Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Clustering Provinces in Indonesia Based on Economic Indicators Using the K-Means Algorithm

Ilham Ilyasa[1]*, Muhamad Fazri Sugara[2], Abdul Aziz[3] , Rani Irma Handayani[4], Risca Lusiana Pratiwi[5], Euis Wida Nengsih[6]

[1,2,3,4,5,6] *Program Studi Informatika, Fakulta Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia*
15230753@bsi.ac.id[1], 15230818@bsi.ac.id[2], 15230207@bsi.ac.id[3], rani.rih@nusamandiri.ac.id[4], risca.ral@nusamandiri.ac.id[5], euis.ewh@bsi.ac.id[6]

Abstract

This study aims to analyze and classify the level of economic development in provinces in Indonesia using the K-Means algorithm. The data used includes three main indicators, namely Gross Regional Domestic Product (GRDP) per capita, percentage of poor population, and Human Development Index (HDI) in 2024 obtained from the Central Statistics Agency (BPS). The data was processed through normalization and analysis using the Elbow method to determine the optimal number of clusters. The results were evaluated using the Davies–Bouldin Index (DBI) to assess the level of separation and compactness between clusters. The results show that the most effective division consists of three groups representing high, medium, and low levels of development. Provinces such as DKI Jakarta and Riau are included in the high development cluster, Central Java and South Sulawesi are in the medium cluster, while Papua and East Nusa Tenggara are in the low cluster. These results show that machine learning methods, particularly K-Means, are capable of identifying patterns of regional economic inequality and provide a useful basis for the government in formulating more targeted and equitable development policies.

Keywords: *K-Means Clustering; Economic Development; Regional Inequality; Human Development Index; GRDP per Capita*

1. Introduction

Economic disparities between regions are a significant problem in Indonesia's development [1]. Although Indonesia has succeeded in increasing economic growth and reducing poverty, this has not been accompanied by a decrease in income inequality in Indonesia [2]. For example, panel research on 34 provinces shows that gross regional domestic product (GRDP) per capita and foreign investment have a significant effect on interprovincial income inequality [3]. Furthermore, another analysis found that during the period before and after the COVID-19 pandemic, income inequality between regions in Indonesia increased when measured using the Theil index, although the main contribution came from inequality within provinces, not between provinces [4].

To understand the dynamics of this inequality and strengthen efforts toward more inclusive development policies, a data-driven analytical approach is needed to describe the socioeconomic characteristics of each province and group together regions with similar conditions. In the field of data science, machine learning is a branch of artificial intelligence that is widely used to solve various problems [5]. In particular, unsupervised learning methods such as clustering offer an effective approach to data-based analysis.

One algorithm that is often used in the context of regional clustering is the K-Means Clustering algorithm. The K-Means Clustering method is one of the unsupervised learning algorithms used to group data into a number of clusters based on similarity of characteristics [6]. Several recent studies in Indonesia have applied it to group provinces based on socioeconomic indicators such as the Human Development Index (HDI) and poverty indicators. For example, an HDI clustering study using K-Means divided Indonesian provinces into several clusters based on HDI categories and showed a match between the clusters formed from the official HDI categories [7].

In the context of regional development, indicators such as the Human Development Index (HDI) are measures that can be used to determine the quality of human resources in a region. Meanwhile, economic variables such as GRDP per capita and poverty rates are important for describing the economic conditions of a region. By combining social and economic indicators into the clustering process, the results are expected to provide a more comprehensive picture of the uneven economic development patterns between provinces.

This study aims to apply the K-Means algorithm in grouping provinces in Indonesia based on economic indicators, including GRDP per capita, poverty rate, and HDI. The clustering results are expected to identify clusters of provinces with similar development characteristics, thereby providing a clear picture of regional disparities and serving as a reference for policymakers in formulating more equitable and targeted development strategies.

2. Research Method

This study uses a quantitative approach with an unsupervised learning-based data mining method. The dataset consists of three main variables, namely Gross Regional Domestic Product (GRDP) per capita, percentage of poor population, and Human Development Index (HDI) obtained from the official publication of the Central Statistics Agency (BPS) in 2024 and covers 38 provinces in Indonesia.

The research stages began with data collection from BPS open sources, followed by the data preprocessing stage. Before clustering, the data was first cleaned of missing values and normalized using the StandardScaler method so that the scale between variables was uniform and no attributes dominated the clustering process. [8] .

Next, the optimal number of clusters was determined using the Elbow method, with a k value between 1 and 10. The optimal k value was determined at the elbow point, where the decline in inertia began to slow significantly. After the optimal number of clusters was obtained, the K-Means algorithm was implemented to group provinces based on their economic characteristics. This process was carried out using the Python programming language with the pandas, numpy, matplotlib, seaborn, and scikit-learn libraries.

To ensure the quality of the resulting grouping, an internal validation stage was carried out using the Davies–Bouldin Index (DBI) method. This index measures how well each cluster is separated from one another by evaluating parameters such as density and distance between clusters [9]. A low DBI value indicates that the clustering results have a large distance between clusters and small internal variation, so that the cluster quality can be categorized as good. The use of DBI as an internal validation measure is in line with unsupervised learning-based research practices that aim to assess the effectiveness of clustering results without requiring external data labels.

In this study, the clustering results are visualized in a three-dimensional (3D scatter plot) graph and an Excel-based map of Indonesia, following the spatial visualization approach that has been applied in previous studies [10]. which displays the distribution of GRDP per capita, poverty rates, and HDI in each province. This map visualization aims to reinforce the interpretation of the analysis results by showing the actual economic conditions between regions.

In general, the stages of this research include: data collection, preprocessing, determining the optimal number of clusters, applying the K-Means algorithm, validating the clustering results, and visualizing and interpreting the results.

3. Result and Discussion

3.1. Determining the Optimal Number of Clusters Using the Elbow Method

Based on the results of the analysis using the Elbow method (Figure 1), the optimal value obtained is k = 3. This value indicates that dividing the data into three clusters is the most efficient representation of the variation in economic indicators between provinces in Indonesia. The selection of these three clusters is based on the elbow point, where the decline in inertia begins to slow down significantly. Thus, all provinces are grouped into three main categories, namely low, medium, and high clusters, based on their economic characteristics. The k value obtained is then used in the application of the K-Means algorithm to produce the most optimal grouping of provinces.

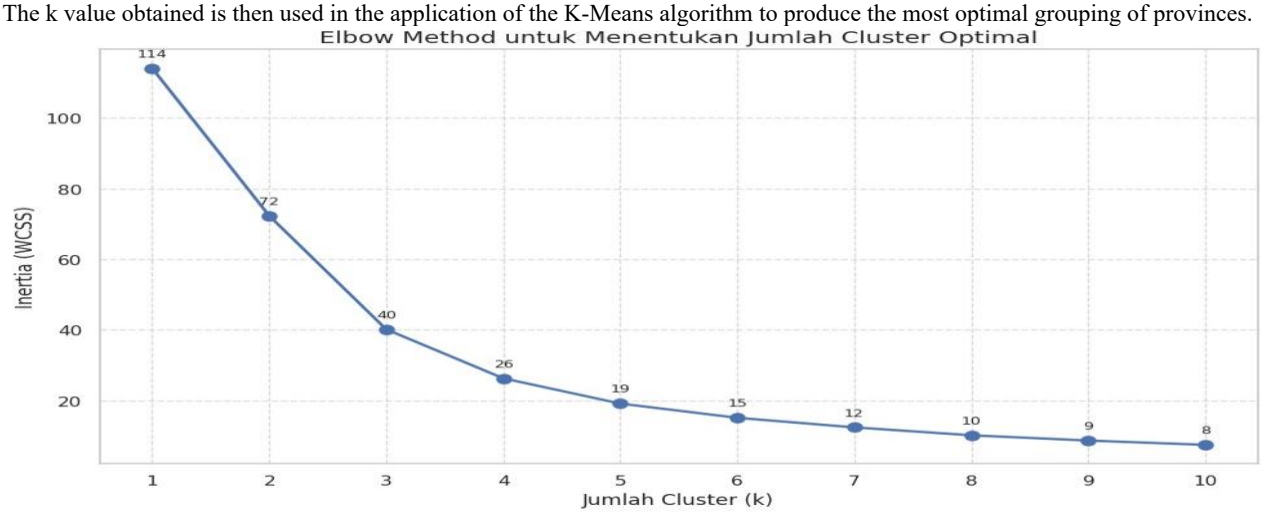

Fig. 1: Elbow Method for Determining the Optimal Number of Clusters

3.2. Statistical Characteristics of Economic Indicators for Each Cluster

Descriptive statistics for each cluster are presented in Table 1, which includes the minimum, maximum, and average values of three main indicators, namely the Human Development Index (HDI), Gross Regional Domestic Product (GRDP) per capita, and the percentage of poor people.

Table 1: Descriptive Statistics of Economic Indicators per Cluster

Cluster	Human Development Index			GRDP per capita			Percentage of Poor Population		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
0	53.42	68.63	64.02	18105.38	131636.45	68905.70	17.54	31.32	22.76
1	68.20	81.55	73.29	32198.14	120749.57	63329.60	3.90	17.68	9.28
2	73.02	83.08	77.54	161424.32	344349.76	216345.62	4.22	6.52	5.46

The analysis shows that Cluster 2 has the highest average GRDP per capita and HDI and the lowest poverty rate, so it can be categorized as a province with a high level of economic development. Conversely, Cluster 0 has the lowest HDI and GRDP values and the highest poverty rate, indicating that its economic development is still lagging behind. Cluster 1 is in the middle with economic indicator values that are fairly balanced between the other two groups. These differences in average values form the basis for further visualization to see the patterns between clusters more clearly through a heat map.

3.3 Visual Analysis of Intercluster Differences

To clarify the differences in characteristics between clusters, a normalized heatmap visualization is used for each indicator, as shown in Figure 2. The normalization process is carried out so that each indicator has a visually comparable scale, allowing differences between clusters to be observed proportionally without being influenced by differences in units or value ranges. In this heatmap, darker colors indicate a higher relative position of the indicator on the normalized scale (0–1). It can be seen that Cluster 2 ranks highest on the Human Development Index (HDI) and GRDP per capita variables, and has the lowest poverty rate compared to other clusters. Conversely, Cluster 0 shows low HDI and GRDP per capita values.

With a much higher poverty rate, while Cluster 1 is in the middle position, reflecting regions with moderate economic conditions. This visualization confirms a consistent pattern of welfare, where provinces with high economic productivity (high GRDP per capita) also tend to have better HDI and lower poverty rates. Thus, this heatmap reinforces the finding that improved economic performance in a region has a positive relationship with community welfare.

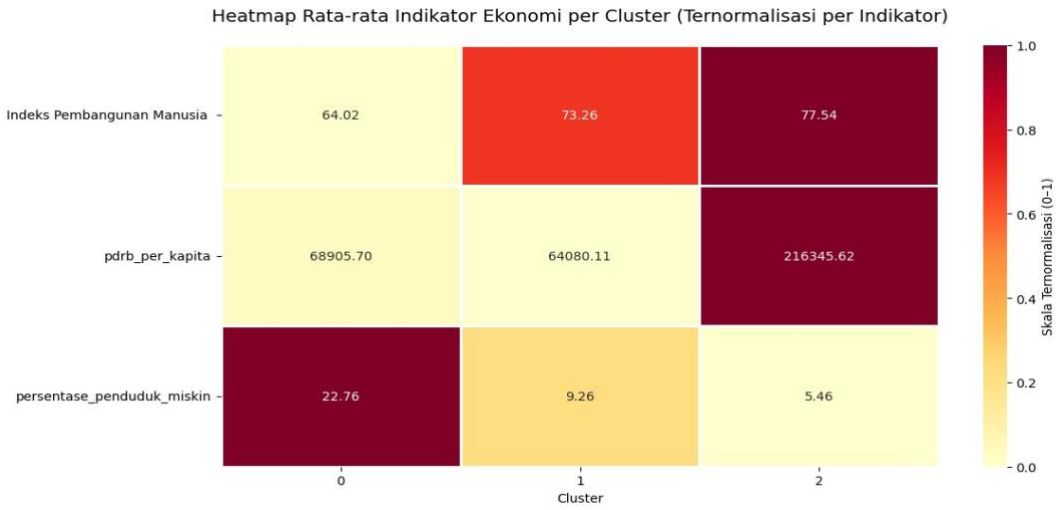

Fig. 2: Average Economic Indicator Heatmap per Cluster

Next, the clustering results are visualized in a three-dimensional (3D scatter plot) graph, as shown in Figure 3. Each color represents a cluster formed from three key economic indicators. The results show clear boundaries between clusters, with most provinces falling into the lower-middle group. This distribution indicates that economic development in Indonesia is still uneven across regions.

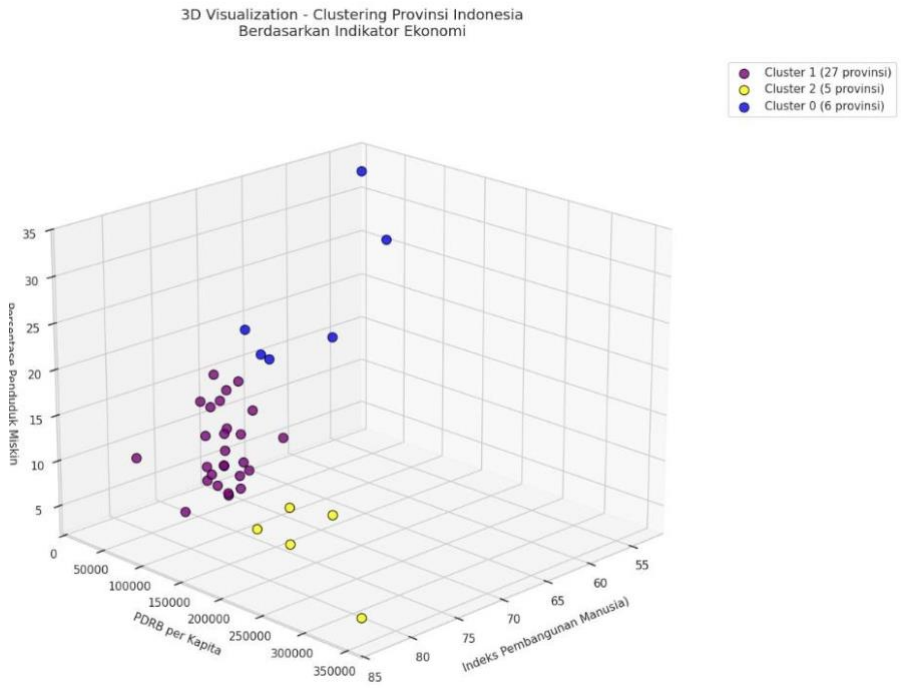

Fig 3: 3D visualization of Indonesian provinces based on economic indicators

3.4. Clustering Results Validation

To measure the validity of the clustering results, an evaluation was conducted using the Davies–Bouldin Index (DBI). The DBI value obtained was 0.79, indicating that the cluster results had fairly good separation and strong internal homogeneity. This value supports the fact that the division into three clusters provides a stable representation of the variation in data between provinces.

3.5. Classification of Provinces Based on Clustering Results

A complete list of provincial clustering results can be seen in Table 2. Based on these results, Cluster 2 is categorized as a group of provinces with high economic development, including regions such as Riau, Riau Islands, DKI Jakarta, East Kalimantan, and North Kalimantan. Meanwhile, Cluster 0 is a group with low economic development, dominated by provinces in eastern Indonesia such as East Nusa Tenggara, Maluku, and Papua. Cluster 1 contains most of the provinces in Indonesia with medium economic characteristics, including Aceh, Central Java, South Sulawesi, and West Kalimantan.

Table 2: List of Provinces Based on Clustering Results

Cluster	Categori	Province
0	Low	East Nusa Tenggara, West Papua, Southwest Papua, South Papua, Central Papua, Papua Mountains
1	Middle	Aceh, North Sumatra, West Sumatra, Jambi, South Sumatra,Bengkulu, Lampung, Bangka Belitung Islands, West Java West Java, Central Java, Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, Papua
2	High	Riau, Riau Islands, Jakarta, East Kalimantan, North Kalimantan

The classification results are visualized spatially in Figure 4, which shows the distribution of clusters in the geographical context of Indonesia. The pattern shows a concentration of provinces with high economic indicators in the regions of Sumatra and Kalimantan, while all provinces in eastern Indonesia (Papua and Nusa Tenggara) are included in the cluster with low indicators. Interestingly, not all provinces on the island of Java are included in the high cluster—only DKI Jakarta falls into this category, while the other provinces of Java are in the middle cluster. This pattern reinforces the finding that Indonesia's economic development is not only concentrated in the west, but more specifically in regions with abundant natural resources and certain economic centers.

Fig. 4: Clustering Map of Indonesian Provinces Based on Economic Indicators

4. Conclusion

This study successfully applied the K-Means algorithm to cluster 38 provinces in Indonesia based on three key economic indicators, namely per capita GRDP, percentage of poor population, and HDI. The analysis results determined three optimal clusters representing high (5 provinces), medium (27 provinces), and low (6 provinces) levels of economic development.

The findings of the study reveal significant spatial patterns: all provinces in eastern Indonesia (Papua and Nusa Tenggara) are concentrated in the low cluster, while provinces with high economic indicators are dominated by the resource-rich regions of Sumatra and Kalimantan. Interestingly, only DKI Jakarta represents Java in the high cluster, indicating the complexity of development factors beyond mere geographical location.

References

[1] G. R. Wahyudi and E. Dini, "Clustering Regencies in Indonesia for Regional Mapping Using the K-Means Algorithm Pengelompokan Kabupaten di Indonesia untuk Pemetaan Daerah Menggunakan Algoritma K-Means," vol. 5, no. July, pp. 1143–1151, 2025.

[2] Mochamad Arief Darmawan, "PENGARUH KEBIJAKAN FISKAL TERHADAP TINGKAT KETIMPANGAN ANTAR PROVINSI DI INDONESI," 2024.

[3] M. Maichal, P. G. Hartono, A. Firman, and I. M. E. K. Yudha, "The Influence of Gross Regional Domestic Product Per Capita and Foreign Direct Investment on Income Inequality: An Empirical Study of 34 Provinces in Indonesia," *J. Econ. Res. Soc. Sci.*, vol. 8, no. 2, pp. 197–206, 2024, doi: 10.18196/jerss.v8i2.23256.

[4] T. Novianti and D. V. Panjaitan, "Income Inequality in Indonesia: Before and during the Covid19 Pandemic," *Int. J. Econ. Financ. Issues*, vol. 12, no. 3, pp. 29–37, 2022, doi: 10.32479/ijefi.12996.

[5] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.

[6] R. Rahmad, S. Defit, and R. Sovia, "BULLETIN OF COMPUTER SCIENCE RESEARCH, Analisis Data Mining dengan Metode K-Means Clustering Dalam Pengelompokan Penggunaan Alat Kontrasepsi," vol. 5, no. 5, pp. 1174–1181, 2025, doi: 10.47065/bulletincsr.v5i5.750.

[7] I. Fahmiyah and R. A. Ningrum, "Human Development Clustering in Indonesia: Using K-Means Method and Based on Human Development Index Categories," *J. Adv. Technol. Multidiscip.*, vol. 2, no. 1, pp. 27–33, 2023, doi: 10.20473/jatm.v2i1.45070.

[8] J. M. Guntur, "Algoritma K-Means untuk Meningkatkan Silhouette Score pada Pengelompokan Data Stok Bahan Manufaktur di PT. XYZ Kabupaten Majalengka," *Multinetics*, vol. 11, no. 1, pp. 11–21, 2025, doi: 10.32722/multinetics.v11i1.7259.

[9] N. A. Y. -, L. E. B. -, G. C. H. R. -, M. F. Z. -, A. -, and F. R. -, "ANALISIS PERBANDINGAN K-MEANS DAN DBSCAN DALAM PENGELOMPOKAN DATA TRAVEL REVIEW RATINGS MENGGUNAKAN EVALUASI SILHOUETTE INDEX DAN DAVIES-BOULDIN INDEX," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, Jul. 2025, doi: 10.23960/jitet.v13i3.6884.

[10] A. Syahputra and M. Ikhsan, "Data Visualization using Tableau With K-Mean Clustering Method uor Identification Of Poor Areas in North Sumatera," *INOVTEK Polbeng - Seri Inform.*, vol. 10, no. 2, pp. 1218–1227, 2025, doi: 10.35314/pqkbwx68.