteristics, offering valuable insights for targeted policy interventions and development planning.

**Table 6. Provincial Grouping with the DBSCAN Method**

| Cluster | Province |
|---|---|
| Cluster 1 (27 provinces) | Aceh, Bali, Banten, Bengkulu, Gorontalo, Jambi, West Kalimantan, South Kalimantan, Central Kalimantan, East Kalimantan, North Kalimantan, Riau Islands, Lampung, Maluku, North Maluku, West Nusa Tenggara, West Papua, Riau, West Sulawesi, South Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi, West Sumatra, South Sumatra, North Sumatra, Special Region of Yogyakarta. |
| Cluster 2 (3 provinces) | West Java, Central Java, East Java |
| Outliers (4 provinces) | Kep. Bangka Belitung, DKI Jakarta, East Nusa Tenggara, Papua |

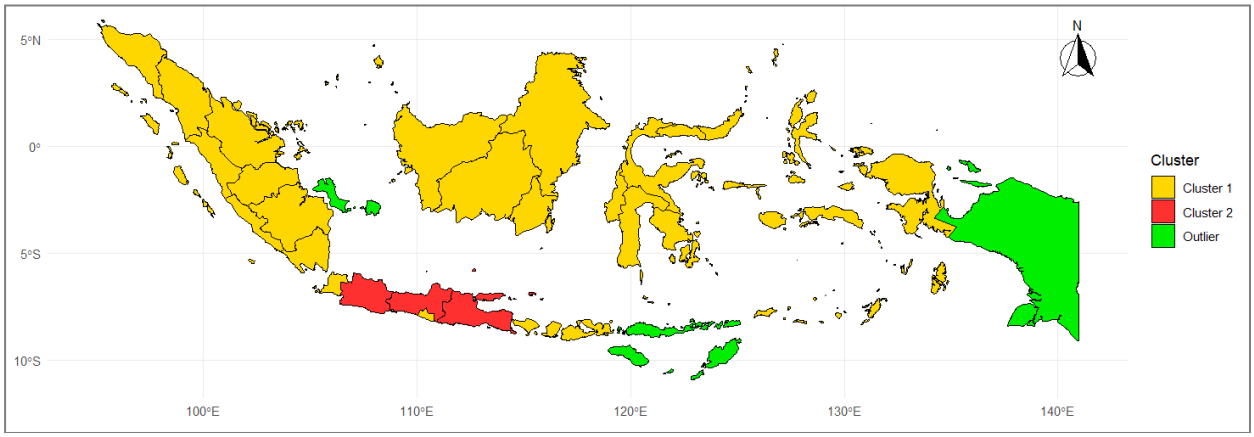



**Figure 4. Clustering of Indonesian Provinces Based on People's Welfare Using DBSCAN**

**Figure 4** presents a visualization of DBSCAN clustering results, categorizing Indonesian provinces based on people's welfare indicators. The provinces are classified into two clusters, with additional provinces identified as outliers. The color-coding represents:

- Cluster 1 (Light Yellow): Provinces with a lower level of people's welfare, characterized by lower population density, lower income levels, and lower access to essential services.
- Cluster 2 (Light Red): Provinces with a higher level of people's welfare, predominantly located in high-density and economically active regions such as Java Island. These regions exhibit higher GDP per capita, better education, and more developed infrastructure.
- Outliers (Light Green): Provinces that do not fit neatly into the identified clusters. Papua, for instance, is categorized as an outlier, likely due to unique socio-economic characteristics, lower population density, and geographical disparities.

The spatial distribution of these clusters highlights significant economic and welfare disparities across Indonesia. Java and parts of Sumatra, being economic hubs, are predominantly classified in Cluster 2, reflecting higher economic activity and better social indicators. In contrast, many provinces in Kalimantan, Sulawesi, and the eastern parts of Indonesia fall under Cluster 1, suggesting lower economic and social welfare conditions.

The presence of outliers, particularly in Papua and several smaller islands, suggests unique socio-economic conditions that deviate from the general clustering patterns. These areas might require special policy interventions, such as targeted economic programs or infrastructure development, to enhance their welfare levels.

### 3.3 Cluster Analysis with GMM

#### 3.3.1 Perform Clustering

This research performs clustering with the number of clusters ranging from 2 to 5, as increasing the number of clusters beyond 5 results in a higher Bayesian Information Criterion (BIC) value, indicating poor model performance. **Table 7** presents the BIC scores for different cluster numbers, where the highest BIC value (-1011.925) is observed when the data is clustered into five groups, suggesting that this configuration provides the best balance between model complexity and data fit.

**Table 7. Cluster Analysis with GMM**

| Cluster | BIC |
|---|---|
| 2 | -1051.872 |
| 3 | -1065.469 |
| 4 | -1086.006 |
| 5 | **-1011.925** |

The Bayesian Information Criterion (BIC) is a widely used statistical metric for model selection, taking into account both goodness of fit and model complexity. A higher BIC value signifies a better model, as it suggests that the model explains the data well while avoiding overfitting. Based on **Table 7**, clustering using Gaussian Mixture Model (GMM) achieves its optimal configuration when divided into five clusters, as this setting results in the highest BIC score, ensuring a well-structured classification of provinces based on people's welfare indicators.

#### 3.3.2 Determining the Optimal Cluster

**Table 8** presents the evaluation metrics for different numbers of clusters using the Gaussian Mixture Model (GMM). The clustering performance is assessed based on three key validation indices: Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The Silhouette Score measures the compactness and separation of clusters, where a higher value indicates better-defined clusters. The Davies-Bouldin Index evaluates the similarity between clusters, where a lower value indicates better

separation. The Calinski-Harabasz Index assesses the ratio of between-cluster variance to within-cluster variance, where higher values indicate better clustering structure.

**Table 8. GMM Optimal Cluster**

| Total cluster | Value Silhouette | Davies Value-Bouldin Index | Calinski Value-Harabasz Index |
|---|---|---|---|
| 2 | 0.26 | 2.64 | 4.58 |
| 3 | 0.14 | 1.89 | 7.21 |
| 4 | 0.23 | 1.32 | 9.08 |
| 5 | **0.28** | **1.12** | **10.90** |

Based on **Table 8**, the optimal clustering configuration is achieved when the number of clusters is 5, as it yields the highest Silhouette value (0.28), the lowest Davies-Bouldin Index (1.12), and the highest Calinski-Harabasz Index (10.9). These results indicate that dividing Indonesian provinces into five clusters using the GMM method provides the best clustering performance, effectively distinguishing regional differences in people's welfare indicators.

### 3.3.3 Results of Cluster Analysis with GMM

**Figure 5** presents a Principal Component Analysis (PCA) cluster plot generated using the Gaussian Mixture Model (GMM) clustering method, where the $x$-axis (Dim1: 37.6%) and $y$-axis (Dim2: 19.1%) represent the first two principal components, capturing the main variance in the dataset.

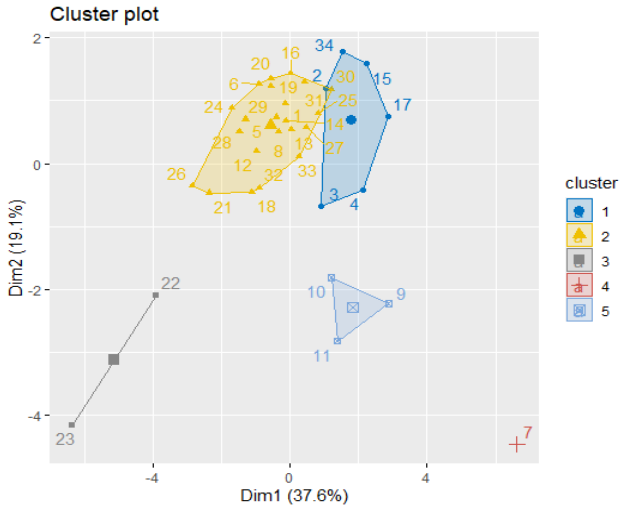



**Figure 5. PCA-Based Cluster Plot Using Gaussian Mixture Model (GMM)**

The visualization shows five distinct clusters: Cluster 1 (Blue, Hexagonal Shape) represents a compact and well-separated group of provinces with similar characteristics, while Cluster 2 (Yellow, Triangular Shape) consists of a larger and more dispersed cluster, indicating moderate variability among its provinces. Cluster 3 (Gray, Connected Line) comprises a few provinces with distinct socio-economic characteristics, and Cluster 4 (Red, Single Point) represents an outlier province with extreme values. Cluster 5 (Light Blue, Small Triangle) forms a compact and closely related subgroup. These clustering results align with previous findings, where five clusters provided the best clustering performance based on the Bayesian Information Criterion (BIC), Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The PCA visualization further supports the validity of GMM-based clustering, demonstrating its ability to flexibly assign provinces to clusters compared to hard clustering methods such as K-Means.

**Table 9. Average of Each Variable in Cluster Results with GMM**

| Variables | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| KP | 600.8 | 104.40 | 64 | **16,047** | 1105 |
| AFKD | **84.49** | 78.59 | 44.63 | 78.08 | 82.36 |
| AHH | 72.32 | 69.64 | 66.76 | **73.27** | 73.27 |
| APS | **76.28** | 74.47 | 68.98 | 73.85 | 72.84 |
| RLS | 9.67 | 9.19 | 7.68 | **11.27** | 8.61 |

| Variables | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|---|
| TPT | 6.25 | 4.83 | 3.48 | **8.29** | 6.76 |
| TPAK | 68.85 | 67.99 | **75.41** | 63.68 | 68.98 |
| PLK | 48.31 | 61.56 | **78.84** | 63.3 | 59.74 |
| PPM | 6.53 | 10.55 | **23.50** | 4.60 | 10.01 |
| PPK | 13,436 | 10,236 | 7,487 | **18,762** | 11,471 |
| PDRB | 377,113,597 | 271,884,481 | 102,780,023 | **3,077,919,696** | 2,149,799,822 |
| PPL | 76.77 | **92.17** | 63.20 | 44.39 | 82.56 |
| RPW | 2.04 | 0.80 | 0.28 | 8.38 | **20.39** |

Based on **Table 9**, the cluster labeling of the analysis results with the GMM method is as follows:

- Cluster 1: **medium** welfare cluster

  Cluster 1 is an area with a balance of people's welfare indicators. Each positive indicator has a high value and the negative indicators have low numbers. This indicates good community welfare in cluster 1.

- Cluster 2: clusters with **moderately low** welfare levels

  Cluster 2 is an area that scores lower on the people's welfare indicators than cluster 1, cluster 4, and cluster 5, but better than cluster 3.

- Cluster 3: clusters with **low** welfare levels

  Cluster 3 is an area of particular concern because it has the lowest poverty, income, and access to basic health facilities of the other clusters.

- Cluster 4: clusters with **high** welfare levels

  Cluster 4 is the region with the highest per capita expenditure, gross regional domestic product, education level and population density. In addition, the percentage of poor people in cluster 4 is the lowest. However, cluster 4 has the highest unemployment rate and the lowest percentage of decent housing. This shows that the population density in cluster 4 causes an imbalance between the number of residents and the number of jobs and decent housing.

- Cluster 5: clusters with a **high** level of welfare

  Cluster 5 is an area that has a fairly good number of indicators of people's welfare. With a high population density, cluster 5 has the highest gross regional domestic product, ease of access to basic health facilities, and percentage of decent housing, as well as the highest number of tourist trips. However, the unemployment and poverty rates in cluster 5 are also quite high and the number of jobs is the second lowest.

**Table 10. Provincial Grouping with GMM Method**

| Cluster | Province |
|---|---|
| Cluster 1 (6 provinces) | Bali, Kep. Bangka Belitung, Banten, East Kalimantan, Kep. Riau, Special Region of Yogyakarta |
| Cluster 2 (22 provinces) | Aceh, Bengkulu, Gorontalo, Jambi, West Kalimantan, South Kalimantan, Central Kalimantan, North Kalimantan, Lampung, Maluku, North Maluku, NTB, West Papua, Riau, West Sulawesi, South Sulawesi, Central Sulawesi, Southeast Sulawesi, North Sulawesi, West Sumatra, South Sumatra, North Sumatra |
| Cluster 3 (2 provinces) | NTT, Papua |
| Cluster 4 (1 province) | DKI Jakarta |
| Cluster 5 (3 provinces) | West Java, Central Java, East Java |

**Figure 6** presents a map visualization of Gaussian Mixture Model (GMM) clustering results, categorizing Indonesian provinces based on people's welfare indicators into five distinct clusters. Cluster 1 (Blue) represents the medium welfare cluster, consisting of six provinces (Bali, Kepulauan Bangka Belitung, Banten, East Kalimantan, Kepulauan Riau, and Special Region of Yogyakarta). This cluster is characterized by a balance in welfare indicators, where positive indicators such as income and education

levels are relatively high, while negative indicators like unemployment and poverty remain low, indicating a generally good level of community welfare. Cluster 2 (Yellow) represents provinces with moderately low welfare levels, covering 22 provinces, including Aceh, Jambi, Kalimantan, Sulawesi, Sumatra, and Maluku. These regions have lower welfare indicators than Clusters 1, 4, and 5 but still perform better than Cluster 3, indicating the need for further development policies to improve welfare conditions.

Cluster 3 (Red) is the low welfare cluster, comprising only two provinces (NTT and Papua), which are among the most vulnerable regions in terms of income levels, poverty rates, and access to basic health facilities. These areas require special government attention and intervention to improve economic and social welfare conditions. Cluster 4 (Green) consists solely of DKI Jakarta, which stands out with the highest per capita expenditure, GDP, education levels, and population density. However, despite its economic dominance, Jakarta also experiences the highest unemployment rate and the lowest percentage of decent housing, highlighting the challenges of overpopulation and economic disparity in urban areas.

Meanwhile, Cluster 5 (Pink) includes three provinces (West Java, Central Java, and East Java), representing a high-welfare cluster. These provinces have the highest GDP, best access to health facilities, and the highest percentage of decent housing, along with the highest number of tourist visits, contributing to economic growth. However, despite these strengths, this cluster also has high unemployment and poverty rates, with the second-lowest number of available jobs, suggesting the need for more employment opportunities and workforce distribution strategies.

The spatial distribution of these clusters illustrates the economic and welfare disparities across Indonesia, where densely populated economic hubs (Clusters 4 and 5) generate high economic output but also struggle with structural issues such as housing shortages and employment imbalances. Meanwhile, underdeveloped provinces in Cluster 3 require focused economic policies and social assistance programs to enhance their living standards. The GMM clustering method effectively groups provinces with similar socio-economic characteristics, providing valuable insights for policymakers to implement targeted development strategies that reduce inequality and enhance overall welfare across the country.

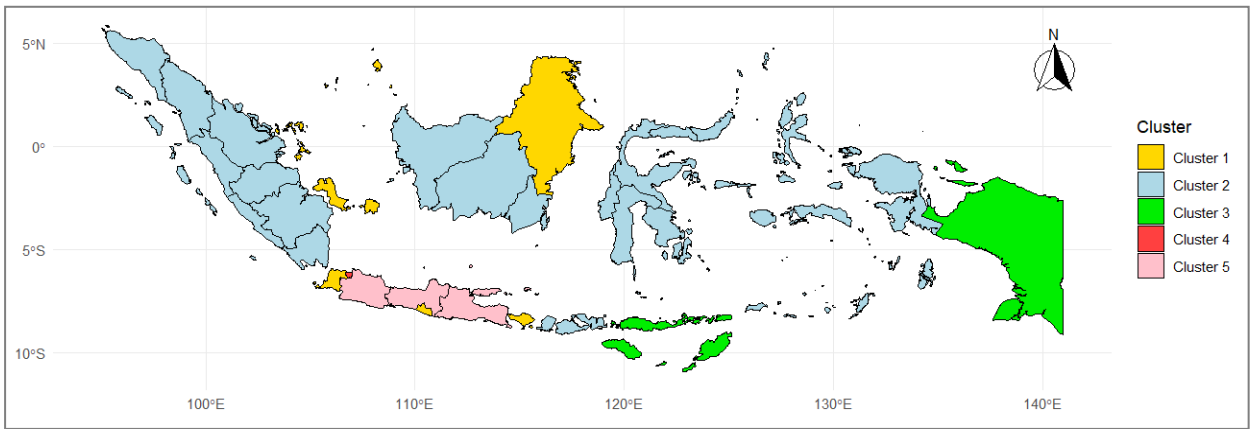



**Figure 6. Clustering of Indonesian Provinces Based on People's Welfare Using GMM**

### 3.4 Determination of the best performing method

Based on the results of determining optimal conditions with the DBSCAN and *Gaussian Mixture Model* (GMM) methods in clustering provinces in Indonesia based on indicators of people's welfare, the GMM method provides more constant results according to the determination of optimal conditions with respect to the Silhouette value, the Davies-Bouldin Index value and the *Calinski-Harabasz Index* value than DBSCAN. Based on these three values, DBSCAN has an optimal number of clusters of two with two different values of *epsilon* and minPts, while GMM has an optimal number of clusters of 5 according to all types of validity. Thus, it can be said that the *Gaussian Mixture Model* has a better performance in clustering provinces in Indonesia based on indicators of people's welfare. This result is consistent with previous research where GMM has a better performance in clustering provinces in Indonesia based on indicators of people's welfare **[12]**.

### 3.5 Discussion

Comparison of optimal cluster results between DBSCAN and *Gaussian Mixture Model* (GMM) can be seen from several points, namely:

1. Based on the Silhouette Index value, Davies-Bouldin Index value and Calinski-Harabasz Index value, DBSCAN has an optimal number of clusters of two clusters with different epsilon values at three validity values considered, while GMM has an optimal number of clusters of five clusters for all validity values considered. This shows that the results of GMM are more constant or stable. The comparison can be seen in **Table 5** and **Table 9**. These results are in accordance with research conducted by Ambasari in 2023 **[12]**.
2. Based on provincial grouping, according to the cluster results with DBSCAN, DKI Jakarta province is in the same group as Kep. Bangka Belitung, NTT, and Papua. This is because DBSCAN has cluster 0 which contains *outlier* data, which is a group of data that has a value far from other data. This result provides an inaccurate explanation, because DKI Jakarta and the other 3 provinces have much different values in each of the people's welfare indicator variables. For example, the percentage of poor people in Bangka Belitung Islands, DKI Jakarta, NTT, and Papua are 4.6%, 4.5%, 20.5%, and 26.5%, respectively. So, it is not right if the four provinces are in one group. Meanwhile, according to the results of clustering with GMM, the province of DKI Jakarta becomes its own cluster, the provinces of NTT and Papua also become their own cluster, while the province of Kep. Bangka Belitung is in one cluster with the provinces of Bali, Banten, East Kalimantan, and DI Yogyakarta. This result is considered more reasonable because the values of each variable in the provinces of NTT and Papua are similar, and DKI Jakarta has a much higher value in each variable, especially in the income and expenditure sections.

Based on the two points above, it can be said that the Gaussian Mixture Model provides better clustering results. The results of the clustering above are in accordance with previous research, where on average, western Indonesia tends to have higher welfare than eastern Indonesia. DKI Jakarta Province, which has the best welfare level, and the Provinces of DI Yogyakarta, Riau Islands, Banten, Bali, and East Kalimantan are in one cluster with good welfare, in accordance with research conducted by Belinda in 2019 **[11]**, **[26]**. Then the provinces of NTT and Papua are in a cluster with high poverty rates, in accordance with research conducted by Saputri in 2023 **[13]**, **[27]**, **[28]**.

The performance of the Gaussian Mixture Model (GMM) is better than DBSCAN in grouping provinces in Indonesia based on indicators of people's welfare, because the GMM results provide optimal cluster results of 5 clusters that are constant or do not change their optimal cluster according to the Silhouette Index, Davies-Bouldin Index, and Calinski-Harabasz Index values, as well as the BIC value **[29]**, **[30]**. This is in accordance with research by Ambarsari in 2023 which states that GMM provides stable results in determining the optimal cluster according to the Silhouette Index, Davies-Bouldin Index, and Calinski-Harabasz Index values **[12]**, **[31]**.

## 4. CONCLUSIONS

Based on the results of this study, several conclusions can be drawn. The comparison of DBSCAN and Gaussian Mixture Model (GMM) in clustering Indonesian provinces based on people’s welfare indicators, using Silhouette Index, Davies-Bouldin Index, and Calinski-Harabasz Index as evaluation metrics, indicates that GMM outperforms DBSCAN in terms of clustering quality. The cluster analysis using GMM resulted in the formation of five distinct clusters, where Cluster 1 represents provinces with medium welfare levels, Cluster 2 consists of provinces with fairly low welfare levels, Cluster 3 includes provinces with low welfare levels, Cluster 4 groups provinces with high welfare levels, and Cluster 5 contains provinces with fairly high welfare levels. These findings highlight the effectiveness of GMM in capturing the variations in welfare conditions across provinces in Indonesia, providing a more detailed and structured classification compared to DBSCAN.

## ACKNOWLEDGMENT

This research was funded by the DIPA grant of Universitas Negeri Yogyakarta in 2023 under the research contract number B/89/UN34.13/PT.01.03/2023. We sincerely appreciate the support provided, which has significantly contributed to the successful completion of this study.

## REFERENCES

[1] F. Basri and H. Munandar, Lanskap Ekonomi Indonesia: KAJIAN DAN RENUNGAN TERHADAP MASALAH-MASALAH STRUKTURAL, TRANSFORMASI BARU, DAN PROSPEK PEREKONOMIAN INDONESIA. Jakarta: Kencana, 2009.

[2] Badan Pusat Statistik (BPS), INDIKATOR KESEJAHTERAAN RAKYAT 2020. Jakarta: Badan Pusat Statistik Indonesia, 2020.

[3] Badan Pusat Statistik (BPS), Indikator Kesejahteraan Rakyat 2023: HUBUNGAN FAKTOR SOSIAL DAN DEMOGRAFI DENGAN PEKERJA LANSIA DI INDONESIA. Jakarta: Badan Pusat Statistik Indonesia, 2023.

[4] E. Setiawan, M. A. Suprayogi, and A. Kurnia, "A COMPARISON OF LOGISTIC REGRESSION, MIXED LOGISTIC REGRESSION, AND GEOGRAPHICALLY WEIGHTED LOGISTIC REGRESSION ON PUBLIC HEALTH DEVELOPMENT IN JAVA," BAREKENG J. Ilmu Mat. Dan Terap., vol. 19, no. 1, pp. 129–140, Jan. 2025, doi: https://doi.org/10.30598/barekengvol19iss1pp129-140.

[5] O. Rahmawati and A. Fauzan, "PROVINCIAL CLUSTERING BASED ON EDUCATION INDICATORS: K-MEDOIDS APPLICATION AND K-MEDOIDS OUTLIER HANDLING," BAREKENG J. Ilmu Mat. Dan Terap., vol. 18, no. 2, pp. 1167–1178, May 2024, doi: https://doi.org/10.30598/barekengvol18iss2pp1167-1178.

[6] P. A. Puspitasari, D. Y. Faidah, and T. Hendrawati, "GROUPING REGENCIES/CITIES IN WEST JAVA PROVINCE BASED ON PEOPLE'S WELFARE INDICATORS USING BIPLOT AND CLUSTERING," BAREKENG J. Ilmu Mat. Dan Terap., vol. 18, no. 3, pp. 1839–1852, Jul. 2024, doi: https://doi.org/10.30598/barekengvol18iss3pp1839-1852.

[7] M. Musa and S. I. Fallo, "HIERARCHICAL CLUSTER ANALYSIS ON PEOPLE'S WELFARE IN SOUTHEAST SULAWESI PROVINCE," BAREKENG J. Ilmu Mat. Dan Terap., vol. 17, no. 2, pp. 1163–1172, Jun. 2023, doi: https://doi.org/10.30598/barekengvol17iss2pp1163-1172.

[8] R. M. Prakash, K. Bhuvaneshwari, M. Divya, K. J. Sri, and A. S. Begum, "SEGMENTATION OF THERMAL INFRARED BREAST IMAGES USING K-MEANS, FCM AND EM ALGORITHMS FOR BREAST CANCER DETECTION," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore: IEEE, Mar. 2017, pp. 1–4. doi: https://doi.org/10.1109/ICIIECS.2017.8276142.

[9] M. B. Johra, "SOFT CLUSTERING DENGAN ALGORITMA FUZZY K-MEANS (STUDI KASUS : PENGELOMPOKAN DESA DI KOTA TIDORE KEPULAUAN)," BAREKENG J. Ilmu Mat. Dan Terap., vol. 15, no. 2, pp. 385–392, Jun. 2021, doi: https://doi.org/10.30598/barekengvol15iss2pp385-392.

[10] Z. Dong, W. Jiang, M. Sun, and Y. Zhang, "SOFT SENSING OF NOX EMISSIONS FROM THERMAL POWER UNITS BASED ON ADAPTIVE GMM TWO-STEP CLUSTERING ALGORITHM AND ENSEMBLE LEARNING," IEEE Trans. Instrum. Meas., vol. 72, pp. 1–19, 2023, doi: https://doi.org/10.1109/TIM.2023.3279913.

[11] N. S. Belinda, I. R. Hg, and H. Yozza, "PENERAPAN ANALISIS CLUSTER ENSEMBLE DENGAN METODE ROCK UNTUK MENGELOMPOKKAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR KESEJAHTERAAN RAKYAT," J. Mat. UNAND, vol. 8, no. 2, p. 108, Jul. 2019, doi: https://doi.org/10.25077/jmu.8.2.108-119.2019.

[12] E. W. Ambarsari, N. Dwitiyanti, N. Selvia, W. N. Cholifah, and P. D. Mardika, "COMPARISON APPROACHES OF THE FUZZY C-MEANS AND GAUSSIAN MIXTURE MODEL IN CLUSTERING THE WELFARE OF THE INDONESIAN PEOPLE," KnE Soc. Sci., May 2023, doi: 10.18502/kss.v8i9.13315.

[13] F. W. Saputri and D. B. Arianto, "PERBANDINGAN PERFORMA ALGORITMA K-MEANS, K-MEDOIDS, DAN DBSCAN DALAM PENGGEROMBOLAN PROVINSI DI INDONESIA BERDASARKAN INDIKATOR KESEJAHTERAAN MASYARAKAT," J. Teknol. Inf. J. Keilmuan Dan Apl. Bid. Tek. Inform., vol. 7, no. 2, pp. 138–151, Aug. 2023, doi: https://doi.org/10.47111/jti.v7i2.9558.

[14] N. Dwitiyanti, N. Selvia, and F. R. Andrari, "PENERAPAN FUZZY C-MEANS CLUSTER DALAM PENGELOMPOKKAN PROVINSI INDONESIA MENURUT INDIKATOR KESEJAHTERAAN RAKYAT," Fakt. Exacta, vol. 12, no. 3, p. 201, Nov. 2019, doi: https://doi.org/10.30998/faktorexacta.v12i3.4526.

[15] N. Dwitiyanti, S. Wulandari, and N. Selvia, "IMPLEMENTASI GRAPH CLUSTERING ALGORITHM MODIFICATION MAXIMUM STANDARD DEVIATION REDUCTION (MMSDR) DALAM CLUSTERING PROVINSI DI INDONESIA MENURUT INDIKATOR KESEJAHTERAAN RAKYAT," Fakt. Exacta, vol. 13, no. 2, p. 73, Aug. 2020, doi: https://doi.org/10.30998/faktorexacta.v13i2.5863.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A DENSITY-BASED ALGORITHM FOR DISCOVERING CLUSTERS IN LARGE SPATIAL DATABASES WITH NOISE," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), Aug. 1996, pp. 226–231.

[17] A. Kassambara, PRACTICAL GUIDE TO CLUSTER ANALYSIS IN R: UNSUPERVISED MACHINE LEARNING, 1st ed. United States: STHDA, 2017.

[18] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN REVISITED, REVISITED: WHY AND HOW YOU SHOULD (STILL) USE DBSCAN," ACM Trans. Database Syst., vol. 42, no. 3, pp. 1–21, Sep. 2017, doi: https://doi.org/10.1145/3068335.

[19] C. M. Bishop, PATTERN RECOGNITION AND MACHINE LEARNING. New York: Springer, 2006.

[20] K. E. Setiawan and A. Kurniawan, "PENGELOMPOKAN RUMAH SAKIT DI JAKARTA MENGGUNAKAN MODEL DBSCAN, GAUSSIAN MIXTURE, DAN HIERARCHICAL CLUSTERING," J. Inform. Terpadu, vol. 9, no. 2, pp. 149–156, Sep. 2023, doi: https://doi.org/10.54914/jit.v9i2.995.
[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "MAXIMUM LIKELIHOOD FROM INCOMPLETE DATA VIA THE EM ALGORITHM," J. R. Stat. Soc. Ser. B Stat. Methodol., vol. 39, no. 1, pp. 1–22, Sep. 1977, doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
[22] G. Schwarz, "ESTIMATING THE DIMENSION OF A MODEL," Ann. Stat., vol. 6, no. 2, Mar. 1978, doi: https://doi.org/10.1214/aos/1176344136.
[23] P. Guenther, M. Guenther, C. M. Ringle, G. Zaefarian, and S. Cartwright, "IMPROVING PLS-SEM USE FOR BUSINESS MARKETING RESEARCH," Ind. Mark. Manag., vol. 111, pp. 127–142, May 2023, doi: https://doi.org/10.1016/j.indmarman.2023.03.010.
[24] C. F. Dormann et al., "COLLINEARITY: A REVIEW OF METHODS TO DEAL WITH IT AND A SIMULATION STUDY EVALUATING THEIR PERFORMANCE," Ecography, vol. 36, no. 1, pp. 27–46, Jan. 2013, doi: https://doi.org/10.1111/j.1600-0587.2012.07348.x.
[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, AN INTRODUCTION TO STATISTICAL LEARNING. New York: Springer, 2013.
[26] M. F. F. Mardianto et al., "GROUPING OF PROVINCES IN INDONESIA BASED ON COMMUNITY WELFARE LEVEL INDICATORS USING HIERARCHICAL CLUSTER ANALYSIS," presented at the 4TH INTERNATIONAL SCIENTIFIC CONFERENCE OF ALKAFEEL UNIVERSITY (ISCKU 2022), Najaf, Iraq, 2023, p. 080015. doi: https://doi.org/10.1063/5.0181024.
[27] P. I. Kontoro, J. Junaidi, N. F. Gamayanti, and A. S. N. Apusing, "CLUSTERING INDONESIAN PROVINCES BASED ON POVERTY LEVELS UTILIZING THE AVERAGE LINKAGE METHOD WITH PRINCIPAL COMPONENT ANALYSIS," in Proceedings of the 5th International Seminar on Science and Technology (ISST 2023), vol. 10, Y. Yuyun, M. Rasyiid, M. S. Zubair, and E. Sesa, Eds., in Advances in Physics Research, vol. 10. , Dordrecht: Atlantis Press International BV, 2024, pp. 78–86. doi: https://doi.org/10.2991/978-94-6463-520-1_13.
[28] K. Vanessa, I. A. Iswanto, K. Wijaya, and M. F. Hidayat, "COMPARING K-MEANS AND DBSCAN ALGORITHMS FOR CLUSTERING POVERTY LEVELS IN PAPUA ISLANDS," in 2024 9th International Conference on Information Technology and Digital Applications (ICITDA), Nilai, Negeri Sembilan, Malaysia: IEEE, Nov. 2024, pp. 1–6. doi: 1 https://doi.org/10.1109/ICITDA64560.2024.10810077.
[29] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "CLUSTER VALIDITY INDICES FOR AUTOMATIC CLUSTERING: A COMPREHENSIVE REVIEW," Heliyon, vol. 11, no. 2, p. e41953, Jan. 2025, doi: https://doi.org/10.1016/j.heliyon.2025.e41953.
[30] I. Ioannou, C. Christophorou, P. Nagaradjane, and V. Vassiliou, "PERFORMANCE EVALUATION OF MACHINE LEARNING CLUSTER METRICS FOR MOBILE NETWORK AUGMENTATION," in 2024 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India: IEEE, Mar. 2024, pp. 1–7. doi: https://doi.org/10.1109/WiSPNET61464.2024.10532825.
[31] D. Y. Faidah, D. Destin, F. A. Anggina, and M. I. Caesar, "ASSESSING THE PERFORMANCE OF K-MEANS AND DBSCAN CLUSTERING METHODS IN TUBERCULOSIS MAPPING," Commun. Math. Biol. Neurosci., 2025, doi: 10.28919/cmbn/9039.