Journal of Artificial Intelligence
and Legal Technology

SAH PUBLISHER

e-ISSN 3123-786X
Vol. 2, No. 2, August 2026

Cluster Analysis of Open Unemployment Rates Across Indonesian Provinces Using the K-Means Algorithm

Egi Dio Bagus Sudewo[1]*, Fadila Khairani[2], Irawati[3], Niken Ayu[4], Andri Syahfikri[5]
[1,2,3,4,5]Sistem Informasi, Universitas Royal, Asahan, Indonesia

Abstract. Unemployment remains a persistent challenge in Indonesia's regional economic development, reflected in the variation of the Open Unemployment Rate (OUR) across provinces. This study aims to identify and classify regional unemployment patterns using a data-driven clustering approach. A quantitative design with an unsupervised learning framework was employed, utilizing secondary data on provincial Open Unemployment Rates for 2025 obtained from Statistics Indonesia (BPS). The K-Means clustering algorithm was applied after data preprocessing and normalization, while the optimal number of clusters was determined using the Elbow Method. The results indicate that three clusters represent the most appropriate grouping structure, categorizing provinces into low (3.03%), moderate (4.41%), and high (6.32%) unemployment clusters. The findings reveal significant regional heterogeneity in labor market conditions and demonstrate that unemployment disparities are structurally differentiated across provinces. This study confirms the effectiveness of K-Means in uncovering latent regional patterns without relying on causal assumptions and provides an empirical foundation for developing targeted, region-sensitive employment policies in Indonesia.

.

Keywords: Open Unemployment Rate; K-Means Clustering; Regional Labor Market; Unsupervised Learning
Received Feb 2026 / **Revised** Aug 2026 / **Accepted** Aug 2026

INTRODUCTION

Unemployment continues to represent a critical challenge in the economic development trajectory of developing countries, including Indonesia. The Open Unemployment Rate (OUR) serves as a key indicator in assessing labor market performance, as it measures the proportion of the labor force that is not actively engaged in productive employment [1]. A persistently high unemployment rate reflects an imbalance between labor force growth and job creation capacity, which may lead to broader macroeconomic and social consequences such as slower economic expansion, increased poverty, and widening regional inequality. Consequently, understanding the structural patterns of unemployment is essential for designing effective and evidence-based development policies [2].

Indonesia's regional diversity in terms of economic structure, industrial concentration, urbanization, and human capital quality contributes to heterogeneous labor market outcomes across provinces. Variations in productive sector dominance and labor absorption capacity result in differing levels of unemployment, indicating that labor market challenges are inherently region-specific [3]. This heterogeneity suggests that standardized national policies may not adequately address localized labor market issues, thereby necessitating analytical approaches capable of identifying regional patterns and similarities.

Previous empirical studies have examined unemployment disparities using quantitative and econometric techniques, including clustering methods such as K-Means. While these approaches have provided insights into relationships between unemployment and macroeconomic indicators, many analyses emphasize

[1]*Corresponding author.
Email addresses: egidio8gali@gmail.com (sudewo)

associative or causal modeling frameworks that rely on specific assumptions [4]. In contrast, unsupervised learning methods offer an alternative perspective by grouping regions based on intrinsic similarities without requiring predefined dependent variables, thus enabling a more exploratory and data-driven segmentation of labor market characteristics [5].

This study applies the K-Means clustering algorithm to classify Indonesian provinces according to similarities in their Open Unemployment Rates using the most recent 2025 data published by Statistics Indonesia. By identifying regional clusters, the research aims to generate a systematic mapping of unemployment patterns that can serve as a foundation for targeted policy formulation. The findings are expected to contribute both to the development of machine learning applications in regional economic analysis and to the formulation of more adaptive and region-sensitive labor market interventions.

METHODS

This study employs a quantitative research design utilizing a data mining approach to examine the regional patterns of the Open Unemployment Rate (OUR) in Indonesia. The primary objective is to classify Indonesian provinces based on their unemployment levels in order to identify structural similarities in regional labor market conditions. An unsupervised learning framework is adopted because the study does not aim to predict future unemployment rates but rather to explore inherent patterns and groupings within the existing data. This exploratory approach enables a systematic segmentation of provinces according to comparable unemployment characteristics. This research flow as shown at Figure 1.

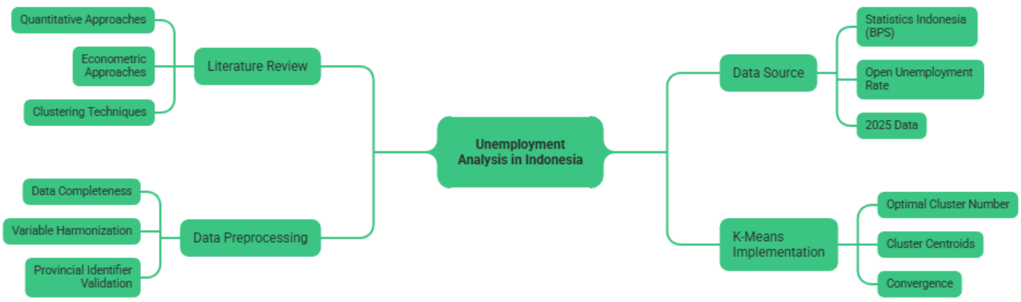

Figure 1. Research Flow

Literature review

Previous studies on unemployment in Indonesia have primarily employed quantitative and econometric approaches to examine the relationship between the Open Unemployment Rate (OUR) and macroeconomic variables such as economic growth, education level, investment, and regional development indicators. Several empirical works have also applied clustering techniques, particularly the K-Means algorithm, to group regions based on economic or labor market similarities, demonstrating the method's effectiveness in revealing latent structural patterns without requiring strict causal assumptions. In the broader context of regional economic analysis, unsupervised learning methods have gained increasing attention due to their ability to identify homogeneous groups within heterogeneous datasets, thereby supporting data-driven policy formulation. These studies collectively underline that spatial disparities in unemployment are structurally embedded and that clustering-based approaches provide a robust analytical framework for understanding regional labor market segmentation.

Data Source

The analysis is based on secondary data obtained from Statistics Indonesia (Badan Pusat Statistik/BPS). The dataset consists of the Open Unemployment Rate by province for the year 2025, expressed in percentage values. Each observation includes the provincial name and its corresponding unemployment rate. The use of official data from BPS ensures the validity, reliability, and credibility of the information employed in this study, thereby strengthening the empirical foundation of the analysis.

Table 1. dataset

Province	**OUR (%)**

Aceh	5.5
North Sumatra	5.1
West Sumatra	5.7
Riau	4.1
Jambi	4.5
South Sumatra	3.9
Bengkulu	3.2
Lampung	4.1
Bangka Belitung Islands	4.2
Riau Islands	6.9
Jakarta Special Capital Region	6.2
West Java	6.7
Central Java	4.3
Special Region of Yogyakarta	3.2
East Java	3.6
Banten	6.6
Bali	1.6
West Nusa Tenggara	3.2
East Nusa Tenggara	3.2
West Kalimantan	4.2
Central Kalimantan	3.5
South Kalimantan	3.9
East Kalimantan	5.3
North Kalimantan	3.9
North Sulawesi	6.0
Central Sulawesi	3.0
South Sulawesi	5.0
Southeast Sulawesi	3.3
Gorontalo	3.1
West Sulawesi	3.2
Maluku	6.0
North Maluku	4.3
West Papua	4.2
Southwest Papua	6.6
Papua	6.9
South Papua	4.9
Central Papua	3.6
Highland Papua	1.7

Data Preprocessing and Cluster Determination

Prior to clustering, the dataset undergoes a preprocessing stage to ensure data consistency and analytical robustness. This stage includes verification of data completeness, harmonization of variable formats, and validation of provincial identifiers [6][7]. To enhance computational stability, the data are normalized using the StandardScaler technique so that all values are transformed onto a consistent scale. Normalization is essential because the K-Means algorithm is sensitive to scale variations, which may otherwise distort distance calculations [8]. The optimal number of clusters is determined using the Elbow Method, which evaluates the within-cluster sum of squares (WCSS) across different cluster configurations. The point at which additional clusters produce diminishing reductions in WCSS—known as the elbow point—is selected as the most representative cluster structure.

K-Means Implementation and Cluster Evaluation

After establishing the optimal number of clusters, the K-Means algorithm is applied to group provinces based on proximity to cluster centroids [9][10]. The algorithm iteratively recalculates centroid positions until convergence is achieved, indicated by minimal changes in centroid movement. All analytical procedures are conducted using Python within the Google Colab environment. To assess clustering quality, the Silhouette Score is employed as an evaluation metric. This measure quantifies the degree of cohesion within clusters and separation between clusters, with values ranging from −1 to 1. Higher Silhouette values indicate clearer cluster boundaries and stronger internal consistency, ensuring that the clustering results are methodologically sound and interpretable within a scientific framework [11].

RESULT AND DISCUSSION

The determination of the optimal number of clusters was conducted using the Elbow Method, as presented in Figure 1. The inertia curve demonstrates a substantial decline between k = 1 and k = 3, followed by a relatively marginal reduction beyond that point. This pattern indicates that three clusters represent the most appropriate structure for capturing the variability of the Open Unemployment Rate (OUR) across provinces. Selecting k = 3 balances explanatory power and model simplicity, ensuring that the clustering structure meaningfully represents regional unemployment disparities without introducing unnecessary segmentation..

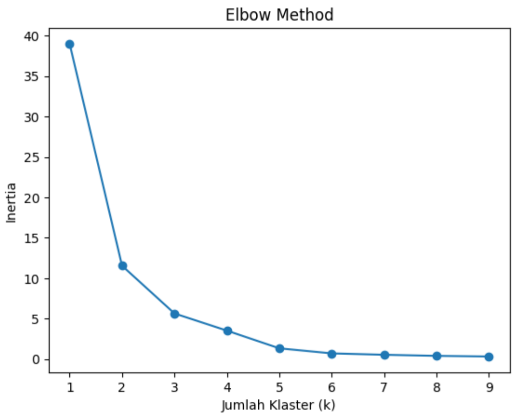

Figure 2. Elbow Method Graph for Determining the Optimal Number of Clusters (k = 3)

Based on this configuration, the K-Means algorithm classified Indonesian provinces into three clusters with distinct unemployment characteristics. Table 2 presents the average OUR for each cluster. Cluster 0 (low) records an average unemployment rate of 3.03%, Cluster 1 (moderate) 4.41%, and Cluster 2 (high) 6.32%. The numerical gap among these averages confirms a clear stratification of labor market conditions, indicating substantial interprovincial heterogeneity in unemployment levels.

Table 2. Average Open Unemployment Rate (OUR) by Cluster.

Cluster	Average OUR (%)
Cluster 0 (Low)	3.03
Cluster 1 (Moderate)	4.41
Cluster 2 (High)	6.32

Table 3 shows the composition of provinces within the low-unemployment cluster. Provinces such as Bali (1.58%), Papua Pegunungan (1.68%), and several provinces in Sulawesi and Kalimantan fall into this category, with unemployment rates consistently below 4%. This distribution suggests relatively stable employment absorption and balanced labor market dynamics in these regions..

Table 3. Provinces in the Low Cluster (Cluster 1)

Province	OUR (%)
Bali	1.58
Highland Papua	1.68
Central Sulawesi	3.02
Gorontalo	3.12
West Sulawesi	3.17
Special Region Of Yogyakarta	3.18
West Nusa Tenggara	3.22
East Nusa Tenggara	3.23
Bengkulu	3.24
Southeast Sulawesi	3.27
Central Kalimantan	3.47
Central Papua	3.55
East Java	3.61

Table 4 presents provinces classified in the moderate cluster, where unemployment rates range approximately between 3.89% and 5.05%. This group includes South Sumatra, Riau, Central Java, North Sumatra. These provinces represent intermediate labor market conditions, characterized by moderate unemployment pressures.

Table 4. Provinces in the Moderate Cluster (Cluster 2)

Province	OUR (%)
South Sumatra	3.89
North Kalimantan	3.90
South Kalimantan	3.94
Lampung	4.07
Riau	4.12
Bangka Belitung Islands	4.17
West Papua	4.21
West Kalimantan	4.23
North Maluku	4.26
Central Java	4.33
Jambi	4.48
South Papua	4.90
South Sulawesi	4.96
North Sumatra	5.05

Table 5 identifies provinces in the high-unemployment cluster, including East Kalimantan (5.33%), DKI Jakarta (6.18%), West Java (6.74%), and Papua (6.92%). Provinces in this cluster exhibit unemployment rates exceeding 5%, reflecting stronger labor market pressures compared to other clusters

Table 5. Provinces with High Unemployment Rate (Cluster 2).

Province	OUR (%)

East Kalimantan	5.33
Aceh	5.50
West Sumatra	5.69
Maluku	5.95
North Sulawesi	6.03
Jakarta Special Capital Region	6.18
Southwest Papua	6.61
Banten	6.64
West Java	6.74
Riau Islands	6.89
Papua	6.92

The interpretative summary presented in Table 6 links each cluster to its corresponding policy orientation. The low cluster emphasizes employment sustainability, the moderate cluster highlights workforce quality improvement, and the high cluster calls for prioritized labor market intervention. Figure 3 further visualizes the clustering results, illustrating clear separation among clusters and confirming the robustness of the grouping structure.

Table 6. Interpretation of Open Unemployment Rate Clusters

Cluster	OUR Characteristics	Policy Implications
Low	Relatively low OUR (< 4%)	Maintenance of existing job opportunities and labor market stability.
Moderate	Moderate OUR (4% – 5%)	Enhancing labor force quality and strengthening education-industry linkages.
High	High OUR (> 5%)	Priority labor interventions and structural job creation strategies.

Table 5 provides a strategic framework for regional employment policies by categorizing provinces into three distinct tiers based on their unemployment risk profiles, ensuring that government interventions are targeted rather than uniform. For the Low Cluster (OUR < 4%), the focus is on maintaining current labor market stability, whereas the oderate Cluster (OUR 4–5%) requires human resource quality improvements and stronger linkages between vocational education and industrial needs. Finally, the High Cluster (OUR > 5%) signifies a critical need for priority interventions, such as structural job creation and industrial incentives, to address the systemic challenges found in Indonesia's major urban and manufacturing hubs

Discussions

The findings demonstrate that unemployment patterns in Indonesia are spatially heterogeneous and can be systematically classified using an unsupervised learning approach. The existence of three distinct clusters indicates that regional labor markets operate under differing structural conditions. Provinces in the low-unemployment cluster exhibit relatively strong employment absorption capacity, suggesting that economic structures in these regions are capable of accommodating labor force growth. Policy strategies in these areas should therefore focus on maintaining employment stability and enhancing productivity rather than large-scale corrective measures. The medium-unemployment cluster reflects transitional labor market conditions. Provinces in this group face moderate imbalances between labor supply and demand. This situation may be associated with structural transformation processes, shifts in sectoral dominance, or demographic pressures. Accordingly, policy responses should prioritize human capital development, vocational training enhancement, and improved alignment between educational outputs and regional industrial needs.

In contrast, provinces within the high-unemployment cluster experience more pronounced structural labor market constraints. Elevated unemployment levels in these areas may stem from limited job creation in formal sectors, urban labor market congestion, or uneven regional economic development. These findings reinforce the argument that unemployment disparities in Indonesia are inherently regional and require geographically differentiated policy interventions rather than uniform national strategies.From a methodological perspective, the results confirm the effectiveness of the K-Means algorithm in uncovering latent regional structures within economic data without relying on causal modeling assumptions. By integrating statistical clustering results with policy interpretation, this study provides a systematic and data-driven framework for understanding regional unemployment dynamics. The cluster-based approach offers practical implications for policymakers, enabling the formulation of targeted, region-sensitive employment strategies grounded in empirical evidence.

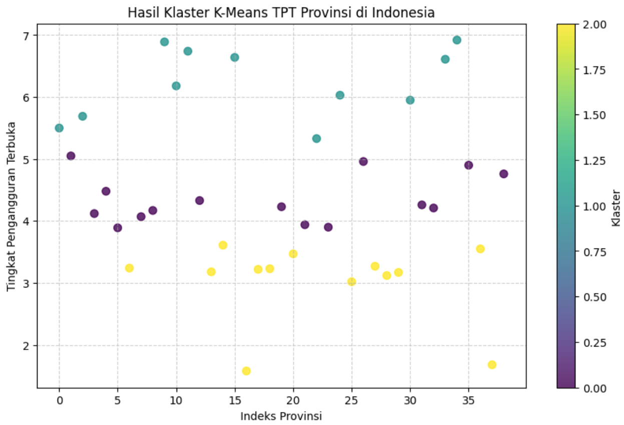

Figure 3. Visualization of K-Means Clustering Results for Open Unemployment Rate by Province in Indonesia

CONCLUSION

This study confirms that the Open Unemployment Rate (OUR) in Indonesia demonstrates significant regional heterogeneity that can be effectively classified using the K-Means clustering approach. The Elbow Method identifies three optimal clusters—low (3.03%), moderate (4.41%), and high (6.32%) unemployment—indicating substantial disparities in provincial labor market conditions. These findings highlight that unemployment challenges are structurally differentiated across regions and therefore require policy responses that are context-specific rather than uniform at the national level. By applying an unsupervised learning framework, this research provides a systematic, data-driven foundation for regional labor market segmentation and offers practical insights for designing targeted and region-sensitive employment policies.

REFERENCES

[1] M. M. Anwar, N. Aida, and A. R. Y. Taher, "The Effect of Open Unemployment and Education on Poverty in Indonesia," *Int. J. Accounting, Manag. Econ. Soc. Sci.*, vol. 3, no. 4, pp. 1384–1393, 2025, doi: 10.61990/ijamesc.v3i4.556.

[2] Y. Tang, "Some Causes of Unemployment and Their Relationship with the Creation and the Destruction of Jobs," 2024, doi: 10.61173/1asfks49.

[3] A. Bina, L. K. Optari, and H. Widiya, "Income Inequality Across Provinces in Indonesia: An Empirical Analysis of Human Development, Economic Performance, and Labor Market Conditions," *Priviet Soc. Sci. J.*, vol. 5, no. 6, pp. 119–126, 2025, doi: 10.55942/pssj.v5i6.617.

[4] A. I. Ramadhan, P. D. Atika, and K. F. Ramdhania, "Analisis Clustering K-Means untuk Pemetaan Tingkat Pengangguran Terbuka di Provinsi-Provinsi Indonesia Tahun 2013–2023," *J. Students' Res. Comput. Sci.*, vol. 5, no. 2, pp. 109–122, 2024, doi: 10.31599/wbpydb62.

[5] F. Rahmawati and S. I. Fallo, "Hierarchical Agglomerative Clustering dengan Metode Ward untuk Pemetaan Pasar Tenaga Kerja Pascapandemi di Jawa Tengah: Pendekatan Machine Learning Berbasis Klasterisasi," *Leibniz*, vol. 5, no. 01, pp. 65–77, 2025, doi: 10.59632/leibniz.v5i01.573.

[6] Y. Jiang, "Research on Optimizing Data Preprocessing Using Clustering Algorithms," 2025. doi: 10.1109/iccea65460.2025.11103230.

[7] W. M. Baihaqi, Chyntia Raras Ajeng Widiawati, and Tegar Insani, "K-Means Clustering Based on Otsu Thresholding For Nucleus of White Blood Cells Segmentation," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 907–914, 2020, doi: 10.29207/resti.v4i5.2309.

[8] J. Wala and R. Umar, “Implementasi K-Means Clustering pada Pengelompokan Pasien Penyakit Jantung,” vol. 9, no. 3, pp. 205–216, 2024.
[9] B. N. Yulisasih, H. Herman, S. Sunardi, and H. Yuliansyah, “Evaluation of K-Means Clustering Using Silhouette Score Method on Customer Segmentation,” *Ilk. J. Ilmiah; Vol 16, No 3 (2024)DO - 10.33096/ilkom.v16i3.2325.330-342*, Dec. 2024, [Online]. Available: https://jurnal.fikom.umi.ac.id/index.php/ILKOM/article/view/2325
[10] J. Arriaga Hernández, B. Cuevas Otahola, A. Diaz Nayotl, A. Jaramillo Núñez, A. Pérez Villegas, and O. Valenzuela, “Machine learning K-means algorithm applied to wavefront sensing in Bi-Ronchi/Hartmann tests with SLM,” *Optik (Stuttg).*, vol. 346, p. 172607, 2026, doi: https://doi.org/10.1016/j.ijleo.2025.172607.
[11] V. S. R. Fatmawaty Imam; Herman, Herman, “Klasterisasi Perguruan Tinggi LLDIKTI V Berdasarkan Indikator Kinerja Utama dan PDDIKTI Menggunakan K-Means Clustering,” *J. MEDIA Inform. BUDIDARMA*, no. Vol 8, No 2 (2024): April 2024, pp. 878–888, 2024, [Online]. Available: https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/7497/3852