A Hybrid Machine Learning Model for Market Clustering

Rendra Gustriansyah
Department of Informatics Engineering, Faculty of Computer and Natural Science, Universitas Indo Global Mandiri, Indonesia
rendra@uigm.ac.id (corresponding author)

Juhaini Alie
Department of Management, Faculty of Economics, Universitas Indo Global Mandiri, Indonesia
juhaini@uigm.ac.id

Nazori Suhandi
Department of Informatics Engineering, Faculty of Computer and Natural Science, Universitas Indo Global Mandiri, Indonesia
nazori@uigm.ac.id

Received: 13 October 2024 | Revised: 31 October 2024 | Accepted: 3 November 2024

 DOI: https://doi.org/10.48084/etasr.9259

ABSTRACT

Market clustering is increasingly important for companies to understand consumer shopping behavior in the context of complex data. This study aims to develop a hybrid model that integrates Principal Component Analysis (PCA) and k-medoids to enhance market clustering based on consumer shopping patterns. The methods used include data preprocessing, PCA application for dimensionality reduction, and clustering using k-medoids. The quality of the clusters is evaluated with various validity indices. The results show that the hybrid model produces clusters with better quality compared to the single k-medoids method, as seen from the Calinski-Harabasz Index (CHI), theSilhouette Width (SW), and the Davies-Bouldin (DB) index. The implications of these findings emphasize the importance of adopting hybrid methods in marketing strategies to improve understanding of consumer behavior dynamics and allow companies to adjust their marketing strategies more effectively. This study provides a strong foundation for further development in clustering analysis across various industry sectors and highlights the potential for innovative techniques to address dynamic market challenges.

Keywords-market clustering; principal component analysis; k-medoids; dimensionality reduction; consumer behavior

I. INTRODUCTION

Market clustering is crucial for companies to identify correlations between specific market clusters and customer shopping patterns and preferences. It allows companies to customize their strategies, messages, promotions, and special offers for these market clusters [1]. With the increasing complexity of market data, hybrid clustering methods have emerged as an innovative approach to enhance the effectiveness of market clustering. Hybrid methods combine the strengths of different clustering methods to overcome the shortcomings of each method. Case studies across various industries, such as marketing and consumer behavior analysis, have demonstrated that hybrid methods not only enhance clustering results but also offer deeper insights for strategic decision-making [2]. Therefore, adopting hybrid methods in market segmentation is becoming increasingly important to address the challenges posed by dynamic and diverse data. Hybrid clustering methods have the potential to enhance the accuracy and robustness of market clustering. However, their implementation often encounters several challenges. Integrating multiple clustering methods requires careful alignment of parameters and algorithm structure. The main difficulty lies in determining an effective way to combine each method's results, including how to incorporate clusters generated by different algorithms without compromising the quality or validity of the clustering results. Furthermore, the technical complexity of hybrid methods often necessitates more intricate parameter settings and a deeper understanding of the interactions between the methods used. Suboptimal implementation can lead to inconsistent results or degrade the overall model performance. Hence, this research aims to develop a hybrid model that integrates dimensionality reduction (Principal Component Analysis-PCA) and machine learning (k-medoids) methods to optimize market clustering

based on customer shopping patterns. The study focuses on how combining different methods can enhance market clustering performance in a dynamic and complex context, unlike previous studies that typically concentrate on a single clustering method or test hybrid method in a limited domain. The PCA method is used to identify patterns and reduce complexity in data without losing important information, aiming to improve clustering performance [3]. Meanwhile, the k-medoids is driven to enhance the classical clustering method, making the clustering results more proportional and optimal. The k-medoids method is selected because it can effectively cluster datasets containing outliers [4].

The rationale for this study is rooted in the pressing need for more adaptable and precise clustering methods given the increasing complexity of market data. Furthermore, there is a lack of understanding regarding the application of hybrid clustering methods. By addressing these gaps in the literature and proposing a new evidence-based approach, this study aims to significantly advance the effectiveness of hybrid clustering methods and establish a foundational literature for future research in market clustering. As a result, this study's contribution will enhance the theoretical aspects of market clustering and positively impact industry practices by offering a more accurate and adaptable approach to market clustering.

II. RELATED WORK

In the e-commerce business, market clustering is a dynamic and demanding area [5]. It drives researchers to specify opportune market clustering techniques [6]. Customer clustering generally involves using the Recency, Frequency, and Monetary (RFM) model to investigate past consumer buying behavior [7-9]. Authors in [10] emphasized that the RFM model can also induce crucial insights for market clustering. The RFM model can be used to develop effective marketing strategies [11] and can be effectively implemented in the sales sector [12, 13].

Classical market clustering methods using RFM models have limitations in processing real-time data, which can result in missing opportunities to identify patterns, automate processes, uncover correlations, and analyze market trends. By applying machine learning methods, businesses can adapt to evolving customer preferences and shopping patterns, ensuring that they can effectively reach the appropriate consumers with the right messages at the right time, thus saving time and resources [14]. Furthermore, combining dimensionality reduction with machine learning methods could lead to the development of diverse market clustering strategies [15-17].

Machine learning enables the real-time analysis of customer shopping history data, allowing businesses to observe patterns and changes in customer preferences [1, 18]. Authors in [19] utilized a priori and k-means methods to group potential buyers according to their characteristics and transaction patterns with specific merchants. Authors in [20] utilized several machine-learning techniques, including the Wald, fuzzy c-means, and k-means methods to cluster customers. This approach enables marketers to provide personalized experiences, including tailored messages, recommendations, and offers based on individual customer preferences, leading to increased customer loyalty, sales, and competitive advantage [21-23].

III. METHODS

The dataset used is a history of 556,121 sales transactions for the duration of one year. This study presumes that the hybrid machine learning methods can enhance the single clustering method and classical RFM analysis to produce more accurate and balanced market clustering. Therefore, a hybrid method that integrates PCA and k-medoids methods is expected to improve the scalability of real-time data processing, identifying patterns, correlations, and dynamics of changes in customer preferences that may be missed by classical method analysis. Figure 1 illustrates the working procedure of the proposed market clustering model to identify shopping patterns and customer preferences. This study utilizes R-Studio as a tool for visualization and data processing.

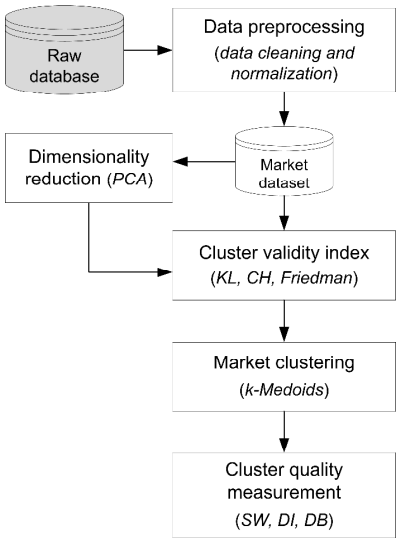

Fig. 1. Working procedure of the proposed method.

A. Data Preprocessing

Data preprocessing involves transforming the raw database into a market dataset that is ready to use. This stage includes cleaning incomplete data, outliers, and duplications as a first step. The second step involves normalizing the data using z-score and natural logarithm (log base 10) [13].

B. Dimensionality Reduction

Dimensionality reduction is a technique used to discover patterns in data. It can simplify without losing important information and enhance clustering performance. This study used the PCA method for data dimensionality reduction. The factoextra package in R programming [24] is implemented as a tool for PCA.

C. Cluster Validity Index

This step is needed to determine the best number of market clustering. To do this, three validity indices: Ratkowsky-Lance Index (RLI), Krzanowski-Lai Index (KLI), and Calinski-Harabasz Index (CHI) are implemented, which are available in

the NbClust package for R programming [25]. The accumulated index values will be used for market clustering.

D. Market Clustering

The next stage involves clustering the market dataset and PCA's result using the k-medoids method [26]. The k-medoids method utilizes the factoextra package in R programming. This process will generate market clusters, each containing a specific number of customers.

E. Cluster Quality Measurement

The quality of the market clusters was assessed using internal validation measures including Silhouette Width (SW), Dunn Index (DI), and Davies-Bouldin (DB) index. The higher SW and DI values indicate better-quality clusters [13]. Likewise, the smaller DB value indicates better-quality clusters. SW for each object i is determined by (1) [27]:

$$SW(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (1)$$

where $a(i)$ represents the average distance between object i and all other objects within the same cluster and $b(i)$ represents the average distance between object i and all objects in different nearest clusters. DI is determined by [28]:

$$DI = \frac{min_{i\neq j}\, d(C_i,C_j)}{max_k\, d(C_k)} \quad (2)$$

where C_i and C_j are different clusters, $d(C_i,C_j)$ represents the distance between clusters C_i and C_j, and $d(C_k)$ is the maximum distance between points in cluster C_k. The DB is determined by [29]:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j\neq i}\left(\frac{s_i+s_j}{d(C_i+C_j)}\right) \quad (3)$$

where s_i is the internal distance of cluster C_i.

IV. RESULTS AND DISCUSSION

The raw dataset contains 556,121 sales transaction entries, comprising seven variables: OrderID, OrderDate, CustomerID, ProductID, Quantity, UnitPrice, and Profit. The dataset was cleaned of duplicates, incomplete, and inconsistent data in the data preprocessing step. In this step, 32 incomplete records were identified and removed. The dataset was then aggregated based on CustomerID and produced a new dataset with six variables: Customer_ID, Quantity, Recency, Frequency, Monetary, and Profit, totaling 56,114 data points. Additionally, two outliers were removed from the final dataset.

To address resource limitations, a random sample equal to 10% of the entire dataset was selected while ensuring that the dataset was representative of the market clustering analysis. As a result, the final dataset consisted of 5,612 data points. The dataset was then normalized using z-score and natural logarithm (log base 10) to simplify the calculations.

Furthermore, the dimensionality reduction step using PCA produces three dimensions that cover more than 80% of the proportion of variation and cumulative in the dataset, as depicted in Figure 2. The PCA results are illustrated in Figure 3 in a biplot form. Vectors further from the origin have a more significant influence on the dimension.

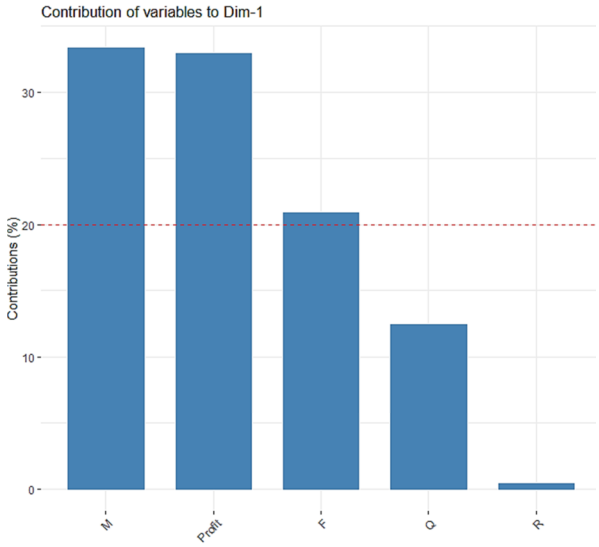

Fig. 2. Contribution of each variable.

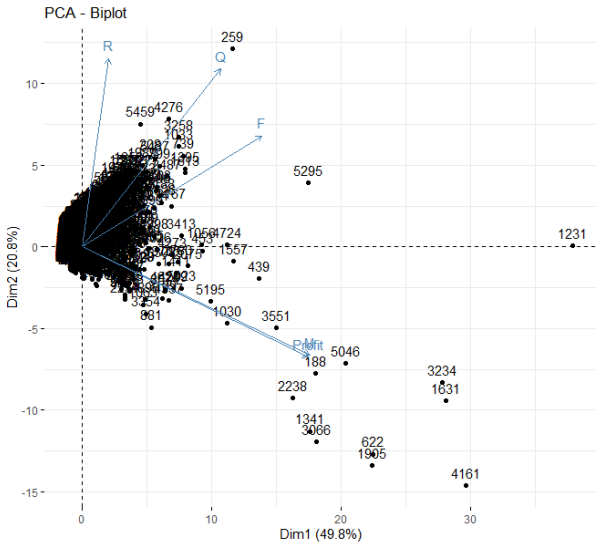

Fig. 3. Contribution of each variable.

Table I shows the results of validity indices using RLI, KLI, and CHI for both the k-medoid method and the PCA and k-medoid model. The best validity index for the k-medoid method is 3. Meanwhile, for the PCA and k-Medoid model, k = 4 is considered the best validity index.

TABLE I. VALIDITY INDICES

Validity Index	k-medoids	PCA and k-medoids
RLI	3	2
CHI	3	4
KLI	8	4

Furthermore, the k-medoids method and the PCA and k-medoids model were utilized for market clustering. The best validity index for each model was used. The market clustering results are visualized in Figures 4 and 5, and the cluster quality measurements using DB, DI, and SW are listed in Table II.

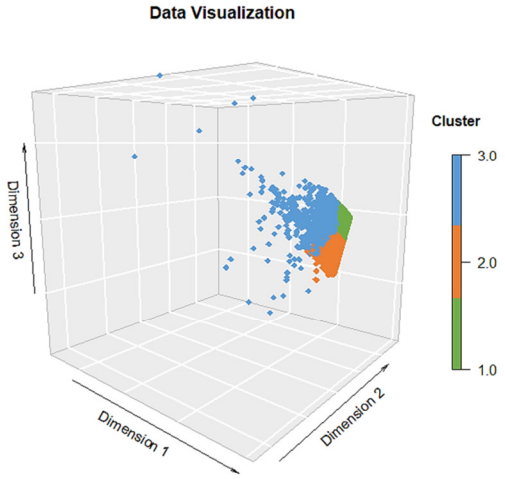

Fig. 4. Market clustering using the k-medoids method.

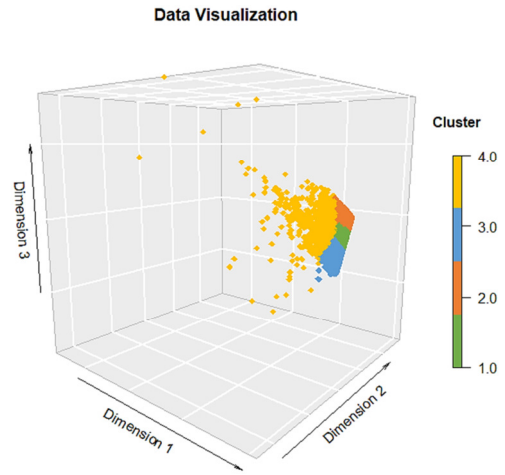

Fig. 5. Market clustering using PCA and k-medoids model.

TABLE II. CLUSTER QUALITY MEASUREMENT

Model	DB	DI	SW
k-medoids	416.6504	0.0004	0.3320
k-medoids and PCA	278.1917	0.0004	0.3654

The measurement results show that the DB, DI, and SW values for the PCA and k-medoids models are higher than the standalone k-medoids method. These results show that the intra-cluster and inter-cluster variances are high, showing that the PCA and k-medoids integration is a hybrid model that produces good market clustering.

The results from using the hybrid model demonstrate a significant improvement in clustering quality compared to the single k-medoids method. The enhanced validity of the indices in the hybrid model, particularly in CHI, indicates that incorporating PCA into the clustering process enables the identification of patterns and structures in more complex data. This is also evident from the increase in SW values and the decrease in DB, indicating that the resulting clusters are more distinct and exhibit higher internal similarities. Integrating PCA reduces the dimensionality of the data while retaining important information, aiding k-medoids in handling data with outliers and high complexity. Therefore, these results confirm that utilizing the hybrid model not only enhances the accuracy and robustness of clustering but also offers a deeper comprehension of consumer shopping behavior dynamics, which is highly valuable for strategic decision-making in a dynamic market.

V. CONCLUSION

This study presents a significant advancement in market clustering methodology by developing a hybrid model that effectively integrates PCA with the k-medoids method. By leveraging a dataset of over half a million sales transactions, the study successfully demonstrates that the proposed hybrid approach can enhance cluster quality and uncover more complex patterns and dynamics in the data compared to traditional k-medoids clustering. The improvement in the CHI and SW, along with a decrease in the DB index, confirms that integrating PCA reduces dimensionality without sacrificing important information effectively. The findings highlight the importance of using hybrid methods in clustering analysis, especially in dynamic and diverse markets, and enabling the formulation of appropriate marketing strategies.

Ultimately, the findings align with previous studies that support the hybrid method, emphasizing their relevance in addressing the growing challenges of data analysis in today's marketing landscape. This study paves the way for further studies on the hybrid approaches implementation in various industry sectors.

ACKNOWLEDGMENT

This study received support from a Fundamental Research Grant from DRTPM Ditjen Diktiristek of the Republic of Indonesia under grant number 104/E5/PG.02.00.PL/2024.

REFERENCES

[1] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, Jan. 2022, Art. no. 7243, https://doi.org/10.3390/su14127243.

[2] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," *Journal of Retailing and Consumer Services*, vol. 61, Jul. 2021, Art. no. 102566, https://doi.org/10.1016/j.jretconser.2021.102566.

[3] A. John, I. F. B. Isnin, S. H. H. Madni, and F. B. Muchtar, "Enhanced intrusion detection model based on principal component analysis and variable ensemble machine learning algorithm," *Intelligent Systems with Applications*, vol. 24, Dec. 2024, Art. no. 200442, https://doi.org/10.1016/j.iswa.2024.200442.

[4] P. D'Urso, M. Mucciardi, E. Otranto, and V. Vitale, "Community mobility in the European regions during COVID-19 pandemic: A partitioning around medoids with noise cluster based on space–time autoregressive models," *Spatial Statistics*, vol. 49, Jun. 2022, Art. no. 100531, https://doi.org/10.1016/j.spasta.2021.100531.

[5] T. Kim and J.-S. Lee, "Maximizing AUC to learn weighted naive Bayes for imbalanced data classification," *Expert Systems with Applications*, vol. 217, May 2023, Art. no. 119564, https://doi.org/10.1016/j.eswa.2023.119564.

[6] J. Salminen, M. Mustak, M. Sufyan, and B. J. Jansen, "How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation,"

Journal of Marketing Analytics, vol. 11, no. 4, pp. 677–692, Dec. 2023, https://doi.org/10.1057/s41270-023-00235-5.

[7] H. Abbasimehr and A. Bahrini, "An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation," *Expert Systems with Applications*, vol. 192, Apr. 2022, Art. no. 116373, https://doi.org/10.1016/j.eswa.2021.116373.

[8] A. Handojo, N. Pujawan, B. Santosa, and M. L. Singgih, "A multi layer recency frequency monetary method for customer priority segmentation in online transaction," *Cogent Engineering*, vol. 10, 2023, Art. no. 2162679, https://doi.org/10.1080/23311916.2022.2162679.

[9] S. Monalisa, Y. Juniarti, E. Saputra, F. Muttakin, and T. K. Ahsyar, "Customer segmentation with RFM models and demographic variable using DBSCAN algorithm," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 21, no. 4, pp. 742–749, Aug. 2023, https://doi.org/10.12928/telkomnika.v21i4.22759.

[10] Y. He and Y. Cheng, "Customer Segmentation and Management of Online Shops Based on RFM Model," in *International Conference on Application of Intelligent Systems in Multi-modal Information Analytics*, Huhehaote, China, Apr. 2021, pp. 34–41, https://doi.org/10.1007/978-3-030-51431-0_6.

[11] S. Monalisa, P. Nadya, and R. Novita, "Analysis for Customer Lifetime Value Categorization with RFM Model," *Procedia Computer Science*, vol. 161, pp. 834–840, Jan. 2019, https://doi.org/10.1016/j.procs.2019.11.190.

[12] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis Based on k-Means," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470–477, Apr. 2020, https://doi.org/10.11591/ijeecs.v18.i1.pp470-477.

[13] R. Gustriansyah, E. Ermatita, and D. P. Rini, "An approach for sales forecasting," *Expert Systems with Applications*, vol. 207, Nov. 2022, Art. no. 118043, https://doi.org/10.1016/j.eswa.2022.118043.

[14] S. Verma, R. Sharma, S. Deb, and D. Maitra, "Artificial intelligence in marketing: Systematic review and future research direction," *International Journal of Information Management Data Insights*, vol. 1, no. 1, Apr. 2021, Art. no. 100002, https://doi.org/10.1016/j.jjimei.2020.100002.

[15] Y. E. Touati, J. B. Slimane, and T. Saidani, "Adaptive Method for Feature Selection in the Machine Learning Context," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14295–14300, Jun. 2024, https://doi.org/10.48084/etasr.7401.

[16] T. Uckan, "Integrating PCA with deep learning models for stock market Forecasting: An analysis of Turkish stocks markets," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 8, Oct. 2024, Art. no. 102162, https://doi.org/10.1016/j.jksuci.2024.102162.

[17] D. Festa *et al.*, "Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering," *International Journal of Applied Earth Observation and Geoinformation*, vol. 118, Apr. 2023, Art. no. 103276, https://doi.org/10.1016/j.jag.2023.103276.

[18] Y. Sun, H. Liu, and Y. Gao, "Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model," *Heliyon*, vol. 9, no. 2, Feb. 2023, Art. no. e13384, https://doi.org/10.1016/j.heliyon.2023.e13384.

[19] M. Riza, K. B. Seminar, and A. Maulana, "Pembentukan Target Pasar Berdasarkan Data Stream Transaksi Kartu Kredit (Clustering dan Association Rule) pada PT Bank Bukopin," *Jurnal Aplikasi Bisnis dan Manajemen*, vol. 4, no. 1, pp. 86–86, Jan. 2018, https://doi.org/10.17358/jabm.4.1.86.

[20] Z.-J. Lee, C.-Y. Lee, L.-Y. Chang, and N. Sano, "Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation," *Symmetry*, vol. 13, no. 9, Sep. 2021, Art. no. 1557, https://doi.org/10.3390/sym13091557.

[21] J. Zhang, P. Lin, and A. Simeone, "Information mining of customers preferences for product specifications determination using big sales data," *Procedia CIRP*, vol. 109, pp. 101–106, Jan. 2022, https://doi.org/10.1016/j.procir.2022.05.221.

[22] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Information Processing & Management*, vol. 59, no. 6, Nov. 2022, Art. no. 103085, https://doi.org/10.1016/j.ipm.2022.103085.

[23] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," *Applied Soft Computing*, vol. 113, Dec. 2021, Art. no. 107924, https://doi.org/10.1016/j.asoc.2021.107924.

[24] A. Kassambara and F. Mundt, "factoextra: Extract and Visualize the Results of Multivariate Data Analyses." Apr. 01, 2020, [Online]. Available: https://cran.r-project.org/web/packages/factoextra/index.html.

[25] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *Journal of Statistical Software*, vol. 61, pp. 1–36, Nov. 2014, https://doi.org/10.18637/jss.v061.i06.

[26] E. Schubert and P. J. Rousseeuw, "Fast and eager k-medoids clustering: O (k) runtime improvement of the PAM, CLARA, and CLARANS algorithms," *Information Systems*, vol. 101, Nov. 2021, Art. no. 101804, https://doi.org/10.1016/j.is.2021.101804.

[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, https://doi.org/10.1016/0377-0427(87)90125-7.

[28] R. Gustriansyah, J. Alie, A. Sanmorino, R. Heriansyah, and M. N. M. M. Noor, "Machine Learning for Clustering Regencies-Cities Based on Inflation and Poverty Rates in Indonesia," *Indonesian Journal of Information Systems*, vol. 5, no. 1, pp. 64–73, Aug. 2022, https://doi.org/10.24002/ijis.v5i1.5682.

[29] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, https://doi.org/10.1109/TPAMI.1979.4766909.