Current Science Research Bulletin
Volume 03 (05), May 2026, pp. 135-142
CrossRef DOI: 10.55677/csrb/06-V03I05Y2026
Available on: https://csrbjournal.org/
e-ISSN: 3050-7103
p-ISSN: 3050-709X

Clustering Indonesian Provinces Based on Sustainable Development Goals Welfare Indicators Using Partitioning Around Medoids

Nabilla Aulia Jenny Dewi Rahmawati[1], Yuciana Wilandari[2], Diah Safitri*[3]
[1,2,3] Department of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro

ABSTRACT: This study aims to classify interprovincial welfare disparities in Indonesia based on selected Sustainable Development Goals (SDGs) indicators and to identify groups of provinces with similar welfare characteristics. A quantitative approach employing cluster analysis was applied to secondary data from 38 provinces in 2024, using eight welfare-related SDG indicators covering food security, health, education, and access to basic services. These indicators were selected to represent key dimensions of human well-being and to reflect multidimensional aspects of regional development. Before analysis, the data were standardized using the z-score method to ensure comparability across indicators with different measurement scales and to prevent dominance by variables with larger variances. The Partitioning Around Medoids (PAM) algorithm was employed due to its robustness to outliers, as it determines cluster centers based on actual observations. The optimal number of clusters was identified using the Gap Statistic method. The results indicate that the optimal solution is achieved at $\boldsymbol{K = 2}$, with a Gap Statistic value of $\mathbf{0.55}$, yielding a partition of provinces into two distinct clusters. Cluster 1 represents provinces with relatively better welfare conditions across the selected indicators. In contrast, Cluster 2 consists of provinces with lower welfare performance and higher vulnerability, with outliers effectively accommodated within the clustering structure. These findings highlight substantial disparities in welfare across provinces, underscoring the need for differentiated policy responses tailored to regional characteristics and development priorities. The results also provide policymakers with an empirical basis for designing more targeted, evidence-based interventions to support the achievement of the SDGs at both national and subnational levels. Nevertheless, further research is recommended to incorporate additional indicators, explore alternative clustering techniques, and examine temporal dynamics to more comprehensively capture changes in welfare.

KEY WORDS: Sustainable Development Goals, Welfare, Clustering, Partitioning Around Medoids, Gap Statistic

1. INTRODUCTION

Sustainable development has become a central global agenda since the adoption of the Sustainable Development Goals (SDGs) in 2015, which continued the Millennium Development Goals (MDGs). The SDGs comprise 17 goals, 169 targets, and more than 230 indicators, emphasizing a balanced integration of economic, social, and environmental dimensions under the principle of leaving no one behind (Putra *et al.*, 2024). In Indonesia, the SDGs serve as a strategic framework for evaluating national development performance. However, their implementation at the subnational level remains constrained by disparities in regional resources, infrastructure, and socioeconomic conditions, which contribute to uneven development outcomes across provinces. Interprovincial welfare inequality continues to represent a critical development challenge in Indonesia, particularly across key dimensions such as food security, healthcare access, educational attainment, and access to safe drinking water and sanitation (BPS, 2024a). For instance, in 2024, access to safe drinking water in Highland Papua was recorded at 30.64%, which is considerably lower than in DKI Jakarta (Special Capital Region of Jakarta) at 99.96% and the national average of 92.64% (BPS, 2024b). Disparities in education are also evident, as reflected in the proportion of out-of-school children, which increases from 0.67% at the primary level to 21.61% at the senior secondary level (BPS, 2023). These conditions indicate that uniform development approaches are insufficient to capture the multidimensional nature of welfare disparities, necessitating more comprehensive, data-driven analytical approaches to support effective policy formulation.

Cite the Article: *Jenny Dewi Rahmawati, N.A., Wilandari, Y., Safitri, D. (2026). Clustering Indonesian Provinces Based on Sustainable Development Goals Welfare Indicators Using Partitioning Around Medoids. Current Science Research Bulletin, 3(5), 135-142. https://doi.org/10.55677/csrb/06-V03I05Y2026*

Publication Date: *May 27, 2026*

Corresponding Author: Diah Safitri

Cluster analysis provides a relevant multivariate approach for identifying patterns of regional inequality by grouping provinces based on similarities in their characteristics (Hair *et al*., 2019). Among non-hierarchical clustering methods, Partitioning Around Medoids (PAM) is recognized for its robustness to outliers, as it determines cluster centers using actual observations (medoids), leading to more stable and interpretable groupings (Kaufman and Rousseeuw, 1990). The optimal number of clusters is determined using the Gap Statistic, which enables a formal comparison between the observed within-cluster dispersion and that expected under a null reference distribution (Tibshirani *et al*., 2001). Previous studies have demonstrated the effectiveness of the K-Medoids approach across various fields, including education (Fialine *et al*., 2021), marketing (Rindiawan *et al*., 2025), and criminology (Dyaherawati *et al*., 2025), particularly when combined with validation techniques such as the Gap Statistic. However, its application to SDG indicators, especially in the context of provincial welfare assessment in Indonesia, remains limited. Therefore, this study aims to classify Indonesian provinces based on welfare-related SDG indicators using the PAM method with Gap Statistic validation, to provide a more robust empirical basis for targeted, policy-relevant development interventions.

II. METHOD

This section describes the data sources, variables, and analytical methods employed in this study. The analysis consists of several stages, including data preparation, assumption testing, clustering using the Partitioning Around Medoids (PAM) method, determining the optimal number of clusters using the Gap Statistic, and cluster profiling. Each stage is designed to ensure the robustness and interpretability of the clustering results in identifying welfare disparities across provinces in Indonesia.

Data and Variables

This study uses secondary data comprising eight welfare-related Sustainable Development Goals (SDGs) indicators for 2024, obtained from the publication *SDGs Welfare Indicators 2024* issued by Statistics Indonesia (BPS). The unit of analysis comprises 38 provinces in Indonesia. All variables are expressed as percentages to ensure comparability across indicators.

The selected indicators represent four dimensions of welfare, namely food security, health, education, and access to basic services. The food security dimension includes the prevalence of inadequate food consumption (X_1) and food insecurity (X_2). The health dimension is measured by the proportion of births assisted by health personnel (X_3) and unmet need for healthcare services (X_4). The education dimension is represented by the proportion of out-of-school children at the junior secondary level (X_5) and the ratio of gross enrollment rates between the lowest and highest quintiles at the senior secondary level (X_6). Access to basic services is measured by the percentage of households with access to safe drinking water (X_7) and adequate sanitation (X_8). Together, these indicators capture the multidimensional nature of welfare disparities across provinces and provide a comprehensive basis for clustering analysis.

Data Preprocessing

Data preprocessing is conducted to ensure comparability across variables and to enhance the reliability of the clustering results. Descriptive statistics are first employed to explore the general patterns, central tendencies, and variability of the data, thereby providing an initial understanding of the dataset (Walpole, 1995). Univariate outliers are subsequently identified using boxplots based on the interquartile range (IQR), a method that does not rely on distributional assumptions and is robust to variations in data distribution (Tukey, 1977).

All variables are then standardized using z-score transformation to eliminate differences in measurement scales and to ensure equal contribution of each variable in distance calculations (Hair *et al*., 2019). The transformation is defined as:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (1)$$

where z_{ij} denotes the standardized value of observation i on variable j, x_{ij} is the observed value, $\bar{x}_j$ represents the mean of variable j, and s_j is the standard deviation of variable j.

Assumption Testing

The adequacy of the data for cluster analysis is evaluated using the Kaiser-Meyer-Olkin (KMO) measure, which assesses sampling adequacy based on the correlation structure among variables (Yamin and Kurniawan, 2014). The KMO statistic is defined as:

$$KMO = \frac{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} r_{x_j x_k}^2}{\sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} r_{x_j x_k}^2 + \sum_{j=1}^{p}\sum_{k=1,k\neq j}^{p} \rho_{x_j x_k \cdot x_l}^2} \quad (2)$$

where p denotes the number of variables, $r_{x_j x_k}$ represents the Pearson correlation coefficient between variables x_j and x_k, and $\rho_{x_j x_k \cdot x_l}$ denotes the partial correlation coefficient between x_j and x_k controlling for the remaining variables. A KMO value greater than or equal to 0.5 indicates that the data are adequate for cluster analysis.

Corresponding Author: Diah Safitri

Multicollinearity among variables is examined using the Variance Inflation Factor (VIF), which measures the extent of linear dependency among predictors (Gujarati and Porter, 2009). The VIF is defined as:

$$VIF_j = \frac{1}{1-R_j^2} \quad (3)$$

where VIF_j is the variance inflation factor for variable j, and R_j^2 is the coefficient of determination obtained by regressing variable x_j on the remaining variables. A VIF value greater than or equal to 10 indicates a high degree of multicollinearity.

Variables exhibiting high multicollinearity are further evaluated using Principal Component Analysis (PCA) to reduce dimensionality and produce uncorrelated components, thereby improving the stability and reliability of the clustering structure (Johnson and Wichern, 2007).

Partitioning Around Medoids (PAM)

Cluster analysis is performed using the Partitioning Around Medoids (PAM) method, a non-hierarchical clustering approach that defines cluster centers as medoids (hereafter denoted as 'medoids'), meaning actual observations. This method is known for its robustness to outliers and for producing more stable clusters than centroid-based techniques (Kaufman and Rousseeuw, 1990).

The dissimilarity between observations is measured using Euclidean distance:

$$d_{ic} = \sqrt{\sum_{j=1}^{p}\left(x_{ij} - m_{cj}\right)^2} \quad (4)$$

where d_{ic} denotes the Euclidean distance between observation i and medoid c, x_{ij} is the value of observation i on variable j, m_{cj} represents the value of medoid c on variable j, and p is the number of variables.

The PAM algorithm begins by selecting K initial medoids from the set of observations. Each observation is then assigned to the medoid with the minimum distance, forming an initial cluster partition. The quality of the clustering is evaluated by the total within-cluster dissimilarity, defined as the sum of distances between observations and their corresponding medoids. Subsequently, candidate medoids are generated by exchanging current medoids with non-medoid observations, and the resulting partition is re-evaluated. This process is repeated iteratively, and a new set of medoids is retained if it reduces the total within-cluster dissimilarity. The algorithm terminates when no further improvement is achieved, indicating that a stable clustering structure has been reached.

Determination of the Optimal Number of Clusters

The optimal number of clusters is determined using the Gap Statistic (Tibshirani *et al*., 2001), which evaluates clustering performance by comparing within-cluster dispersion in the observed data with that expected under a reference (null) distribution. The within-cluster dispersion is first defined at the cluster level as the sum of squared Euclidean distances between observations and their corresponding medoid.

$$D_c = \sum_{i \in C_c} d^2_{ic} \quad (5)$$

where D_c denotes the total within-cluster dispersion for cluster c, C_c represents the set of observations in cluster c, and d_{ic} is the Euclidean distance between observation i and the medoid of cluster c.

The total within-cluster dispersion for K clusters is then defined as:

$$W_K = \sum_{c=1}^{K} \frac{1}{2n_c} D_c \quad (6)$$

where W_K denotes the overall within-cluster dispersion, and n_c is the number of observations in cluster c.

The Gap Statistic is then defined as:

$$Gap(K) = E^*\{log(W_K)\} - log(W_K) \quad (7)$$

where $E^*[\log(W_K)]$ is the expected value of $\log(W_K)$ obtained from B Monte Carlo simulations under a reference distribution, and W_K is the within-cluster dispersion for the observed data.

To account for simulation variability, the standard error is computed as:

$$s_K = \sqrt{1 + \frac{1}{B}}\, sd(K) \quad (8)$$

where $\text{sd}(K)$ is the standard deviation of $\log(W_K^*)$ across the simulated datasets.

The optimal number of clusters is selected as the smallest value of K satisfying the following criterion:

$$Gap(K) \geq Gap(K+1) - s_{K+1} \quad (9)$$

This criterion indicates that increasing the number of clusters beyond K does not yield a statistically meaningful improvement in clustering quality. In other words, the additional clusters fail to provide substantial new information or enhance the separation

Corresponding Author: Diah Safitri

between groups. Therefore, the selected value of K is considered optimal, as it best represents the data's underlying structure while balancing model simplicity and interpretability.

Cluster Profiling

Cluster profiling involves systematically interpreting the characteristics of each group by examining mean values and identifying the dominant patterns of variables within each cluster (Simamora, 2005). Through this approach, the distinctive features of each cluster can be more clearly understood. Comparing profiles across clusters helps reveal disparities in welfare levels across regions and provides a stronger analytical basis for developing evidence-based, region-specific policy recommendations aligned with each group's unique characteristics.

III. RESULTS

The analysis begins with an exploratory assessment to provide an overview of the eight SDG-related welfare indicators across 38 provinces. Univariate outliers are identified using boxplots based on the interquartile range (IQR), as illustrated in Figure 1.

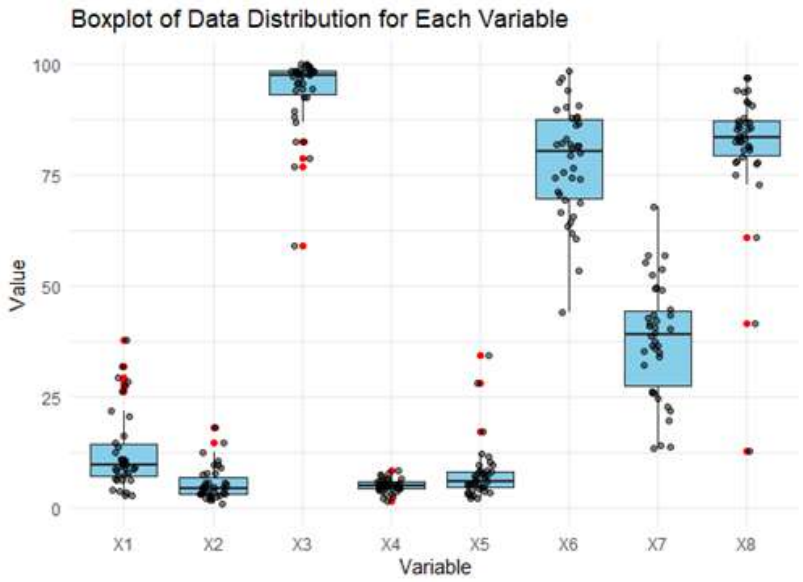

Figure 1. Boxplot of Research Variables

As shown in Figure 1, several variables exhibit outliers. These observations are retained, as they reflect actual disparities across provinces. Since all variables are expressed as percentages but differ in scale, z-score standardization is applied to ensure proportional contribution in distance calculations prior to clustering.

The adequacy of the data is then evaluated through assumption testing. The Kaiser-Meyer-Olkin (KMO) measure yields a value of 0.73, indicating that the data are sufficiently suitable for further multivariate analysis. All variables meet the minimum sampling adequacy requirements, suggesting that none need to be excluded. The multicollinearity assessment indicates that all variables meet the acceptable threshold, suggesting that redundancy among variables does not pose a significant issue for clustering.

Cluster analysis is subsequently performed using the Partitioning Around Medoids (PAM) method on standardized data. The number of clusters is initially evaluated over $K = 2$ to $K = 6$ to identify the most representative grouping structure. The results show a decreasing trend in the objective function value as the number of clusters increases, indicating improved within-cluster homogeneity.

Table 1. PAM Objective Function Values (Swap Stage)

Number of Clusters (K)	2	3	4	5	6
Objective Function	2.10	1.84	1.61	1.47	1.34

Table 1 shows that the objective function decreases as the number of clusters increases. However, this reduction alone is not sufficient to determine the optimal number of clusters, as the objective function generally decreases with increasing K.

Corresponding Author: Diah Safitri

To determine the optimal number of clusters, the Gap Statistic method is applied with $B = 100$ bootstrap replications. The evaluation results are presented in Table 2.

Table 2. Gap Statistic Results ($B = 100$)

K	$\log(W_K)$	$E^*(\log W_K)$	$Gap(K)$	s_K
2	3.24	3.79	0.55	0.03
3	3.13	3.71	0.58	0.04
4	2.98	3.63	0.65	0.04
5	2.88	3.56	0.68	0.04
6	2.81	3.5	0.69	0.04

As shown in Table 2, the selection criterion is first satisfied at $K = 2$, indicating that two clusters provide the most appropriate representation of the data structure. This finding is further supported by the Gap Statistic plot in Figure 2.

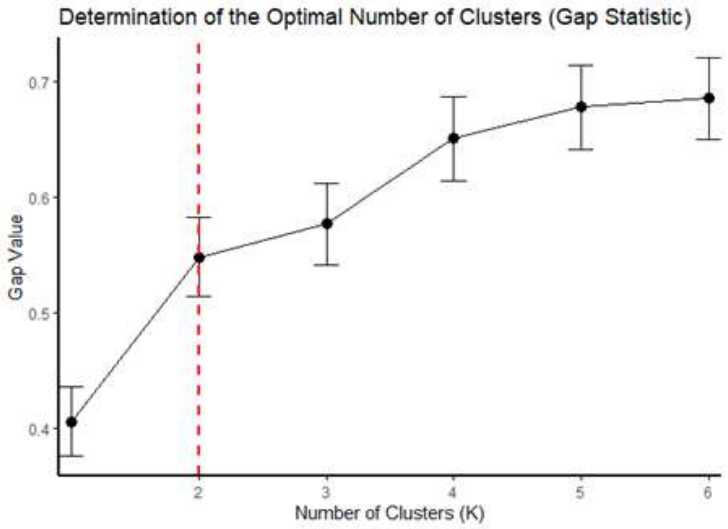

Figure 2. Gap Statistic Plot ($B = 100$)

As shown in Figure 2, the Gap Statistic value increases sharply at $K = 2$ and exhibits a more gradual trend for higher values of K. This pattern indicates that the most substantial improvement in clustering structure occurs at $K = 2$, after which additional clusters provide only marginal gains. This result supports the selection of $K = 2$ as the optimal number of clusters.

Based on this optimal solution, the provinces are grouped into two distinct clusters. A summary of the clustering results is presented in Table 3.

Table 3. Summary of PAM Clustering Results ($K = 2$)

Cluster	Medoid Province	Number of Members
Cluster 1	South Sulawesi	29
Cluster 2	North Maluku	9

Based on Table 3, Cluster 1 contains the majority of provinces, whereas Cluster 2 consists of a smaller group with distinct characteristics.

The clustering results are first visualized using a cluster plot to illustrate the distribution of provinces across clusters, as shown in Figure 3.

Corresponding Author: Diah Safitri

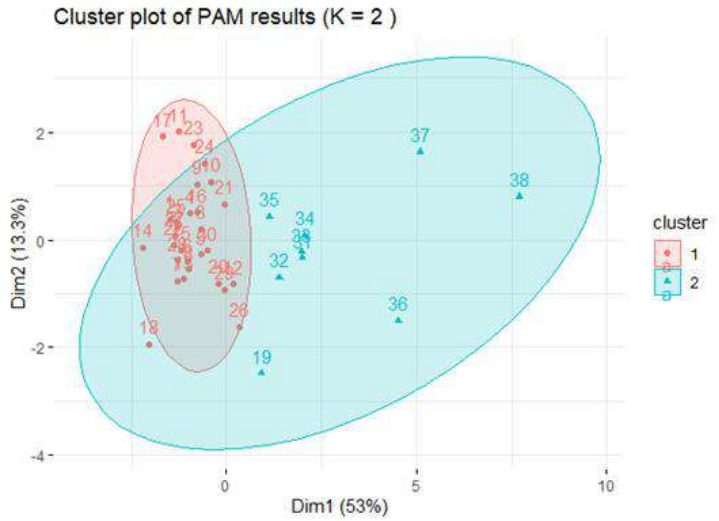

Figure 3. Cluster Plot of PAM Results ($K = 2$)

The spatial distribution of the clustering results is further presented using a thematic map, as shown in Figure 4.

Figure 4. Thematic Map of Clustering Results ($K = 2$)

To further examine the characteristics of each cluster, the mean values of the variables are calculated and presented in Table 4.

Table 4. Mean Values of Variables by Cluster

Variable	Cluster 1 (South Sulawesi)	Cluster 2 (North Maluku)
Prevalence of inadequate food consumption (X_1)	8.35	26.10
Prevalence of food insecurity (X_2)	3.67	10.80
Births assisted by health personnel (X_3)	97.60	81.80
Unmet need for healthcare services (X_4)	5.22	3.76
Proportion of out-of-school children at the junior secondary level (X_5)	6.18	12.20
Ratio of gross enrollment rates between the lowest and highest quintiles at the senior secondary level (X_6)	81.30	67.00
Households with access to safe drinking water (X_7)	39.60	32.10
Households with access to adequate sanitation (X_8)	85.90	65.90

Table 4 shows that Cluster 1 is characterized by relatively better welfare conditions across most indicators, whereas Cluster 2 exhibits higher vulnerability in several dimensions.

Corresponding Author: Diah Safitri

IV. DISCUSSION

The clustering results reveal a clear and systematic pattern of welfare disparities across provinces in Indonesia, reflecting persistent inequalities in SDG-related development outcomes. Cluster 1 represents provinces with relatively better welfare conditions, characterized by lower levels of food insecurity, higher access to healthcare services, better educational participation, and more adequate access to sanitation and basic services. Despite these relatively favorable conditions, several development gaps remain, particularly in improving access to safe drinking water and ensuring more equitable distribution of healthcare services across regions.

In contrast, Cluster 2 reflects provinces with higher vulnerability across multiple welfare dimensions, particularly in food security, education, and access to basic services. These provinces exhibit higher levels of food insecurity, lower educational participation, and more limited access to essential services, including healthcare, safe drinking water, and sanitation. Such conditions indicate persistent structural constraints on development that may hinder progress toward achieving key Sustainable Development Goals, particularly those related to poverty reduction, health, education, and clean water and sanitation.

The observed disparity between clusters underscores the importance of adopting region-specific, needs-based development strategies rather than uniform policy approaches. From a policy perspective, provinces in Cluster 1 may require strategies focused on improving service quality, efficiency, and sustainability, whereas those in Cluster 2 require priority interventions to expand basic service coverage, strengthen food security systems, and improve access to education, particularly for vulnerable populations. These findings underscore the need for more targeted, evidence-based policy design to reduce interprovincial inequality and accelerate the achievement of the SDGs in Indonesia.

V. CONCLUSION

The clustering results of Indonesian provinces using the Partitioning Around Medoids (PAM) method, validated by the Gap Statistic, indicate that the welfare structure is optimally divided into two distinct clusters. This finding reveals a clear, consistent pattern of disparities in welfare conditions across provinces, as measured by SDG-related indicators.

Cluster 1 comprises provinces with relatively better performance in food security, healthcare services, education, and sanitation access, although challenges remain in improving access to safe drinking water and ensuring the equitable distribution of health services. In contrast, Cluster 2 comprises provinces with higher vulnerability in food security, educational participation, and access to basic services, underscoring the need for more targeted, inclusive, and integrated policy interventions. Overall, these results highlight the importance of region-specific strategies to reduce interprovincial disparities and to accelerate the achievement of welfare-related SDG targets in Indonesia.

VI. ACKNOWLEDGEMENTS

The authors express sincere gratitude to all who contributed to this study. Special thanks go to Statistics Indonesia (BPS) of Indonesia for providing access to the secondary data. Appreciation is also given to academic supervisors and colleagues for their guidance, feedback, and insightful discussions. Their support greatly improved the quality and rigor of this study.

VII. DISCLOSURE

The authors declare that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Badan Pusat Statistik. (2023). *Angka anak tidak sekolah menurut jenjang pendidikan dan jenis kelamin, 2023*. Jakarta: Badan Pusat Statistik. https://www.bps.go.id/id/statistics-table/2/MTk4NiMy/angka-anak-tidak-sekolah-menurut-jenjang-pendidikan-dan-jenis-kelamin.html.
2. Badan Pusat Statistik. (2024a). *Indikator SDGs kesejahteraan rakyat 2024*. Jakarta: Badan Pusat Statistik.
3. Badan Pusat Statistik. (2024b). *Persentase rumah tangga yang memiliki akses terhadap sumber air minum layak menurut provinsi (persen), 2024*. Jakarta: Badan Pusat Statistik. https://www.bps.go.id/id/statistics-table/2/ODQ1IzI=/persentase-rumah-tangga-yang-memiliki-akses-terhadap-sumber-air-minum-layak-menurut-provinsi.html.
4. Dyaherawati, O., Martha, S., & Imro'ah, N. (2025). Penerapan algoritma K-medoids dengan optimasi gap statistics dalam pengelompokan daerah rawan kriminalitas di Indonesia. *Buletin Ilmiah Matematika, Statistika dan Terapannya*, 14(1), 103-112.
5. Fialine, A. P., Alodia, D. A., Endriani, D., & Widodo, E. (2021). Implementasi metode K-medoids clustering untuk pengelompokan provinsi di Indonesia berdasarkan indikator pendidikan. *Journal of Mathematics Education and Applied*, 2(2), 1-13.
6. Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.

Corresponding Author: Diah Safitri

7. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning EMEA.
8. Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall.
9. Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
10. Putra, A. A., Hasibuan, H. S., Tambunan, R. P., & Lautetu, L. M. (2024). Integration of the Sustainable Development Goals into a regional development plan in Indonesia. *Sustainability, 16*(23), 1-21.
11. Rindiawan, S. Z., Rachman, A. N., & Purwayoga, V. (2025). Optimasi jumlah cluster untuk analisis penjualan barang kosmetik menggunakan K-medoids. *Jurnal Sistem dan Teknologi Informasi, 13*(1), 148-165.
12. Simamora, B. (2005). *Analisis multivariat pemasaran*. Gramedia Pustaka Utama.
13. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411-423.
14. Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
15. Walpole, R. E. (1995). *Pengantar statistika* (Edisi ke-3). Gramedia Pustaka Utama.
16. Yamin, S., & Kurniawan, H. (2014). *SPSS complete: Teknik analisis statistik terlengkap dengan software SPSS*. Salemba Infotek.

Corresponding Author: Diah Safitri