Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index
DOI: https://doi.org/10.24127/sciencestatistics.v4i1.XXXX
Vol. 5, No. 2, 2026, 77-89
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Gaussian Mixture Models for Human Development-Based Regional Clustering of East Java

Article Info

Article History:
Received: 04-06-2026
Revised: 25-06-2026
Accepted: 30-06-2026
Available online: 03-07-2026

Keywords:
clustering;
gaussian mixture model;
human development

ABSTRACT

Regional disparities in human development require an analytical approach that can identify latent patterns in development achievements. This study applies the Gaussian Mixture Model (GMM) to cluster regencies and cities in East Java based on the Human Development Index (HDI). GMM was chosen because it offers a probabilistic and distribution-based clustering framework, assigns regions using posterior membership probabilities, and provides interpretable parameters such as mixing proportions, component means, and variances. Univariate GMMs with two, three, and four components were fitted to HDI data from 38 regencies/cities in East Java at two time points, 2015 and 2025, which represent a ten-year interval. Model selection was conducted using internal cluster-validity measures, namely the Silhouette Coefficient and the Davies–Bouldin Index. The results show that the two-component GMM is selected as the best model for both years. The selected model produces the same membership structure in 2015 and 2025, forming a larger cluster of 31 regencies/cities with relatively lower and more variable HDI values and a smaller cluster of 7 regencies/cities with higher and more homogeneous HDI values. Over the ten-year interval, HDI increased in all regions, while the two-cluster structure remained evident. These findings can support regional development planning, policy evaluation, and the formulation of more targeted development strategies across regencies and cities.

INTRODUCTION

The Human Development Index (HDI) is an important indicator for assessing human development achievements in a region (Ramadhan & Wijayanto, 2023). It reflects not only economic progress but also the quality of life through the dimensions of health, education, and decent living standards (Lind, 2019; Pratama & Ciptawaty, 2022). In regional development, HDI is a strategic measure because it can reveal differences in development achievements across regencies/cities and support more targeted policy formulation.

East Java consists of many regencies/cities with diverse regional characteristics. Differences in social, economic, educational, health, and public service conditions may lead

to unequal HDI achievements across regions (Nurhasanah, 2023; Qolbi et al., 2024). Some regencies/cities may have relatively high HDI values, while others remain at lower levels. This condition indicates that HDI analysis should not rely only on provincial averages or administrative rankings, but should also identify the grouping structure of regions based on similarities in human development achievements.

Grouping regencies/cities based on HDI is important because it provides a more systematic picture of human development disparities across regions (Comim & Martori, 2026). Regions with relatively similar HDI characteristics can be placed in the same group, allowing clearer interpretation of regional development patterns. The resulting groups may also help identify areas with relatively low, medium, or high HDI, providing an initial basis for making more accurate development plans.

The choice of grouping method should be aligned with the characteristics of the data and the research objective. Various clustering methods can be used for regional grouping, including K-means (Fahmiyah & Ningrum, 2023; Unggul et al., 2023), K-medoids (Robiati et al., 2024; Salsabila et al., 2025), hierarchical clustering (Astuti & Rezania, 2022; Khikmah & Sofro, 2021), and Fuzzy C-Means (Latifah et al., 2022; Rahmanesta et al., 2024). However, these methods differ in their grouping principles and are not always designed to model the underlying distribution of the data. K-means and hierarchical clustering generally assign each region to a single group based on distance or similarity, while Fuzzy C-Means allows partial membership but still defines membership mainly through distance to cluster centers.

For this reason, this study prioritizes a distribution-based grouping approach using the Gaussian Mixture Model (GMM). GMM is a probabilistic model that represents the data distribution as a mixture of several Gaussian components (Pravitasari et al., 2020). Each component can be interpreted as a latent group consisting of regencies/cities with relatively homogeneous HDI characteristics. Thus, GMM does not merely divide the data into groups; it also models the distributional structure of HDI more formally. This approach is suitable when HDI variation across regions may arise from several latent subpopulations with different levels of human development.

A key advantage of GMM over conventional clustering methods is its probabilistic nature. GMM produces posterior probabilities, which indicate the probability that a regency/city belongs to each component or group. Therefore, classification is not only expressed as a group label but is also accompanied by a measure of membership certainty. This is important because differences in human development across regions are not always sharply separated. Some regencies/cities may have HDI values close to the boundary between groups, so additional information is needed to assess how strongly a region belongs to a particular group.

In addition to membership probabilities, GMM provides distributional parameters for each group, including mixing proportion, mean, and variance. These parameters offer a stronger basis for interpreting the characteristics of each group. For example, a group with a lower mean HDI can be interpreted as a group of regions with relatively lower human development achievement, while a group with a higher mean HDI can be interpreted as a more developed group. In this sense, GMM provides not only a grouping result but also a statistical explanation of the HDI distribution within each group.

An important issue in applying finite mixture models, including the GMMs, is determining the most appropriate number of components (Manole & Khalili, 2021). The number of

groups should not be chosen subjectively, because too few components may oversimplify the data structure, while too many components may produce an overly complex model. Therefore, this study evaluates several component configurations, namely $K = 2$, $K = 3$, and $K = 4$. The best model is selected using internal cluster-validity measures, namely the Silhouette Coefficient (SC) and the Davies–Bouldin Index (DBI). These measures assess the quality of the resulting clusters in terms of within-cluster compactness and between-cluster separation, allowing the selected number of groups to have a more objective basis.

This study focuses on the 2025 HDI data for regencies/cities in East Java as the main analysis. The 2015 HDI data are used as an additional comparison to observe changes in the grouping structure over a ten-year period. This comparison provides not only a current picture of regional grouping but also a historical context regarding how HDI group patterns may change over time.

Based on this background, this study aims to apply the Gaussian Mixture Model to group regencies/cities in East Java based on the Human Development Index. Specifically, this study aims to determine the best number of components using internal cluster-validity measures, interpret the model parameters and group characteristics, and compare the 2025 HDI grouping results with those of 2015 as an additional reference. This approach is expected to provide a methodological contribution to regional development analysis through distribution-based grouping and to offer a more informative understanding of human development heterogeneity across regencies/cities in East Java.

METHODS

Data and Variable

The data used in this study consist of HDI data from 38 regencies/cities in East Java Province, obtained from the official public website of BPS of East Java Province. The main focus of the analysis is the 2025 HDI data, while the 2015 HDI data are used as an additional comparison to examine changes in the grouping structure over a ten-year period. HDI itself is a continuous variable that represents human development achievements in a region. In the context of this study, HDI values are used as the basis for forming groups of regencies/cities with relatively similar human development characteristics. In general, the HDI observations can be expressed as $x_1, x_2, \dots, x_{38}$, where x_i denotes the HDI value of the i-th regency/city, $i = 1,2, \dots ,38$.

Gaussian Mixture Model

The Gaussian Mixture Model assumes that the data distribution is generated from a mixture of several Gaussian components (Chassagnol et al., 2023; Ren et al., 2025). If there are K components, the mixture density function for observation x_i is expressed in Equation (1) with $\boldsymbol{\Theta}$ denoting the parameter set defined in Equation (2).

$$f(x_i|\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \, \phi\left(x_i|\mu_k, \sigma_k^2\right) \tag{1}$$

$$\boldsymbol{\Theta} = \left\{\pi_k, \mu_k, \sigma_k^2\right\}_{k=1}^{K} \tag{2}$$

In this model, π_k denotes the mixing proportion of the k-th component with $\pi_k > 0$ for all $k = 1,2, \dots, K$ and $\sum_{k=1}^{K} \pi_k = 1$ (You et al., 2023). On the other hand, μ_k and σ_k^2 denote the mean and the variance of the k-th component, respectively. Meanwhile, $\phi\left(x_i|\mu_k, \sigma_k^2\right)$ represents the normal density function evaluated at observation x_i for the k-th component.

Parameter Estimation of GMM

The parameters of the GMM can be estimated by maximizing the likelihood function. For n independent observations, the likelihood and log likelihood function of the GMM are respectively shown in equation (3) and (4):

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \, \phi\left(x_i | \mu_k, \sigma_k^2\right) \right] \tag{3}$$

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \, \phi\left(x_i | \mu_k, \sigma_k^2\right) \right] \tag{4}$$

Because the mixture model includes unobserved component membership, parameter estimation is executed by using the Expectation-Maximization (EM) algorithm (McLachlan et al., 2019). Let z_{ik} be a latent indicator variable defined as in equation (5) (Unggul et al., 2026).

$$z_{ik} = \begin{cases} 1, & \text{if } i\text{-th observation belongs to } k\text{-th component} \\ 0, & \text{if } i\text{-th observation does not belong to } k\text{-th component} \end{cases} \tag{5}$$

The EM algorithm consists of two main steps, namely the E-step and the M-step (Jannah & Saputro, 2022). In the E-step, the posterior probability or responsibility, notated by τ_{ik}, is calculated. This value represents the probability that observation x_i belongs to component k, given the parameter values at iteration r. It is given in equation (6) by applying Bayes' rule.

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \, \phi\left(x_i | \mu_k^{(r)}, \left(\sigma_k^2\right)^{(r)}\right)}{\sum_{s=1}^{K} \pi_s^{(r)} \, \phi\left(x_i | \mu_s^{(r)}, (\sigma_s^2)^{(r)}\right)} \tag{6}$$

In the M-step, the model parameters are updated based on the posterior probabilities obtained from the E-step. The updated mixing proportion is given in equation (7). The updated component mean and variance are shown in equation (8) and (9), respectively.

$$\pi_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)} \tag{7}$$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \, x_i}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{8}$$

$$\left(\sigma_k^2\right)^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \left(x_i - \mu_k^{(r+1)}\right)^2}{\sum_{i=1}^{n} \tau_{ik}^{(r)}} \tag{9}$$

The E-step and M-step are repeated until convergence is achieved (Oduor, 2025). Convergence is reached when the change in the log-likelihood value between iterations becomes very small or when the maximum number of iterations is reached. After the final parameter estimates are obtained, the group membership of each regency/city is determined based on the largest posterior probability. The group allocation for observation i is expressed as in equation (10) . In addition to producing group labels, GMM also provides a measure of classification certainty through $\max_{k \in \{1,2,\ldots,K\}} \tau_{ik}$, which is the maximum posterior

probability. Values close to 1 indicate that this region has a strong probability of belonging to one particular group.

$$\hat{z}_i = \arg \max_{k \in \{1,2,\ldots,K\}} \tau_{ik} \tag{10}$$

Model Evaluation and Selection

Model evaluation is conducted using two internal cluster-validity measures, namely the Silhouette Coefficient (SC) and the Davies–Bouldin Index (DBI). Unlike information-based criteria, these measures assess the quality of the resulting partition directly, based on how compact each cluster is and how well separated the clusters are from one another. The SC is given in Equation (11) (Rousseeuw, 1987). For each observation i, let $a(i)$ be the average distance between i and the other members of its own cluster, and let $b(i)$ be the average distance between i and the members of the nearest neighbouring cluster. The SC is the mean of the individual silhouette values over all observations. It ranges from −1 to 1, and a higher value indicates better-defined clusters. The DBI is given in Equation (12) (Davies & Bouldin, 1979). Let S_k be the average distance of the members of cluster k to its center c_k, and let $d(c_k, c_j)$ be the distance between the centres of clusters k and j. A smaller DBI indicates better-defined clusters.

$$\text{SC} = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{11}$$

$$\text{DBI} = \frac{1}{K} \sum_{k=1}^{K} \max_{j \neq k} \left[\frac{S_k + S_j}{d(c_k, c_j)} \right] \tag{12}$$

Stages of Analysis

The analysis was carried out through the following stages:

1. Data preparation
   The HDI data of 38 regencies/cities in East Java for 2025 and 2015 were prepared.
2. Descriptive exploration
   The distribution of HDI was explored descriptively to observe the initial distributional pattern of the data.
3. Model estimation
   GMMs were estimated using several configurations, namely $K = 2$, $K = 3$, and $K = 4$.
4. Model evaluation
   Each model was evaluated using SC and DBI.
5. Best model selection
   The best model was selected based on cluster quality together with the interpretability of the resulting clusters and the consistency of the structure across the two years.
6. Parameter interpretation
   The parameters of the selected model were interpreted, including the mixing proportions, component means, and component variances.
7. Group allocation
   Regencies/cities were assigned to groups based on the largest posterior probability.
8. Comparison of grouping results
   The 2025 HDI grouping results were analyzed and compared with the 2015 grouping results as an additional reference.
9. Conclusion
   The final conclusions were drawn based on the selected model and the resulting HDI group characteristics.

RESULT AND DISCUSSIONS

This section describes the results of this study. First, Table 1 summarizes the descriptive statistics of HDI across the 38 regencies and cities in each year. In 2025 the mean HDI is 76.0645 with a standard deviation of 4.6983, ranging from 67.23 to 85.65. The median is 75.32, and the first and third quartiles are 72.45 and 79.265. In 2015 the mean is 69.1132 with a standard deviation of 5.3338, ranging from 58.18 to 80.05. The median is 68.62, with quartiles at 64.95 and 72.31. Between the two years the average HDI is higher by about 6.95 points, and the standard deviation is smaller in 2025, so the units are slightly less dispersed.

Table 1. Descriptive statistics for HDI across regions in East Java.

Year	Mean	SD	Min	Max	Q1	Median	Q3
2025	76.0645	4.6983	67.23	85.65	72.4500	75.3200	79.2650
2015	69.1132	5.3338	58.18	80.05	64.9500	68.6150	72.3125

For each year, univariate Gaussian mixture models were fitted with K = 2, 3, and 4 components on all n = 38 units. The number of free parameters is $q = 3K - 1$, giving q = 5, 8, and 11. The models were compared using the SC and the DBI, reported in Table 2.

Table 2. Model performance across different numbers of components.

Year	K	n	q	SC	DBI
2025	2	38	5	0.5946	0.4132
	3	38	8	0.5189	0.5438
	4	38	11	0.5051	0.4238
2015	2	38	5	0.5645	0.4640
	3	38	8	0.5604	0.4638
	4	38	11	0.5723	0.4275

In 2025, the two-component model has the highest Silhouette Coefficient (0.5946) and the lowest Davies–Bouldin Index (0.4132) among the three configurations, so both measures point to K=2 as the best-defined partition. In 2015, the two measures change only slightly across the configurations (SC between 0.560 and 0.572; DBI between 0.428 and 0.464), so no number of components is clearly better than the others. Given this near-equivalence in 2015 and the aim of comparing the same cluster structure across the two years, the two-component model is also retained for 2015. The two-component model ($K = 2$) is therefore selected as the best model for both 2025 and 2015.

Table 3 reports the estimated parameters of the selected two-component models, which are the mixing proportion ($\hat{\pi}_k$), mean ($\hat{\mu}_k$), and variance ($\hat{\sigma}_K^2$) of each component. In both years the data split into one large component at lower HDI and one small component at higher HDI. In 2025, Component 1 holds about 82% of the units with a mean of 74.36 and a variance of 10.53, while Component 2 holds the remaining 18% with a higher mean of 83.66 and a smaller variance of 2.91. The 2015 estimates show the same structure: Component 1 covers 84% of the units (mean 67.49, variance 16.38) and Component 2 the other 16% (mean 77.74, variance 4.10). The two years differ mainly in level. Both component means are higher in 2025 (74.36 vs. 67.49 and 83.66 vs. 77.74), matching the overall rise in HDI, and both variances are smaller, so units within each component are more alike than before. The mixing proportions barely change, and the gap between the two component means stays close to 10 points (9.30 in 2025, 10.25 in 2015).

Table 3. Parameter estimates of the best models.

Year	Component	$\hat{\pi}_k$	$\hat{\mu}_k$	$\hat{\sigma}_k^2$
2025	1	0.8167	74.3601	10.5254
	2	0.1833	83.6601	2.9078
2015	1	0.8415	67.4898	16.3810
	2	0.1585	77.7421	4.0980

Substituting the estimates in Table 3 into the two-component Gaussian mixture gives the fitted density for each year. The fitted density for 2025 is given in Equation (13) and the fitted density for 2015 in Equation (14). In each equation the first bracketed term is Component 1 and the second is Component 2, each weighted by its mixing proportion so that the two add up to the overall density. In both equations the value substituted for each component is its variance σ_k^2, so no standard-deviation conversion is applied.

$$f_{2025}(x_i|\widehat{\boldsymbol{\Theta}}) = 0.8167\left[\frac{1}{\sqrt{2\pi(10.5254)}}\exp\left\{-\frac{(x_i-74.3601)^2}{2(10.5254)}\right\}\right] + 0.1833\left[\frac{1}{\sqrt{2\pi(2.9078)}}\exp\left\{-\frac{(x_i-83.6601)^2}{2(2.9078)}\right\}\right] \quad (13)$$

$$f_{2015}(x_i|\widehat{\boldsymbol{\Theta}}) = 0.8415\left[\frac{1}{\sqrt{2\pi(16.3810)}}\exp\left\{-\frac{(x_i-67.4898)^2}{2(16.3810)}\right\}\right] + 0.1585\left[\frac{1}{\sqrt{2\pi(4.098)}}\exp\left\{-\frac{(x_i-77.7421)^2}{2(4.098)}\right\}\right] \quad (14)$$

Furthermore, Figure 1 shows the fitted mixture densities, with 2025 in the top panel and 2015 in the bottom panel. In each panel the solid curve is the overall mixture density, the two shaded curves are the contributions of Component 1 and Component 2, and the vertical lines mark the component means. Component 1 is shown in red and Component 2 in blue.

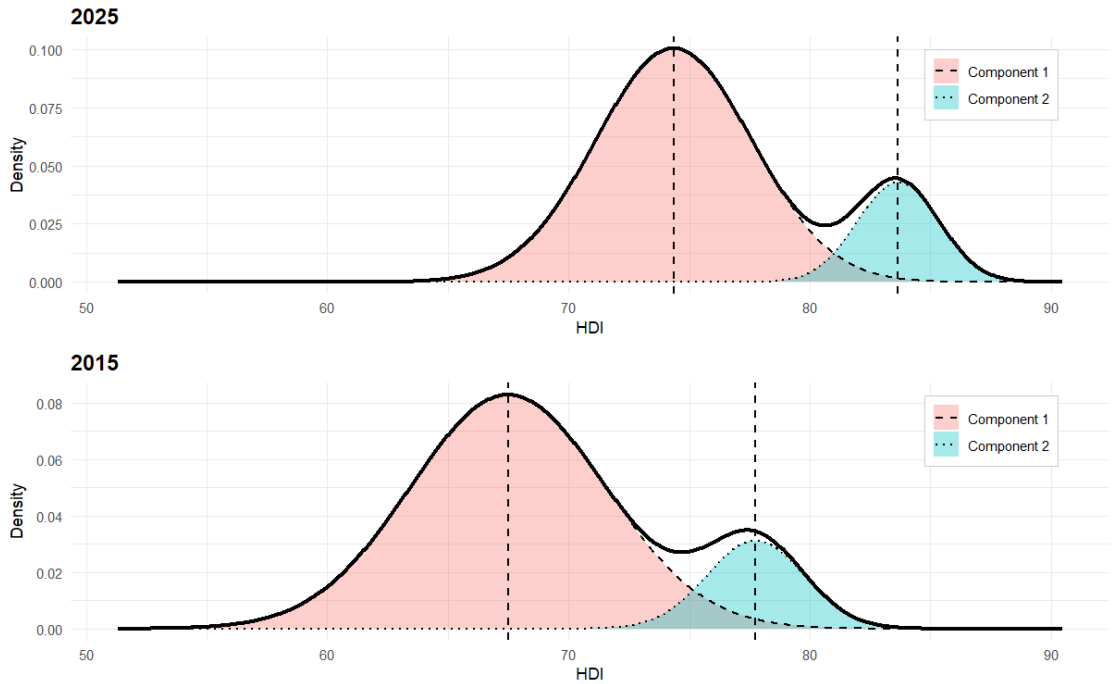

Figure 1. Fitted two-component Gaussian mixture densities for East Java HDI.

Both years show the same two-part shape. A large component covers the lower and middle range of HDI, while a smaller component sits at the upper end. Relative to 2015, the 2025 density is shifted toward higher HDI, since both component means are larger. The upper component's peak is higher in 2025: its larger mixing proportion (0.18 versus 0.16) and smaller variance (2.91 versus 4.10) together raise the modal height of the weighted component from about 0.031 in 2015 to about 0.043 in 2025, an increase of roughly 39%. The figure is therefore consistent with the parameter estimates, showing an overall increase in HDI between the two years alongside a persistent split between a large lower-HDI component and a small higher-HDI component.

Each component forms one cluster, with each unit assigned to the cluster it most likely belongs to. Figure 2 maps these assignments for all 38 regencies and cities, with 2025 in the top panel and 2015 in the bottom panel, Cluster 1 shown in red and Cluster 2 in blue. The intensity of each colour reflects the maximum posterior probability, so darker shading indicates more certain assignment. The spatial pattern is identical across the two years: every one of the 38 units keeps the same cluster label in both 2025 and 2015, as confirmed by the unchanged membership reported in Table 4, so no region changes cluster between the two time points. Cluster 2 consists of the same seven units, namely Sidoarjo, Surabaya City, Mojokerto City, Kediri City, Blitar City, Malang City, and Madiun City, while the remaining 31 form Cluster 1. Geographically, Cluster 2 units are concentrated in the Surabaya metropolitan core. All regencies, all four Madura island units (Bangkalan, Sampang, Pamekasan, and Sumenep), and all eastern regencies fall into Cluster 1.

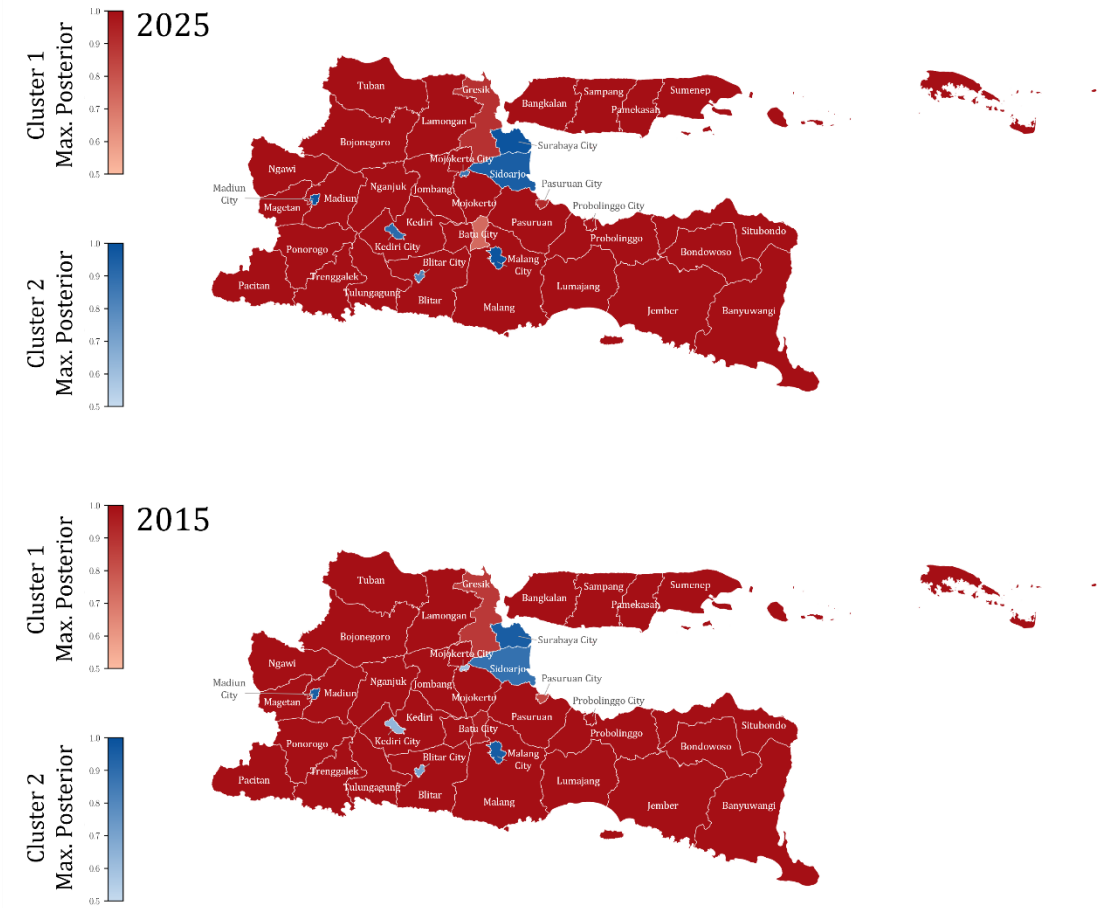

Figure 2. Visualization of cluster allocation and maximum posterior based on the selected models.

The shading reveals a consistent pattern in both years. Most Cluster 1 units are assigned with near-certainty, appearing as solid deep red, with only Gresik, Pasuruan City, and Batu City visibly lighter. Within Cluster 2, the three lower-HDI cities (Mojokerto City, Kediri City, and Blitar City) carry the weakest assignments and appear as lighter blue, while Malang City, Madiun City, Sidoarjo, and Surabaya City are near-solid. Comparing the two panels, the 2025 assignments are more confident: the lighter blues of Mojokerto City, Kediri City, and Blitar City in 2015 (posterior 0.60, 0.63, and 0.71, respectively) darken noticeably in 2025 (0.87, 0.91, and 0.82). The one unit moving in the opposite direction is Batu City, which is clearly the palest red in 2025 (posterior 0.74) compared to a much more certain assignment in 2015 (0.97), consistent with its HDI rising from 72.62 to 80.35, bringing it closer to the Cluster 2. Despite these cases, the two-cluster partition remains stable across both years, with no unit changing its assignment. Table 4 presents the complete cluster membership.

Table 4. Cluster membership for both years.

Cluster	Regencies/Cities
1	Bangkalan, Banyuwangi, Batu City, Blitar, Bojonegoro, Bondowoso, Gresik, Jember, Jombang, Kediri, Lamongan, Lumajang, Madiun, Magetan, Malang, Mojokerto, Nganjuk, Ngawi, Pacitan, Pamekasan, Pasuruan, Pasuruan City, Ponorogo, Probolinggo, Probolinggo City, Sampang, Situbondo, Sumenep, Trenggalek, Tuban, Tulungagung
2	Blitar City, Kediri City, Madiun City, Malang City, Mojokerto City, Sidoarjo, Surabaya City

Cluster 2 consists of six cities and one regency (Sidoarjo), while Cluster 1 covers the remaining 28 regencies and three cities, namely Batu City, Pasuruan City, and Probolinggo City. The partition broadly follows the administrative distinction between cities and regencies, though not entirely. Sidoarjo, an industrial regency within the Surabaya metropolitan area, is assigned to Cluster 2. Its inclusion reflects Sidoarjo's socioeconomic profile: although administratively a regency, it is a densely populated, industrialised area within the Surabaya metropolitan region (Gerbangkertosusila), with extensive manufacturing activity, well-developed infrastructure, and ready access to the education and health facilities of the metropolitan core. These conditions raise its human development achievements to a level comparable with the cities, which explains its placement in the high-HDI cluster. Table 5 reports some descriptive statistics of the HDI distribution within each cluster for both years.

Table 5. Some characteristics for each cluster.

Cluster	Year	n	Mean	SD	Min	Max	Q1	Median	Q3
1	2025	31	74.3126	3.2015	67.23	80.35	71.80	74.43	76.23
	2015	31	67.1826	3.7872	58.18	73.78	64.24	68.08	69.87
2	2025	7	83.8229	1.5732	82.03	85.65	82.53	83.35	85.34
	2015	7	77.6629	1.9817	75.54	80.05	75.84	77.43	79.48

In 2025, Cluster 1 has a mean HDI of 74.31 compared to 83.82 for Cluster 2, a difference of approximately 9.5 points. In 2015, this gap was slightly wider at approximately 10.5 points (67.18 versus 77.66). Both clusters improved over the decade, with the gap narrowing somewhat due to a slightly larger gain in Cluster 1. Regarding within-cluster variation, Cluster 1 is more dispersed than Cluster 2 in both years, with standard deviations of 3.20 in 2025 and 3.79 in 2015, compared to 1.57 and 1.98 for Cluster 2, indicating that the units

within Cluster 2 are more homogeneous. Figure 3 shows the HDI improvement of each unit from 2015 to 2025, with the bold dashed lines representing the cluster means.

As seen in Figure 3, all 38 units show a positive HDI change from 2015 to 2025, and the two clusters remain clearly separated throughout the period. No unit crosses from one cluster to the other, which is consistent with the stable membership reported in Table 4. The gap between the two cluster means narrows slightly over the decade, reflecting the modest convergence noted in Table 5, yet the two-tier structure of HDI distribution in East Java remains intact by 2025. Overall, the results indicate that while all regencies and cities in East Java have made progress, the divide between the higher-HDI cluster and the broader lower-HDI cluster has persisted. From a policy perspective, the stability of this two-tier structure suggests that uniform, province-wide measures may not be sufficient to close the development gap. Sustained and targeted support directed at the lower-HDI cluster, particularly in widening access to education and health services, would be needed to accelerate convergence, since a decade of general progress alone did not move any unit out of its group.

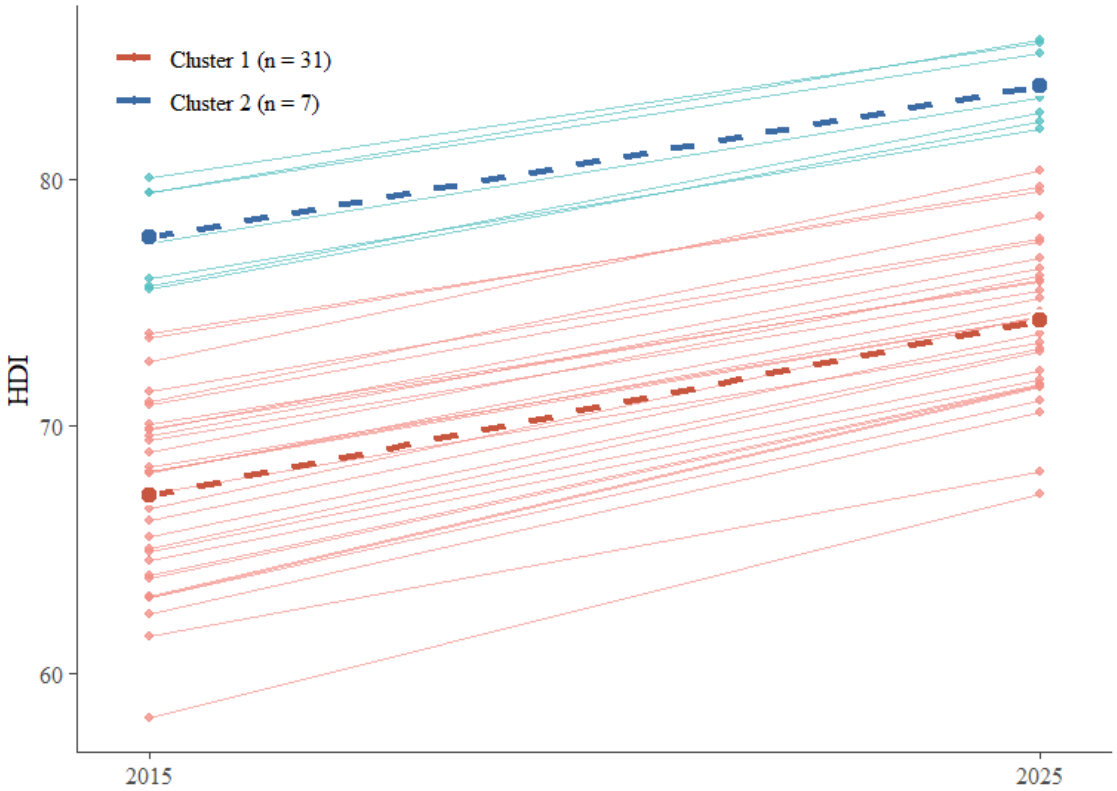

Figure 3. HDI improvement across East Java regencies and cities from 2015 to 2025, colored by cluster.

CONCLUSIONS AND SUGGESTIONS

The two-component GMM ($K = 2$) is selected as the best model for both 2025 and 2015 data, based on internal cluster-validity measures, namely the Silhouette Coefficient and the Davies–Bouldin Index. The model identifies two distinct clusters across all 38 regencies and cities in East Java, with the same membership in both years. Cluster 1 consists of 31 regencies/cities at a lower HDI values with greater variability, while Cluster 2 consists of 7 regencies/cities at a higher HDI values and is more homogeneous. Although all units show

improvement from 2015 to 2025, the division between the two clusters remains, with the gap narrowing only slightly over the decade.

The emergence of only two latent groups has a clear practical interpretation. Human development in East Java is concentrated in a small set of highly urbanised cities, together with the industrial regency of Sidoarjo in the Surabaya metropolitan area, where access to education, health services, and economic opportunity is strongest. The remaining regencies form a broad and fairly homogeneous group at more moderate development levels. This center–periphery pattern, essentially a divide between urbanised centres and the wider regencies, produces two well-separated latent groups rather than a finer gradient, which is consistent with the cluster-validity measures not supporting a third or fourth component.

Future research could extend the analysis to more time points to better capture the trend of HDI development over time. Incorporating spatially-aware models or additional socioeconomic variables may also provide a richer understanding of the factors driving the persistent divide between the two clusters.

Author Contributions

Didik Bani Unggul: Conceptualization, data curation, methodology, formal analysis, writing-original draft, software, validation, visualization. Muhammad Rusli Baharuddin: Conceptualization, writing-review & editing, validation. Miftah Fahira: Conceptualization, methodology, writing-review & editing, validation. Muhammad Zulfadhli: Conceptualization, writing-review & editing, validation.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

-

Declarations

The authors declare that they have no conflicts of interest.

REFERENCE

Astuti, C. C., & Rezania, V. (2022). Cluster Analysis for Grouping Districts in Sidoarjo Regency Based on Education Indicators. *KnE Social Sciences*, 311–317. https://doi.org/10.18502/kss.v7i10.11233

Chassagnol, B., Bichat, A., Boudjeniba, C., Wuillemin, P.-H., Guedj, M., Gohel, D., Nuel, G., & Becht, E. (2023). Gaussian Mixture Models in R. *The R Journal*, *15*(2), 56–76. https://doi.org/10.32614/RJ-2023-043

Comim, F., & Martori, F. (2026). Mapping HDI's Geostructures of Inequality: A Poset Analysis. *Social Indicators Research*, *183*(2), 14. https://doi.org/10.1007/s11205-026-03860-6

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Fahmiyah, I., & Ningrum, R. A. (2023). Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories. *Journal of Advanced Technology and Multidiscipline*, *2*(1), 27–33. https://doi.org/10.20473/jatm.v2i1.45070

Jannah, W., & Saputro, D. R. S. (2022). *Parameter estimation of Gaussian mixture models (GMM) with expectation maximization (EM) algorithm*. 040002. https://doi.org/10.1063/5.0117119

Khikmah, K. N., & Sofro, A. (2021). Clustering Regency and City in East Java Based on Population Density and Cumulative Confirmed COVID-19 Cases. *ComTech: Computer, Mathematics and Engineering Applications*, *12*(2), 111–121. https://doi.org/10.21512/comtech.v12i2.6891

Latifah, U. W., Surono, S.-, & Suparman, S. (2022). K-means and fuzzy c-means algorithm comparison on regency/city grouping in Central Java Province. *Desimal: Jurnal Matematika*, *5*(2), 155–168. https://doi.org/10.24042/djm.v5i2.12204

Lind, N. (2019). A Development of the Human Development Index. *Social Indicators Research*, *146*(3), 409–423. https://doi.org/10.1007/s11205-019-02133-9

Manole, T., & Khalili, A. (2021). Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *The Annals of Statistics*, *49*(6). https://doi.org/10.1214/21-AOS2072

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Nurhasanah, N. (2023). Classifying regencies and cities on human development index dimensions: Application of K-Means cluster analysis. *Jurnal Sains Sosio Humaniora*, *7*(2), 169–174. https://doi.org/10.22437/jssh.v7i2.15752

Oduor, O. K. (2025). Application of Finite Univariate Gaussian Mixture Models for High-Frequency Financial Data Modelling. *Asian Journal of Probability and Statistics*, *27*(1), 81–95. https://doi.org/10.9734/ajpas/2025/v27i1705

Pratama, A. D., & Ciptawaty, U. (2022). Economic Spatial Patterns and Human Development Index Districts and Cities in Five Southern Sumatera Provinces. *Jurnal Pembangunan Wilayah Dan Kota*, *18*(2), 192–208. https://doi.org/10.14710/pwk.v18i2.36430

Pravitasari, A. A., Iriawan, N., Nurul Solichah, S. A., Irhamah, I., Fithriasari, K., Purnami, S. W., & Ferriastuti, W. (2020). Gaussian Mixture Model for MRI Image Segmentation to Build a Three-Dimensional Image on Brain Tumor Area. *MATEMATIKA*, 217–234. https://doi.org/10.11113/matematika.v36.n3.1222

Qolbi, N. L., Indriyana, M., & Wulandari, S. P. (2024). Determining Factors of Human Development Index and Quality of Life by Regency/City in East Java Province in 2023 Using Factor Analysis Method. *Indonesian Journal of Interdisciplinary Research in Science and Technology (MARCOPOLO)*, *2*(11), 1487–1506. https://doi.org/10.55927/marcopolo.v2i11.12214

Rahmanesta, F., Martha, Z., Vionanda, D., & Zilrahmi, Z. (2024). Implementation of Fuzzy C-Means Algorithm for Clustering Provinces in Indonesia Based on Micro and Small Industry Ratio in Village Areas. *Indonesian Journal of Statistics and Its Applications*, *8*(2), 178–190. https://doi.org/10.29244/ijsa.v8i2p178-190

Ramadhan, R., & Wijayanto, A. W. (2023). Integrating Satellite Imageries and Multiple Geospatial Big Data for Granular Mapping of Spatial Distribution of Human Development Index in East Java, Indonesia. *Proceedings of The International Conference on Data Science and Official Statistics*, *2023*(1), 274–295. https://doi.org/10.34123/icdsos.v2023i1.369

Ren, X., Wang, M., Dai, G., Peng, L., Chen, X., & Song, Z. (2025). Balancing convergence and diversity: Gaussian mixture models in adaptive weight vector strategies for multi-objective algorithms. *Information Sciences*, *700*, 121858. https://doi.org/10.1016/j.ins.2024.121858

Robiati, S., Fitria, D., Vionanda, D., & Sulistiowati, D. (2024). Comparison of K-Means and K-Medoids in Clustering Regency/City in West Sumatra Province Based on Environmental Indicators. *Indonesian Journal of Statistics and Its Applications*, *8*(2), 191–201. https://doi.org/10.29244/ijsa.v8i2p191-201

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salsabila, T., Nurdin, N., & Retno, S. (2025). Comparison of K-Medoids and K-Means Result for Regional Clustering of Capture Fisheries in Aceh Province. *International Journal of Engineering, Science and Information Technology*, *5*(2), 282–289. https://doi.org/10.52088/ijesty.v5i2.829

Unggul, D. B., Ainy, K. N., & Jannah, R. (2023). Profiling the Inequality of School Water, Sanitation, and Hygiene Facilities Among Indonesian Regions Using Cluster Analysis. *JURNAL KESEHATAN LINGKUNGAN*, *15*(1), 27–36. https://doi.org/10.20473/jkl.v15i1.2023.27-36

Unggul, D. B., Iriawan, N., Irhamah, I., Fahira, M., & Nuraini, U. S. (2026). Mixture Models for Biomedical Image Segmentation: A Systematic Review. *IEEE Access*, *14*, 57521–57539. https://doi.org/10.1109/ACCESS.2026.3681967

You, J., Li, Z., & Du, J. (2023). A new iterative initialization of EM algorithm for Gaussian mixture models. *PLOS ONE*, *18*(4), e0284114. https://doi.org/10.1371/journal.pone.0284114