Volume 11 Issue 1 Year 2026 | Page 452-462 ISSN: 2527-9866
Received: ***30-01-2026*** | *Revised:* ***14-02-2026***| *Accepted:* ***28-02-2026***

Applying Clustering Techniques for Customer Segmentation Based on Shipping Behavior, Cost, and Satisfaction in Logistics Services

Jaka Sunara[1*], Agus Purnomo[2], Maniah[3]
[1,2,3]Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia, 40151
e-mail: jaksun0673@ulbi.ac.id[1], aguspurnomo@ulbi.ac.id[2], maniah@ulbi.ac.id[3]
*Correspondence: jaksun0673@ulbi.ac.id

Abstract: *In modern logistics operations, behavioral data-based customer segmentation plays a crucial role in optimizing service delivery and achieving competitive differentiation. This study proposes a clustering-based approach using K-Means, Agglomerative, and Gaussian Mixture to segment sender-level customer profiles in a logistics network based on shipping cost and delivery duration, while customer satisfaction is used for post-cluster profiling and interpretative analysis. A comprehensive preprocessing pipeline is implemented, including temporal feature engineering and sender-based statistical aggregation. Grid search is used for hyperparameter tuning, and clustering performance is evaluated using the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The results indicate that K-Means with two clusters achieve the highest silhouette score (0.843), outperforming the aggregative and Gaussian mixture models. Principal Component Analysis (PCA) reveals clear separability between clusters labeled as Efficient Senders and Costly & Slow Senders. These findings provide actionable information for logistics service providers to improve pricing strategies, delivery efficiency, and customer satisfaction through intelligent segmentation.*

Keywords: *Clustering, K-Means, Customer Segmentation, Logistics Optimization, Data-Driven Logistics.*

1. INTRODUCTION

In the realm of modern logistics management, particularly in relation to customer segmentation, it is increasingly recognized that aligning service differentiation with distinct customer behaviors is paramount to maintaining operational efficiency and competitive edge [1]. Segmentation practices in logistics operations, especially during last-mile deliveries and B2B services, reveal critical variations in shipping frequency, delivery duration, costs, and customer satisfaction. Effective segmentation empowers logistics firms to efficiently allocate resources, tailor their service offerings, and prioritize high value customers, significantly improving service level performance and supporting long-term profitability [2], [3], [4].

In the midst of the growing complexity of customer data, traditional segmentation strategies that rely on basic financial metrics or manual classifications have proven inadequate. Modern logistics management requires the application of data-driven techniques that can reveal latent patterns within multidimensional customer data sets. Clustering algorithms such as KMeans and DBSCAN represent scalable methodologies for identifying behavioral segments without predefined labels, facilitating informed decision making in customer prioritization, pricing strategies, and delivery optimization [5], [6], [7], [8]. Such algorithms play a crucial role in addressing key managerial tasks such as demand forecasting and service design, thereby converting customer data into actionable segments that improve operational effectiveness [9].

Despite the promising development of clustering methodologies in logistics analytics, previous research has focused predominantly on isolated variables such as purchasing frequency or order size. There is a notable gap in the literature on integrated segmentation frameworks that consider multiple dimensions, including delivery behavior, shipping cost, and customer satisfaction concurrently. This multi-faceted approach is essential for logistics managers, especially in dynamic and cost-sensitive environments where both operational efficiency and customer-centric strategies are critical [10]. Addressing this gap, the development of a clustering-based segmentation framework is proposed that is tailored to the profiling of logistics customers. Using real-world operational data, a comprehensive model is constructed that incorporates diverse attributes, such as delivery patterns and service ratings, employing KMeans, Agglomerative Clustering, and Gaussian Mixture Models for improved accuracy. Internal validation is performed using clustering evaluation metrics such as the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index to ensure robustness and segmentation reliability [11], [12].

This study contributes three significant advances to the field: (1) it proposes a practical segmentation model tailored to the strategic objectives of logistics organizations, (2) it illustrates the efficacy of clustering techniques in capturing diverse customer segments within the logistics domain, and (3) it offers strategic insights that support logistics managers in implementing targeted service planning, differentiated pricing strategies, and performance monitoring [12], [13]. Collectively, these findings underscore the transformative impact of sophisticated customer segmentation on logistics service delivery within today's complex marketplace. The remainder of this paper is structured as follows. The Method section details the methodology, including data processing and model optimization. The Results and Discussion section presents results and comparisons. The Conclusion section concludes with key findings and future directions.

2. METHODS

This study adopts a CRISP-DM–inspired analytical framework tailored for unsupervised learning. The research design focuses on customer segmentation using clustering techniques applied to aggregated logistics operational data. The workflow consists of data collection, preprocessing (including temporal transformation, feature engineering, aggregation, and normalization), hyperparameter tuning using grid search, clustering model implementation (K-Means, Agglomerative, and Gaussian Mixture), and internal cluster validation using Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. The overall methodology is illustrated in Figure 1.

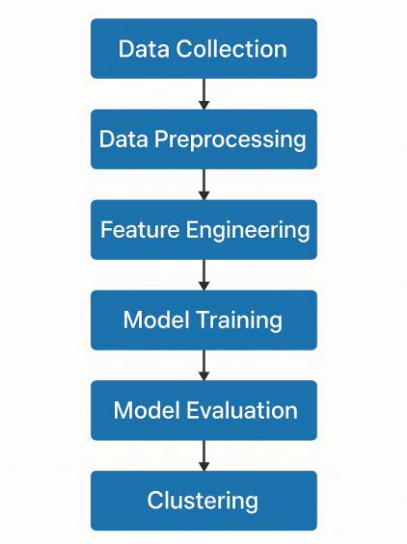

Figure 1. The proposed research methodology.

A. Data Collection

The data set used in this study was obtained from internal shipment records of a government-owned logistics company in Indonesia. The data were extracted from the company's official transaction system known as POSAJA, which records real-time logistics operations. This data set comprises a total of 209,320 delivery transaction records collected over a seven-month period, from January to July 2025. Each entry includes detailed information such as shipment and delivery dates, shipping costs, package weights, customer satisfaction ratings, and unique identifiers for both senders and receivers.

B. Preprocessing

The preprocessing stage is crucial to prepare the data before applying clustering algorithms. Several key steps are performed in this study, including temporal transformation, feature engineering, statistical aggregation, and normalization. Although the original dataset consists of transaction-level shipment records, the unit of analysis in this study is the sender (customer) profile. All transaction records were aggregated by Sender ID to construct sender-level behavioral profiles, including average shipping cost, average delivery duration, shipment frequency, and average customer rating. Therefore, clustering was performed on aggregated sender profiles rather than individual transactions.

1. Datetime Conversion of Temporal Columns

In the shipping dataset, the shipping-date and delivery-date columns are converted to date-time format to enable precise numerical operations. This conversion is necessary to calculate the delivery duration. Mathematically, the delivery duration *D* in hours is defined as:

$$D = \frac{t_{\text{received}} - t_{\text{shipped}}}{3600} \quad (1)$$

Equation 1 indicates that $t_{received}$ and $t_{shipped}$ are time points in seconds.

2. Feature Engineering: Delivery Duration

The duration of delivery serves as a derived feature and is computed from the temporal difference between the shipping and the delivery times. Capture the operational efficiency and responsiveness of each sender.

3. Statistical Aggregation by Sender ID

The data were first grouped by sender, and for each sender we computed several aggregate statistics: total and average shipping costs, average package weight, average delivery duration, average customer rating, and total number of shipments. Although the original dataset is transaction-level, the unit of analysis in this study is the sender profile obtained through aggregation by Sender ID. This aggregation step transforms transaction-level data into sender-level analytical units, which serve as the input observations for clustering. These aggregated values are computed using the following general formula.

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij} \quad (2)$$

Equation 2 explains that $\bar{x}_i$ is the average value of the characteristic *x* of the sender *i*, and N_i is the number of shipments for that sender.

4. Data Normalization with Robust Scaler
To ensure fair clustering, normalization is applied to features with varying scales. Instead of using Z-score standardization, this study applies the *Robust Scaler* method, which transforms each feature by removing the median and scaling according to the interquartile range (IQR). The transformation is defined as follows.

$$z = \frac{x - Q_2}{Q_3 - Q_1} \quad (3)$$

Equations 3 explain Q_1, Q_2, and $Q3$ are the first, median, and third quartiles of the distribution, respectively. This method is more robust to outliers compared to standard Z-score normalization and is suitable for skewed distributions observed in the dataset.

5. Feature Selection for Clustering
Only two features average shipping cost and average delivery duration were used as clustering input variables. Customer satisfaction ratings were excluded from the clustering process and were instead utilized for post-cluster behavioral profiling.

C. Hyperparameter Tunning
Hyperparameter tuning is a crucial step in optimizing the performance of machine [24] learning models. It involves selecting the best combination of parameters that are not learned from the training data but can significantly affect the outcome of the model. In clustering algorithms such as KMeans, Agglomerative Clustering, and Gaussian Mixture Model (GMM), examples of hyperparameters include the number of clusters *K*, the linkage criterion in hierarchical clustering (e.g., ward, average, complete), and the number of Gaussian components or covariance type in GMMs.

In this study, hyperparameter optimization is performed using *Grid Search*, a commonly used brute-force method that evaluates all possible combinations within a predefined search space to determine the most optimal configuration [14], [15].

Mathematically, the grid search process can be expressed as follows:

$$\theta^* = \arg\max_{\theta \in \Theta} \mathcal{M}\left(f(D;\theta)\right) \quad (4)$$

Equation 4 indicates that the optimal hyperparameter configuration θ^* is the one that maximizes the evaluation metric $\mathcal{M}$ in the search space Θ. In this formulation, θ refers to a candidate combination of hyperparameters, and Θ denotes the complete set of configurations explored by the grid search. The function $f(D;\theta)$ represents the clustering model trained on the dataset D using parameters θ, and $\mathcal{M}$ is the evaluation metric, such as the silhouette score, used to quantify model quality. The grid search provides an exhaustive and interpretable approach*.

To ensure transparency and reproducibility, the grid search was conducted over predefined parameter ranges. The number of clusters (K) was explored from 2 to 6 for all models. For K-Means, k-means++ initialization was used with n_init = 10, a maximum of 300 iterations, and a fixed random seed. Agglomerative Clustering evaluated ward, complete, and average linkage using Euclidean distance. For the Gaussian Mixture Model (GMM), the number of components ranged from 2 to 6, covariance types (full, tied, diag, spherical) were tested, initialization was set to k-means, and the maximum EM iterations were set to 200 with a fixed random seed. The final configuration was selected based on the highest mean Silhouette Score from the 6-fold stability-based resampling validation, supported by DBI and CHI as complementary metrics.

D. Models Clustering

K-Means is a widely used clustering algorithm that partitions data into nonoverlapping Kclusters by minimizing intra-cluster variance, where each cluster is represented by its centroid, the mean of its assigned point. The algorithm iteratively assigns each point to the nearest centroid and updates the centroids until convergence. Its objective function is given in Equation 5:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} | x_i - \mu_k |^2 \quad (5)$$

Equation 5 minimizes the total squared distance between each data point x_i and its cluster centroid μ_k, resulting in compact and well-separated clusters. In contrast, agglomerative clustering and the Gaussian mixture model (GMM) represent two alternative approaches to unsupervised clustering. Agglomerative clustering follows a hierarchical bottom-up approach where each observation starts in its own cluster, and pairs of clusters are iteratively merged based on a linkage criterion such as the Ward method or average linkage [16]. Meanwhile, GMM is a probabilistic model that assumes the data is generated from a mixture of several Gaussian distributions with unknown parameters. It assigns probabilities for each point belonging to a cluster and uses the Expectation-Maximization (EM) algorithm for optimization [17].

E. Semantic Labeling Based on Efficiency Score

After the clustering process, the resulting clusters are interpreted and labeled on the basis of the operational efficiency of each sender [7]. Two normalized indicators are computed: average shipping cost and average delivery duration. These indicators are scaled using min-max normalization, and a composite efficiency score is computed as the arithmetic mean of the two.

$$\text{efficiency\_score} = \frac{1}{2}(\text{norm\_cost} + \text{norm\_duration}) \quad (6)$$

Senders with lower scores are considered to be more efficient. Based on the sorted efficiency scores of the centroids of the cluster, descriptive labels such as *"Effective Sender"* and *"Costly & Slow Sender"* are assigned to each cluster. This semantic labeling improves the interpretability of the clustering outcomes in the logistics context.

F. Evaluation Metrics

To evaluate the quality of our clustering solutions, we employ three complementary metrics: the Silhouette Score [12], [18], the Davies–Bouldin Index (DBI) [19], and the Calinski–Harabasz Index (CHI) [20]. Equations 7, 8, and 9 define these metrics formally.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (7)$$

Here, $a(i)$ is the average distance from sample i to all other points in its own cluster, and $b(i)$ is the minimum average distance from i to points in any other cluster. The silhouette score $s(i)$ ranges from -1 (poor clustering) to $+1$ (strong clustering), with values close to 0 indicating overlapping clusters.

$$\text{DBI} = \frac{1}{K} \sum_{p=1}^{K} \max_{q \neq p} \frac{S_p + S_q}{M_{pq}} \quad (8)$$

In equation 8, K is the number of clusters, S_p is the average distance of all points in cluster p to its centroid, and M_{pq} is the distance between the centroids of clusters p and q. Lower DBI values indicate better cluster separation and compactness.

$$\text{CHI} = \frac{\text{trace}(B)/(K-1)}{\text{trace}(W)/(N-K)} \quad (9)$$

In equation 9, N is the total number of samples, K the number of clusters, B the scatter matrix between clusters, and W the scatter matrix within clusters. Higher CHI values reflect greater dispersion between clusters relative to within-cluster dispersion, indicating more distinct clusters.

3. RESULTS AND DISCUSSION

A. Cluster Distribution and Interpretability

The K-Means clustering algorithm successfully identified two distinct clusters: *Efficient Senders* and *Costly & Slow Senders*. This classification is illustrated in Figure 2.

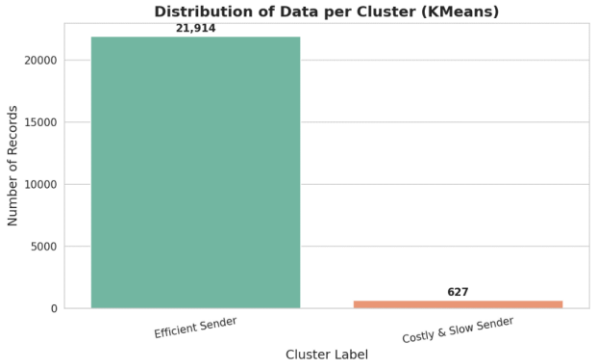

Figure 2. Distribution of Senders per Cluster (KMeans)

Based on figure 2, the distribution indicates that most senders fall into the efficient category (21,914 unique senders), while only 627 records are associated with Costly & Slow Senders. All cluster membership counts refer to aggregated sender-level profiles rather than individual transaction records The silhouette analysis confirms that the optimal number of clusters is K = 2, with a silhouette score of 0.843. The results of this analysis are presented in Figure 3.

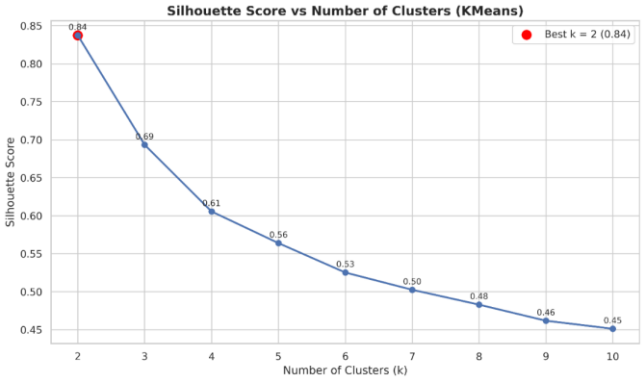

Figure 3. Silhouette Score vs. Number of Clusters (KMeans)

K-Means outperforms the Agglomerative (0.788) and Gaussian Mixture (0.625) clustering models in terms of clustering performance. Table 1 provides a detailed comparison of these models.

Table 1. Evaluation Metrics for Clustering Models (6-Fold CV)

Model	Silhouette Score	DBI Score	CHI Score
K-Means	**0.843**	0.57	**18337.454**
Agglomerative	0.788	**0.49**	16941.088
Gaussian Mixture	0.625	1.01	9977.388

To assess the robustness and stability of the clustering results, 6-fold resampling validation procedure was applied at the sender level. The aggregated sendr-level dataset was randomly partitioned into six equal subsets. For each iteration, clustering models were trained on five folds and evaluated on the remaining fold.

Internal validation metrics, including Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI), were computed for each fold. The final reported values correspond to the mean performance across the six resampling iterations. This approach does not represent supervised cross-validation but rather a stability-based validation mechanism designed to evaluate clustering consistency under repeated data partitioning.

B. Geographical Segmentation

The chart highlights the distribution of senders per product and group. This distribution is illustrated in Figure 4.

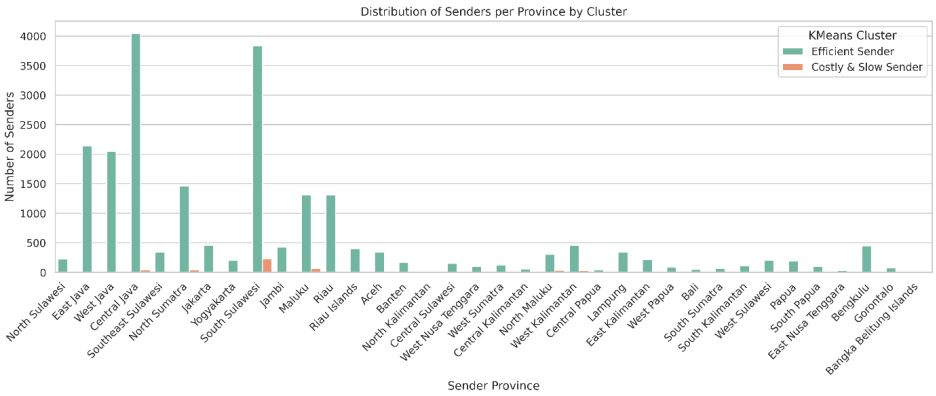

Figure 4. Distribution of Senders per Product by Cluster

Radar charts for selected provinces: Central Java, South Sulawesi, East Java, West Java, and North Sumatra illustrate that senders in the Costly & Slow category tend to have longer average delivery durations and higher shipping costs, with Central Java and North Sumatra exhibiting the most noticeable differences. These provincial comparisons are depicted in Figures 5–9.

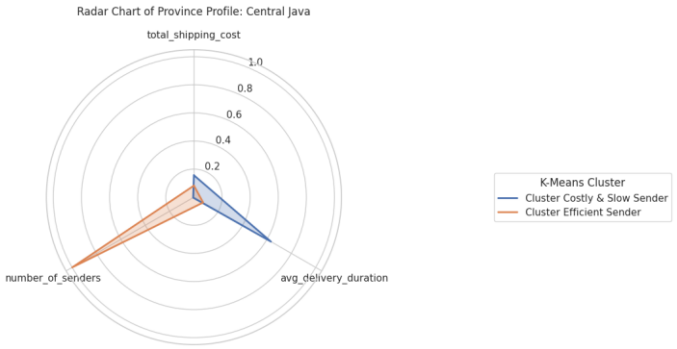

Figure 5. Radar Chart of Central Java Sender Profile Based on K-Means Clustering

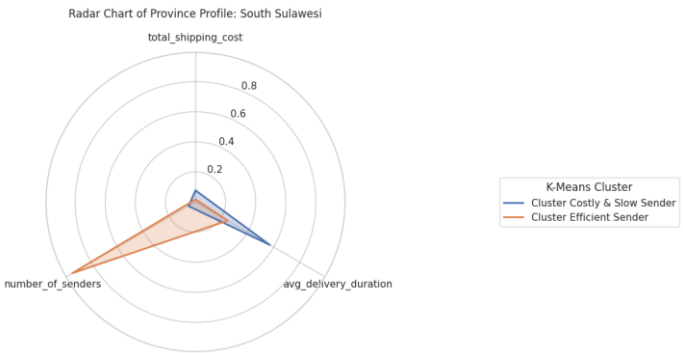

Figure 6. Radar Chart of South Sulawesi Sender Profile Based on K-Means Clustering

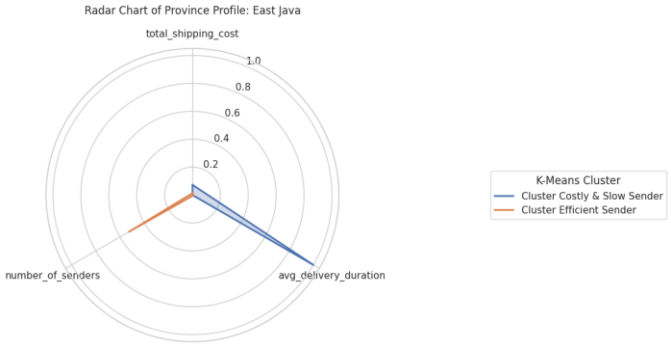

Figure 7. Radar Chart of East Java Sender Profile Based on K-Means Clustering

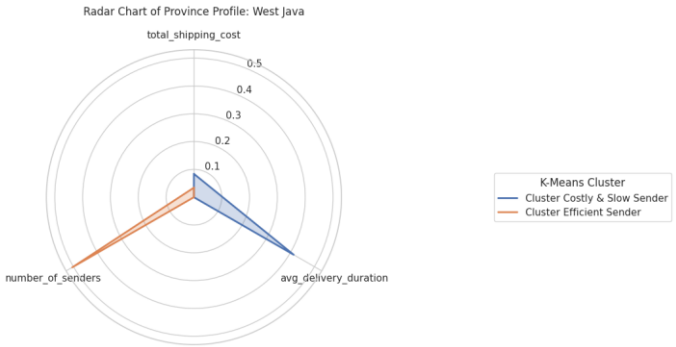

Figure 8. Radar Chart of West Java Sender Profile Based on K-Means Clustering

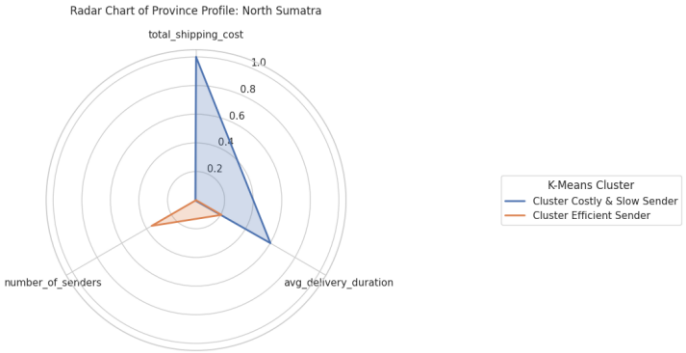

Figure 9. Radar Chart of North Sumatra's Sender Profile Based on K-Means Clustering

C. Sender Profile Characteristics

Figure 10 shows the radar chart of the average sender profile per cluster. Costly & Slow Senders are associated with higher average shipping costs, delivery durations, and package weights, while receiving lower average customer ratings. In contrast, efficient senders are characterized by faster delivery, lower costs, and higher satisfaction.

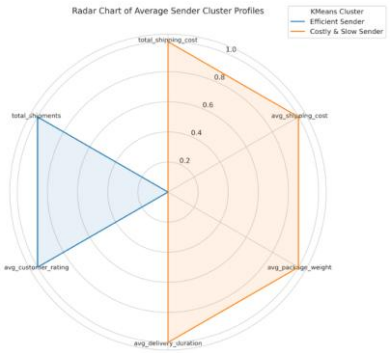

Figure 10. Radar Chart of Average Sender Profiles per Cluster

D. Cluster Visualization Using PCA

The PCA plot visualizes the two clusters in reduced dimensions, showing a clear spatial separation between Efficient and Costly & Slow senders. This confirms that the selected features provide good discriminative power for clustering sender behavior. This spatial separation is visualized in Figure 11.

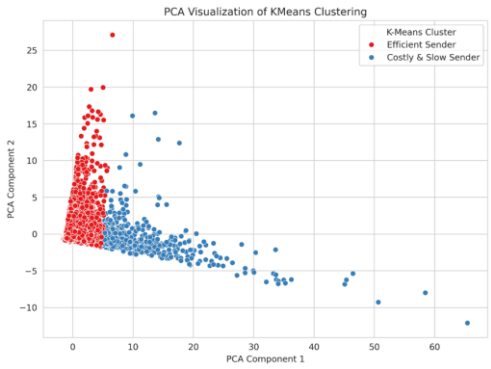

Figure 11. PCA Visualization of KMeans Clustering

E. Discussion

1. Implications for Strategic Segmentation

The clustering results highlight meaningful distinctions in customer behavior, offering an opportunity for logistics companies to implement data-driven segmentation strategies. The ability to identify Efficient versus Costly & Slow Senders allows for customized pricing models, personalized service strategies, and better performance monitoring across geographic regions.

2. Operational Recommendations

Regions with a high concentration of inefficient senders can be targeted for intervention through operational improvements such as route optimization, courier training, or incentive-based programs. Radar charts serve as valuable tools for regional managers to assess performance and prioritize improvements.

Table 2. Comparison of Related Works and Proposed Method

Ref	Model Used	Feature Engineering	Grid Search	Cross-Validation	PCA	Cluster Validity	Evaluation Metrics	Best Silhouette Score
Cahyana et al. [21]	Hybrid Clustering (Average Linkage + K-Means)	✓	×	×	✓	✓	Segment Profile, Centroid Gap	–
Trappey et al. [22]	Ward + K-Means (Two-Stage Clustering)	✓	×	×	×	✓	Euclidean Distance, Customer Preference Stats	–
Wu & Chou, [3]	Fuzzy C-Means (Soft Clustering)	✓	×	×	✓	✓	Silhouette, Dunn Index	0.804
This Study	**K-Means**	**✓**	**✓**	**✓**	**✓**	**✓**	**Silhouette, CHI, DBI**	**0.843**

Table 2 summarizes the comparative analysis between the proposed study and existing related works. The study by Cahyana et al. [21] employed a hybrid clustering approach combining average linkage and K-Means, focusing on interpretative segment profiles and centroid gaps without involving validation through cross-validation or automated parameter tuning. Similarly, Trappey et al. [22] implemented a two-stage clustering method using Ward's method followed by K-Means, with an emphasis on distance-based evaluation and customer preference ranking, but lacked

dimensionality reduction or cluster quality scoring, such as silhouette Purnomo et al. [23]. In contrast, Wu et al. [3] applied a soft clustering method (Fuzzy C-Means) and integrated PCA and multiple internal validation metrics including the silhouette and Dunn index, obtaining a silhouette score of 0.804. However, none of these studies applied systematic grid search or cross-validation mechanisms to maintain model robustness.

The proposed study distinguishes itself by incorporating comprehensive pre-processing and feature engineering, hyperparameter optimization using grid search, and model robustness via cross-validation. Furthermore, PCA is used for dimensionality reduction, and multiple cluster validity indices (Silhouette, Calinski-Harabasz Index, and Davies-Bouldin Index) are used for evaluation. Although multiple evaluation metrics were considered, the final model selection followed a prioritized multi-metric strategy. The Silhouette Score was defined as the primary criterion due to its ability to simultaneously measure intra-cluster cohesion and inter-cluster separation. The Davies–Bouldin Index (DBI) and Calinski–Harabasz Index (CHI) were used as complementary indicators. While Agglomerative Clustering achieved a slightly lower DBI, K-Means obtained the highest mean Silhouette Score (0.843) and the highest CHI value across the 6-fold stability-based validation. Therefore, K-Means was selected as the best-performing model under the defined evaluation framework. This highlights the effectiveness of combining systematic tuning and validation techniques in improving customer segmentation in logistics

4. CONCLUSIONS

This study uses K-Means clustering to identify two customer segments: efficient senders (low cost, fast delivery, high satisfaction) and Costly and Slow Senders (high cost, slow delivery, low satisfaction), allowing logistics providers to tailor pricing, performance monitoring and partnerships. Key contributions include a practical operations-focused segmentation model, validation of clustering for sender profiling, and strategic guidance for targeted service planning. Future work will incorporate additional features (e.g. package dimensions, seasonality, real-time tracking), compare alternative clustering methods, analyze segment stability over time, and link segmentation to predictive models for proactive logistics management.

REFERENCES

[1] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010, doi: 10.1016/j.patrec.2009.09.011.

[2] Z. Liu, Y. Li, Z. Zhang, and R. Zhao, “The Impact of Business Strategic Orientation on Innovation-Driven Mergers and Acquisitions: An Empirical Study,” *Discret. Dyn. Nat. Soc.*, 2021, doi: 10.1155/2021/5254365.

[3] R.-S. Wu and P.-H. Chou, “Customer segmentation of multiple category data in e-commerce using a soft-clustering approach,” *Electron. Commer. Res. Appl.*, vol. 10, no. 3, pp. 331–341, 2011, doi: 10.1016/j.elerap.2010.11.002.

[4] P. Rajapandian, A. Karunamurthy, V. Vasanth, and M. Meganathan, “E-Commerce Customer Segmentation: A Clustering Approach in A Web-Based Platform,” *J. Eng. Technol. Appl. Phys.*, vol. 7, no. 1, pp. 71–79, 2025, doi: 10.33093/jetap.2025.7.1.12.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012. doi: 10.1016/C2009-0-61819-5.

[6] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley \& Sons, Inc., 1990. doi: 10.1002/9780470316801.

[7] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005, doi: 10.1109/TNN.2005.845141.

[8] S. F. Novia, E. A., Rahayu, W. I., & Pane, “Implementasi Algoritma K - Means Clustering Tingkat Kepentingan Tagihan Rumah Sakit Di PT Pertamina (Persero),” *J. Ilm. Inform.*, vol. 8, no. 1, pp. 44–52, 2020, doi: 10.33884/jif.v8i01.1844.

[9] H. Luo, K. Wang, Z. Wang, and L. Cheng, “A novel clustering-based algorithm for distribution network design in pharmaceutical retail logistics,” *Ind. Manag. \& Data Syst.*, vol. 126, pp. 1–24,

2025, doi: 10.1108/IMDS-03-2025-0387.
[10] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999, doi: 10.1145/331499.331504.
[11] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On Clustering Validation Techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2–3, pp. 107–145, 2001, doi: 10.1023/A:1012801612483.
[12] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
[13] K.-S. Chen and C.-M. Yu, "Fuzzy test model for performance evaluation matrix of service operating systems," *Comput. \& Ind. Eng.*, vol. 140, p. 106240, 2020, doi: 10.1016/j.cie.2019.106240.
[14] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, no. 10, pp. 281–305, 2012.
[15] A. G. Putrada, E. K. Laeli, S. F. Pane, N. Alamsyah, and M. N. Fauzan, "TPOT on Increasing The Performance of Credit Card Application Approval Classification," in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, IEEE, 2022, pp. 216–221. doi: 10.1109/ICICyTA57421.2022.10038063.
[16] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2785–2797, 2015, doi: 10.1016/j.eswa.2014.09.054.
[17] W. Zhang, J. Zhao, L. Yu, and S. Wang, "GMM Enhanced Anchor-Based Spectral Clustering for Large-Scale Data," *IEEE Trans. Neural Networks Learn. Syst.*, 2025, doi: 10.1109/TNNLS.2025.3571473.
[18] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, 2013, doi: 10.1016/j.patcog.2012.07.021.
[19] I. K. Khan *et al.*, "Standardization of expected value in gap statistic using Gaussian distribution for optimal number of clusters selection in K-means," *Egypt. Informatics J.*, vol. 30, p. 100701, 2025, doi: 10.1016/j.eij.2025.100701.
[20] J. Li, D. Hassan, S. Brewer, and R. Sitzenfrei, "Is clustering time-series water depth useful? An exploratory study for flooding detection in urban drainage systems," *Water*, vol. 12, no. 9, p. 2433, 2020, doi: 10.3390/w12092433.
[21] B. E. Cahyana, U. Nimran, H. N. Utami, and M. Iqbal, "Hybrid cluster analysis of customer segmentation of sea transportation users," *J. Econ. Financ. Adm. Sci.*, vol. 25, no. 50, pp. 321–337, 2020, doi: 10.1108/JEFAS-07-2019-0126.
[22] C. V Trappey, A. J. C. Trappey, A.-C. Chang, and A. Y. L. Huang, "Clustering analysis prioritization of automobile logistics services," *Ind. Manag. \& Data Syst.*, vol. 110, no. 5, pp. 731–743, 2010, doi: 10.1108/02635571011044759.
[23] Purnomo, A., Putrada, A.G., Habibi, R., Syafrianita. (2024). MDI and PI XGBoost Regression-Based Methods: Regional Best Pricing Prediction for Logistics Services. TELKOMNIKA (Telecommunication, Computing, Electronics and Control). 22(5), Page 1157~1166, ISSN: 1693-6930, e-ISSN: 2302-9293, https://doi.org/10.12928/TELKOMNIKA.v22i5.26037
[24] Tedyyana, A., Ahmad, A. A., Idrus, M. R., Mohd Shabli, A. H., Abu Seman, M. A., Ghazali, O., & Abd Razak, A. H. (2024). Enhance Telecommunication Security Through the Integration of Support Vector Machines. *International Journal of Advanced Computer Science & Applications*, *15*(3).