Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam
Volume 4, Number 2, August 2025
p-ISSN: 2828-9382; e-ISSN: 2828-9390, Hal. 292-304
DOI: https://doi.org/10.55606/jurrimipa.v4i2.6657
Available Online at : https://prin.or.id/index.php/JURRIMIPA

K-Means Clustering to Classify Indonesian Provinces Based on School Participation and Socio-Economic Indicators

Nilam Novita Sari[1*], Khaola Rachma Adzima[2], Sahiba Sahila[3], Tiara Husnul Khotimah[4]
[1] Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, Indonesia
[2,4] Prodi Pendidikan Matematika, Fakultas Matematika Ilmu Pengetahuan Alam, Universitas Negeri Jakarta, Indonesia
[3] Prodi Teknologi Rekayasa Cetak dan Grafis 3 Dimensi, Jurusan Teknik Grafika dan Penerbitan, Politeknik Negeri Jakarta, Indonesia
*Email : nilam.novita@unj.ac.id[1]**

Alamat: Jl. Rawamangun Muka Raya No 11, East Jakarta, 13220, Indonesia
**Korespondensi penulis*

***Abstract**. Education serves as a fundamental pillar in national development, as it not only enhances individual capacities but also improves overall social welfare. Despite this crucial role, Indonesia continues to face disparities in both access to and quality of education among its regions, as can be seen from variations in school participation indicators and socio-economic backgrounds. To analyze these differences, this study applied the K-Means Clustering method to categorize provinces in Indonesia using six variables: School Participation Rate, Net Enrollment Rate, Gross Enrollment Rate, Poverty Rate, High School Ratio, and Teacher Ratio. To identify the most suitable number of clusters, three validation indices were utilized, namely Dunn Index, C-Index, and Davies-Bouldin Index, with cluster counts tested from three to six. The results indicated that the best clustering solution was five clusters, as reflected in the highest Dunn Index (0.1569), lowest C-Index (0.0321), and lowest Davies-Bouldin Index (0.5062). The robustness of this clustering was further supported by the ratio between within-cluster and between-cluster standard deviation (S(w)/S(b) = 0.33). Each cluster revealed unique characteristics of education and socio-economic conditions, where Cluster 4 displayed the most favorable outcomes with high participation and low poverty levels, whereas Cluster 5 highlighted the weakest performance across all observed indicators.*

***Keywords:** Davies-Bouldin index; K-Means clustering; School participation; Socio-economic indicator; Validation indices.*

1. BACKGROUND

Education was a consciously and systematically designed process aimed at creating an environment and learning experiences that encouraged students to develop their full potential (Munna, 2021). It involved strengthening spiritual and religious values, self-control, character building, cognitive development, cultivation of noble character, and the acquisition of skills relevant to personal life and contributions to society (Turienzo, 2024). Education played a crucial role in individual life and served as a driving force for national progress (Haleem, 2023). It was also a strategic factor in determining success and achieving a better future (Chigbu, 2021). However, along with the advancement of time and technology, the field of education faced various challenges, particularly in developing countries such as Indonesia. One of the major challenges in Indonesian education was the inequality in access to and quality of education, where disparities persisted between urban and rural areas as well as across regions in obtaining proper educational services (Baharuddin, 2025). In addition, the quality of

Manuscript Submission: July 31, 2025; Revised: August 16, 2025; Accepted: August 27, 2025; Published: August 29 2025;

educators, the availability of infrastructure, and the utilization of technology were also critical factors that significantly influenced the quality of education. Many teachers in remote areas lacked adequate competencies and faced limited access to training and up-to-date information sources (Nguyen, 2026). This situation negatively impacted the effectiveness of the teaching and learning process in schools. Therefore, it was essential to establish indicators that can be used to assess the quality and success of education at the national level.

One of the indicators used to measure the success of the Indonesian government in the field of education was the School Participation Rate defined as the ratio of students within a specific age group who were currently enrolled in education at various levels, compared to the total population in that age group, expressed as a percentage (BPS, 2024). This measure served as an indicator of the effectiveness of the education system and could be used to assess educational performance in a region. However, as the level of education increased, the participation rate tended to decrease. In addition to regional disparities in educational access and quality, the rate might also have been influenced by economic factors. Higher levels of education were generally associated with increased costs, which could have limited individuals from lower economic backgrounds from continuing their education to higher levels (Batool, 2021).

To improve educational equity, regions could be grouped based on school participation and socio-economic factors using relevant variables. This clustering aimed to identify regions with similar characteristics in terms of school participation and socio-economic conditions, thereby assisting the government in formulating more targeted intervention strategies, especially in areas with low school participation rates and socio-economic status.

Cluster analysis was a data analysis method that aimed to group objects into several clusters based on certain shared characteristics (Suraya, 2023). In the clustering process, objects with similar characteristics were grouped into the same cluster, while objects with differing characteristics were placed in separate clusters. This ensured that the objects within each cluster were highly similar (homogeneous), while the clusters themselves were distinctly different (heterogeneous) (Schröder, 2022). The degree of similarity between data points was determined by the distance between them; smaller distances indicated higher similarity, while larger distances reflected lower similarity (Schröder, 2022).

K-Means was a non-hierarchical clustering method used to divide *n* objects into *k* clusters based on their similarity. The similarity among cluster members was measured based on their proximity to the cluster's average value (centroid) (Tabianan, 2022). The greater the similarity between objects, the smaller the distance between them (Junhui, 2020). The Euclidean distance formula was used to measure proximity between each data point and the centroid (Abou-Moustafa, 2016). Each data point was then assigned to the cluster with the nearest centroid (Muhajir, 2018). The advantages of the K-Means method included its ability to efficiently cluster large datasets, fast computation time, conceptual simplicity, and ease of implementation, making it widely used across various fields (Chong, 2021).

Numerous studies examined educational disparities and regional socio-economic differences in Indonesia. These studies typically relied on regression analysis to identify inequalities in school participation and access to educational facilities. However, there was a lack of research that systematically classified provinces based on school participation indicators and socio-economic indicators. The application of K-Means Clustering in this context remained underexplored, despite its potential to reveal hidden patterns and segment regions with similar educational and socio-economic profiles.

Moreover, existing clustering studies in education rarely applied multi-index validation techniques to determine the optimal number of clusters. Most previous research relied on a limited set of cluster validation metrics, which might have affected the robustness and interpretability of the clustering results.

Based on the identified research gaps, this study aimed to group Indonesian provinces based on educational participation and socio-economic indicators using the K-Means Clustering method. It classified all 38 Indonesian provinces using a comprehensive set of indicators at the senior high school level. This study also incorporated multiple cluster validity indices to determine the optimal number of clusters. The clustering results were expected to produce an optimal classification solution that provided insights into provinces with similar characteristics in terms of school participation. Furthermore, the findings served as a foundation for policy formulation and decision-making to improve educational access and equity across different regions.

2. THEORETICAL STUDY

Education and Socio-Economic Inequality

Education is one of the fundamental factors in national development because it can improve the quality of human resources and promote community welfare (Munna & Kalam, 2021). However, in Indonesia, there are still gaps in access to and quality of education, both between regions and between urban and rural areas (Baharuddin & Burhan, 2025) Indicators such as the School Participation Rate (SPR), Pure Participation Rate (PPR), and Crude Participation Rate (CPR) are often used to measure the achievement of educational equity.

Economic factors, such as poverty levels, also significantly influence access to education, where high education costs can limit opportunities for low-income communities (Batool & Liu, 2021).

Differences in teacher quality and the availability of educational infrastructure further widen the educational gap between regions. Teachers in remote areas often face limitations in training and access to up-to-date information, which impacts the effectiveness of the learning process (Nguyen et al., 2024). Thus, the relationship between socioeconomic conditions and educational achievement reveals patterns of inequality that need to be addressed through data-driven policies.

Cluster Analysis Theory and Concepts

Cluster analysis is a statistical method used to group objects into several clusters based on certain common characteristics (Suraya et al., 2023). The fundamental principle of this method is to group data such that members within a cluster exhibit a high degree of homogeneity, while clusters differ significantly from one another (Schröder & Kiko, 2022). This approach is effective in identifying hidden patterns in complex data, making it useful for mapping regions based on educational and socioeconomic conditions. One popular method is K-Means Clustering, a non-hierarchical technique that divides data into k clusters based on the distance to the centroid or center point (Tabianan et al., 2022). The clustering process is performed by calculating the proximity of each data point to the centroid using a distance measure, typically Euclidean Distance (Abou-Moustafa, 2016). The advantages of K-Means include efficiency in handling large datasets, fast computation time, and simple implementation (Chong, 2021)

3. RESEARCH METHODS

Data

The data used in this study were obtained from the Central Bureau of Statistics (Badan Pusat Statistik/BPS) for the year 2023, covering all 38 provinces in Indonesia. The data specifically focused on the senior secondary school age group. The variables used were as follows: (1) School Participation Rate: measured as the proportion of the school-age population currently enrolled in senior secondary education. A higher rate indicated better accessibility to education. (2) Gross Enrollment Rate (GER): compared the number of students, regardless of age, who were enrolled in senior secondary education with the total population of the official school-age group, expressed as a percentage. A higher Gross Enrollment Rate showed broader access to education, though it could also include students outside the official age group. (3) Net Enrollment Rate (NER): defined as the ratio between the total number of students enrolled in senior secondary education who were of the official age group and the population in that same age group, expressed as a percentage. A higher Net Enrollment Rate meant more students were attending school at the appropriate age. (4) Poverty Rate: indicated the percentage of the population whose income was below the poverty line, meaning they could not afford basic needs such as food, shelter, clothing, education, and healthcare. A lower poverty rate was considered better socio-economic conditions. (5) Teacher Ratio: referred to the number of students per teacher in senior secondary schools. A lower teacher ratio (e.g., 50 students per teacher) was preferable, as it allowed for better learning quality and interaction. (6) High School Ratio: referred to the average number of students per school at the senior secondary level. A high ratio might have indicated a shortage of high schools in a province. For example, a ratio of 100 implied that there were 100 students per high school.

Research Design

In this study, the provinces of Indonesia were grouped using the K-Means Clustering method. Prior to clustering, a validation test was conducted to determine the optimal number of clusters using the Dunn Index, C-Index, and Davies-Bouldin Index. Once the optimal number of clusters was identified, K-Means clustering was applied to group the provinces accordingly. The final step involved comparing the clustering results by calculating the ratio of within-cluster sum of squares ($S_{(w)}$) to between-cluster sum of squares ($S_{(b)}$) in order to determine the best clustering outcome. The steps were follows (Sinaga, 2020):

Determining the optimal number of clusters using internal validation indices: C-Index, Dunn Index, and Davies-Bouldin Index, based on their respective formulas as follows (Fernandes et al., 2022):

Dunn Index:

The Dunn Index was calculated using the equation as shown in Equation (1):

$$C = \frac{d_{min}}{d_{max}} \qquad (1)$$

where:

d_{min}: The smallest distance between objects in different clusters

d_{max}: The largest distance in the cluster

Davies-Bouldin Index

The Davies-Bouldin Index was computed based on the formulation provided in Equation (2):

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq j}\left[\frac{d'(c_i)+d'(c_j)}{d(c_i,c_j)}\right] \qquad (2)$$

where:

n: number of groups

$d(c_i, c_j)$: the distance between groups c_i and c_j

$d'(c_k)$: the distance in groups c_k

C-Index

The C-Index using the formulation shown in Equation (3):

$$C - Index = \frac{S_{(w)}-S_{min}}{S_{max}-S_{(w)}}, S_{min} \neq S_{max} \in (0,1) \qquad (3)$$

Where $S_{(w)}$ denotes the sum of intra-cluster distances and computed using Equation (4):

$$S_{(w)} = \sum_{k=1}^{q} \sum_{\substack{i,j\in C_k \\ i<j}} d(x_i, x_j) \qquad (4)$$

The minimum value of the index was used to indicate the optimal number of clusters. (1) Randomly selected K data points as the initial centroids (cluster centers). (2) Grouping the Objects by calculate the Euclidean distance from each data point to all centroids using the following formula as shown in Equation (5):

$$d(A,B) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x - y_n)^2} \qquad (5)$$

Each data point is then assigned to the cluster with the nearest centroid. For each cluster, the centroid was recalculated by taking the average position of all its members. These new centroids were used for the next iteration. Repeated the grouping and centroid recalculation steps. The iteration stopped when one of the following conditions was met: (a) There were no changes in cluster membership. (b) The centroid updates were minimal (convergence is reached). (c) The maximum number of iterations was reached.

After obtaining the clustering results, the next step was to compare the results by using the ratio of the average within-cluster standard deviation to the between-cluster standard deviation, using the following formulas.

Average within-cluster standard deviation ($S_{(w)}$)

The average within-cluster distance was calculated using the approach outlined in Equation (6).

$$S_{(w)} = \frac{1}{c}\sum_{k=1}^{c} S_k \quad (6)$$

Average between-cluster standard deviation ($S_{(b)}$)

The average between-cluster distance was derived using the method illustrated in Equation (7).

$$S_{(b)} = \left[\frac{1}{c-1}\sum_{k=1}^{c}(\bar{X}_k - \bar{X})^2\right]^{\frac{1}{2}} \quad (7)$$

All data analysis and clustering procedures were conducted using R programming language. The “diceR” packages were utilized for implementing K-Means and performing validation the clustering results.

4. RESULT AND DISCUSSION

In cluster analysis, determining the optimal number of clusters was done using several cluster validation indices, including the Dunn Index, C-Index, and Davies-Bouldin Index. A higher Dunn Index value indicated a better clustering result, whereas for the C-Index and Davies-Bouldin Index, lower values were preferred, as they indicated better cluster separation. In this study, the number of clusters tested ranged from 3 to 6. The validation results using the Dunn Index, C-Index, and Davies-Bouldin Index were presented in Table 1.

Table 1. Cluster validation test.

Number of Clusters	Dunn Index	C-Index	Davies-Bouldin Index
3	0.0182	0.1463	0.6604
4	0.1569	0.0455	0.5759
5	0.1569	0.0321	0.5062
6	0.0459	0.0649	0.6225

Table 1 showed that, based on the Dunn Index using cluster counts from 3 to 6, the optimal number of clusters was either 4 or 5, as these had the highest Dunn Index values. Using the C-Index, the optimal number of clusters was found to be 5, since cluster 5 had the lowest C-Index value. Similarly, using the Davies-Bouldin Index, the optimal cluster number was also 5, as it had the lowest Davies-Bouldin Index value. Based on these three validation indices, the optimal number of clusters for the K-Means Clustering method was either 4 or 5 clusters.

Therefore, clustering was conducted on the school participation data using both 4 and 5 clusters, and the results were compared to determine the most optimal solution. The clustering results using 4 clusters were shown in Table 2 below.

Table 2. K-Means with 4 clusters.

Cluster	Provinces
1	Aceh, North Sumatra, West Sumatra, Bengkulu, Riau Islands, Yogyakarta, Bali, West Nusa Tenggara, East Nusa Tenggara, East Kalimantan, North Kalimantan, Southeast Sulawesi, Maluku, North Maluku, West Papua, Southwest Papua, Papua
2	Gorontalo, West Sulawesi
3	Central Papua, Highland Papua
4	Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Islands, Jakarta, West Java, Central Java, East Java, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, South Papua

Table 2 presented the clustering result using 4 clusters. As shown, Cluster 1 included 16 provinces, Cluster 2 included 1 province, Cluster 3 included 2 provinces, and Cluster 4 included 17 provinces. The distribution indicated an imbalance in grouping the provinces because some clusters, such as Cluster 2 and Cluster 3, consisted of only one and two provinces, which could have reduced the interpretability and stability of centroids in the smaller clusters. Subsequently, the clustering results using five clusters were presented in Table 3.

Table 3. K-Means with 5 clusters.

Cluster	Provinces
1	Aceh, East Nusa Tenggara, East Kalimantan, Maluku, North Maluku, West Papua, Southwest Papua, Papua
2	Jambi, South Sumatra, Lampung, Bangka Belitung Islands, Jakarta, West Java, Central Java, East Java, Banten, West Kalimantan, Central Kalimantan, South Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, South Papua
3	Gorontalo, West Sulawesi
4	North Sumatra, West Sumatra, Riau, Bengkulu, Riau Islands, Yogyakarta, Bali, West Nusa Tenggara, North Kalimantan
5	Central Papua, Highland Papua

Table 3 showed the clustering result using 5 clusters, with Cluster 1 consisting of 8 provinces, Cluster 2 of 18 provinces, Cluster 3 of 2 provinces, Cluster 4 of 9 provinces, and Cluster 5 of 2 provinces. Compared with the four-cluster solution, the five-cluster solution reduced the heterogeneity within the largest groups and produced a more balanced distribution of provinces among clusters, while still retaining a few small clusters, such as Cluster 3 and Cluster 5, that captured outlying provincial profiles.

To evaluate the clustering performance, the ratio of within-cluster sum of squares ($S_{(w)}$) to between-cluster sum of squares ($S_{(b)}$) was calculated for both solutions. The $S_{(w)}/S_{(b)}$ ratios for 4 and 5 clusters were presented in Table 4.

Table 4. Goodness of the cluster method.

Number of Clusters	$S_{(w)}/S_{(b)}$
4	0.52
5	0.33

As shown in Table 4, the $S_{(w)}/S_{(b)}$ ratio for 4 clusters was 0.52, while the ratio for 5 clusters had a significantly lower value of 0.33. A lower $S_{(w)}/S_{(b)}$ ratio indicated better clustering performance, as it reflected more compact clusters and greater separation between clusters [15]. Therefore, the 5-cluster solution was determined to be the most optimal clustering result. Taken together with the validation indices (Table 1), these results supported the selection of five clusters for substantive interpretation and policy insights. The characteristics of each cluster were shown in Table 5.

Table 5. Cluster characteristics.

Cluster	School Participation Rate	Net Enrollment Rate	Gross Enrollment Rate	Poverty Rate	High School Ratio	Teacher Ratio
1	78.71	94.36	66.59	14.39	150.20	7.46
2	72.71	85.55	61.68	8.27	357.30	13.95
3	72.39	87.93	61.66	12.29	383.50	138.63
4	82.04	94.34	72.00	7.53	381.57	14.31
5	51.87	58.96	41.22	28.63	256.37	16.49
🟨	Lowest			🟩	Highest	

Table 5 showed the characteristics of each cluster, including the average values of School Participation Rate, Net Enrollment Rate, Gross Enrollment Rate, Poverty Rate, High School Ratio, and Teacher Ratio. The provinces in cluster 1 exhibited relatively good school participation, particularly with a very high Net Enrollment Rate (NER) of 94.36, indicating that a large proportion of school-age children (senior high school level) were enrolled in the education system. The poverty rate in this cluster was categorized as moderate compared to other clusters, suggesting that there were still groups of people whose income fell below the poverty line. The student-to-school ratio and student-to-teacher ratio in cluster 1 were the lowest, at 150.20 and 7.46, respectively. The student-to-teacher ratio in this cluster was notably low, with only seven students per teacher. This could have indicated two possibilities: the availability of adequate teaching resources and facilities, or a relatively small number of students enrolled in schools.

Provinces in Cluster 2 had moderate average values for School Participation Rate, Net Enrollment Rate, and Gross Enrollment Rate, indicating that a significant portion of school-aged children were either not enrolled or were not attending the appropriate level. The poverty rate in this cluster was low, but the high school ratio was relatively high, suggesting that existing schools might have been overburdened, with an average of over 350 students per school. The teacher ratio remained reasonably low.

Cluster 3 showed similar patterns to Cluster 2 in terms of School Participation Rate, Net Enrollment Rate, and Gross Enrollment Rate. However, the poverty rate was relatively high at 12.29%. Notably, Cluster 3 had the highest high school and teacher ratios, at 383.50 and 138.63, respectively. This indicated limited educational resources and an excessive teacher workload, that each teacher was responsible for an extremely high number of students. This likely hampered the quality of teaching and learning.

Cluster 4 had the highest School Participation Rate, Net Enrollment Rate, and Gross Enrollment Rate values, as well as the lowest poverty rate. These indicated excellent educational and socio-economic conditions. However, this cluster suffered from a high high school ratio (381.57), indicating a potential shortage of schools despite otherwise favorable conditions.

Cluster 5 had the lowest values across all educational indicators (School Participation Rate, Net Enrollment Rate, and Gross Enrollment Rate) and a high poverty rate of 28.63%, reflecting significant economic vulnerability. The high school and teacher ratios in this cluster were in the medium range. The low educational participation rates and high poverty suggested that provinces in Cluster 5 faced serious challenges in accessing education, likely due to limited access, inadequate facilities, lack of qualified educators, and economic hardship.

These findings aligned with several previous studies cited in the Introduction. Baharudin and Burhan (Nguyen, 2024), found that teachers in rural areas faced challenges in infrastructure and access to resources, yet remained adaptive to policy changes. This supported the present study's finding that provinces with low school participation tended to have low socio-economic indicators, suggesting that limited access to quality education was linked to regional disparities in resources and development. Similarly, Xu (2024) highlighted how digital transformation in rural China was hindered by low digital literacy and inadequate infrastructure. These challenges reflected similar conditions observed in Indonesian provinces with low socio-economic scores, where limited access to educational and technological facilities likely contributed to low participation rates. Furthermore, Batool (Batool & Liu, 2021) emphasized the influence of socio-economic variables on higher education participation in Pakistan, identifying government expenditure and unemployment as critical factors. This was consistent with the results of the present study, which showed that socio-economic indicators—particularly poverty and education spending—played a significant role in shaping school participation patterns. By clustering provinces based on these variables, the study offered a regional classification that extended previous regression-based findings and revealed broader patterns of educational inequality across Indonesia.

The distribution of provinces in Indonesia based on the optimal clustering result, which used five clusters, was shown in Figure 1.

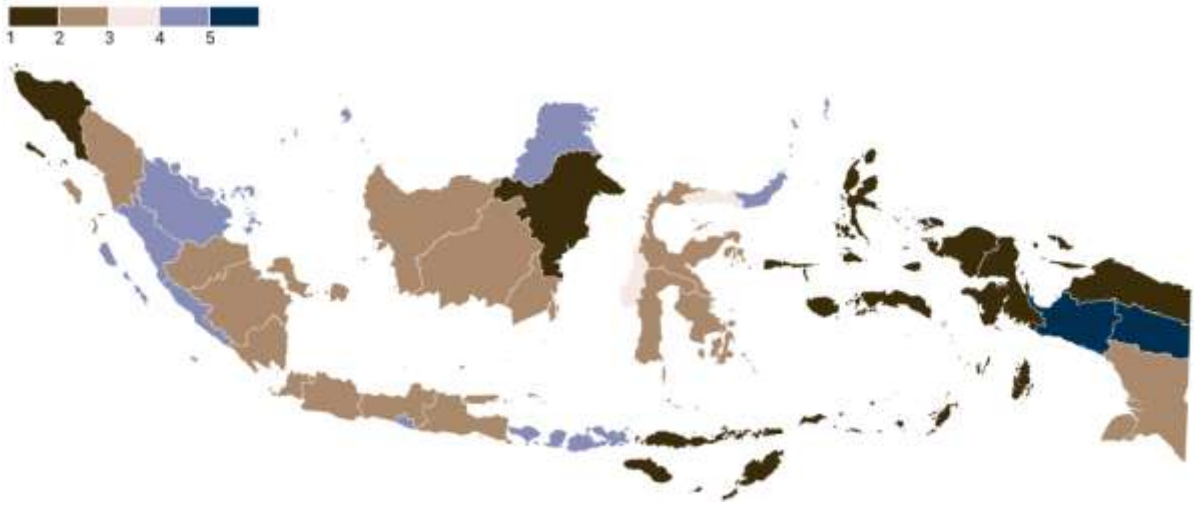

Figure 1. Cluster of provinces in Indonesia.

Figure 1 illustrates the distribution of provinces in Indonesia based on the clustering results using five clusters. Provinces in dark brown belonged to Cluster 1; brown represented Cluster 2; cream represented Cluster 3; purple represented Cluster 4; and dark blue represented Cluster 5. From the figure, it could be seen that Cluster 3 had the fewest members. Meanwhile, the cluster with the most members was Cluster 2, followed by Cluster 4, as shown on the map where provinces in these two clusters were spread across various regions in Indonesia.

5. CONCLUSION

The study found that the optimal clustering was achieved with five clusters using the K-Means method, supported by multiple validation indices (Dunn Index, C-Index, and Davies-Bouldin Index) and the smallest S(w)/S(b) ratio. This addressed the research gap in systematically grouping regions by educational and socio-economic indicators. Cluster 1 was characterized by good senior secondary participation and moderate poverty but low school and teacher ratios. Cluster 2 showed moderate education access, low poverty, but high student-to-school ratios. Cluster 3 combined moderate education access with high poverty and very high student-to-school and student-to-teacher ratios. Cluster 4 reflected strong education and economic conditions, although still school shortages persisted. Finally, Cluster 5 represented the most disadvantaged group, with the lowest educational indicators and highest poverty. These findings highlighted clear disparities, from provinces with strong participation and low poverty to those with severe educational and economic disadvantages. By revealing these hidden regional patterns, the study proposed a clustering-based framework to better target educational interventions, revealing regional groupings that can inform equitable, geographically sensitive policies in Indonesia. Future research was recommended to investigate contextual factors within clusters, evaluate educational quality beyond quantitative ratios, and conduct focused studies on disadvantaged regions to uncover structural barriers and guide targeted interventions.

REFERENCE

Abou-Moustafa, K. (2016). What is the distance between objects in a data set? A brief review of distance and similarity measures for data analysis. *IEEE Pulse, 7*(2), 41–47. https://doi.org/10.1109/MPUL.2015.2513727

Badan Pusat Statistik. (2024). *Indonesia statistics 2024* (Vol. 52). Statistics Indonesia. https://www.bps.go.id/id/publication/2024/02/28/c1bacde03256343b2bf769b0/statistik-indonesia-2024.html

Baharuddin, & Burhan. (2025). Urban and rural teacher perspectives on Indonesian educational reform: Challenges and policy implications. *Cogent Education, 12*(1), 2497142. https://doi.org/10.1080/2331186X.2025.2497142

Batool, S. M., & Liu, Z. (2021). Exploring the relationships between socioeconomic indicators and student enrollment in higher education institutions of Pakistan. *PLOS ONE, 16*(12), e0261577. https://doi.org/10.1371/journal.pone.0261577

Chigbu, B. I., & Nekhwevha, F. H. (2021). High school training outcome and academic performance of first-year tertiary institution learners: Taking "Input-Environment-Outcomes model" into account. *Heliyon,* 7(7), e07700. https://doi.org/10.1016/j.heliyon.2021.e07700

Chong, B. (2021). K-means clustering algorithm: A brief review. *Academic Journal of Computer and Information Sciences,* 4(5), 37–40. https://doi.org/10.25236/ajcis.2021.040506

Fernandes, A. A. R., Solimun, Efendi, E. C. L., Badung, N. M. A. A., & Krisnawati, E. (2022). Cluster analysis study on various cluster validity indexes with various linkages and Euclidean distance (Study on compliant paying behavior of Bank X customers in Indonesia 2021). *Journal of Statistics Applications & Probability, 11*(3), 875–882. https://doi.org/10.18576/jsap/110311

Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Understanding the role of digital technologies in education: A review. *Sustainable Operations and Computers, 3,* 275–285. https://doi.org/10.1016/j.susoc.2022.05.004

Jie, C., Jiyue, Z., Junhui, W., Yusheng, W., Huiping, S., & Kaiyan, L. (2020). Review on the research of K-means clustering algorithm in big data. In *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)* (pp. 107–111). IEEE. https://doi.org/10.1109/ICECE51594.2020.9353036

Muhajir, M., & Sari, N. N. (2018). K-Affinity Propagation (K-AP) and K-means clustering for classification of earthquakes in Indonesia. In *2018 International Symposium on Advanced Intelligent Informatics (SAIN)* (pp. 6–10). IEEE. https://doi.org/10.1109/SAIN.2018.8673344

Munna, A. S., & Kalam, M. A. (2021). Teaching and learning process to enhance teaching effectiveness: Literature review. *International Journal of Humanities Innovation, 4*(1), 1–4. https://doi.org/10.33750/ijhi.v4i1.102

Nguyen, H. T. M., Bui, N. A., Ngo, N. T. H., & Luong, T. Q. (2024). Surviving and thriving: Voices from teachers in remote and disadvantaged regions of Vietnam. *Asia Pacific Journal of Education, 00*(00), 1–16. https://doi.org/10.1080/02188791.2024.2336246

Novianti, P., Setyorini, D., & Rafflesia, U. (2017). K-means cluster analysis in earthquake epicenter clustering. *International Journal of Advanced Intelligent Informatics, 3*(2), 81–89. https://doi.org/10.26555/ijain.v3i2.100

Schröder, S. M., & Kiko, R. (2022). Assessing representation learning and clustering algorithms for computer-assisted image annotation: Simulating and benchmarking MorphoCluster. *Sensors, 22*(7), 2775. https://doi.org/10.3390/s22072775

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8,* 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Suraya, S., Sholeh, M., & Lestari, U. (2023). Evaluation of data clustering accuracy using K-Means algorithm. *International Journal of Multidisciplinary Approach in Research and Science, 2*(1), 385–396. https://doi.org/10.59653/ijmars.v2i01.504

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability, 14*(12), 7243. https://doi.org/10.3390/su14127243

Turienzo, J. (2024). A transversal and practical education as a business success factor: Literature review of learning process of basic design through ICT tools. *Journal of Management and Business Education,* 7(1), 70–89. https://doi.org/10.35564/jmbe.2024.0005

Xu, Q. (2024). The impact of new media technology applications on educational equity in rural areas. *Education Journal, 13*(5), 284–293. https://doi.org/10.11648/j.edu.20241305.15