Technologia Journal: Jurnal Informatika
E-ISSN:3046-9163
Vol.2.No.4, November 2025
DOI : https://doi.org/10.62872/xe3zrt53

Analysis of Order Data Customer Segmentation in Logistics Companies Using K-Medoids and DBSCAN Algorithms

Mujiono Sadikin[1], Nanda Azvita[2]
Program Studi Informatika Fakultas Ilmu Komputer Universitas Bhayangkara Jakarta Raya[1,2]
e-mail: *mujiono@dsn.ubharajaya.ac.id

Inputed : October 10, 2025 *Revised : October 29, 2025*
Accepted : October 28, 2025 *Published : November 03, 2025*

ABSTRACT

The development of the logistics industry makes the use of customer data to understand market behavior and needs increasingly important. This study aims to segment customers based on logistics company order data using the *K-Medoids algorithm and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)*. This approach is used to identify customer groups with similar characteristics to support more effective marketing and service strategies. This study uses 12,000 customer order data entries from the past year, with variables including order, cost, and receiving location. The data is processed through preprocessing stages (cleaning, transformation, and normalization) before being applied to two clustering models. The analysis results show that the K-Medoids algorithm produces a Silhouette Score of 0.3559, while DBSCAN obtained a score of 0.3233. These values indicate that K-Medoids has more compact and well-separated clusters than DBSCAN. Thus, the K-Medoids method is more effective in segmenting customers to support strategic decisions of logistics companies.

Keywords: *Customer Segmentation, K-Medoids, DBSCAN, Data Mining, Logistics.*

INTRODUCTION

The advancement of the digital era has brought significant changes to the way companies operate, including in the logistics sector. Global competition demands that companies intelligently manage customer data to improve service quality and operational efficiency (Shadani & Nasution, 2024). The abundance of customer data from ordering activities (order data) is a valuable source of information that, when analyzed properly, can help companies understand customer behavior patterns, predict demand, and formulate more effective marketing strategies.

In the logistics industry, each customer has different characteristics and preferences, such as service type, delivery frequency, and cost level. Therefore, an analytical strategy is needed to group customers into specific segments based on their shared characteristics (Masbullah & Bahri, 2024). This process is known as customer segmentation, which is the main foundation for companies to personalize services, determine pricing strategies, and optimize resources.

One relevant approach for this purpose is clustering analysis in data mining. Clustering techniques are used to group data without initial labels or categories, thereby revealing hidden structures within the data. Two widely used algorithms are K-Medoids and DBSCAN. The K-Medoids algorithm is known to produce stable clusters and is resistant to outliers, while DBSCAN is effective for finding irregularly shaped clusters and is capable of detecting noise in the data (Aulia et al., 2025).

Several previous studies have demonstrated the effectiveness of this method in various business and industrial contexts. For example, Triyanto (2015) applied K-Medoids to determine product marketing strategies and obtain representative customer clusters. Meanwhile, Muhajir and Sari (2020) combined K-Means and DBSCAN to cluster films based on their characteristics, demonstrating significant differences in clustering results. However, studies on the simultaneous application of these two algorithms in the context of the Indonesian logistics industry are still rare (Azzahra & Akbar, 2024).

Thus, the research gap lies in the need for a comparative evaluation between the K-Medoids and DBSCAN algorithms for customer segmentation analysis based on logistics company order data (Yosia & Siregar, 2024). The novelty of this research is the simultaneous application of both algorithms to large-scale real customer data, with an in-depth analysis of the differences in the resulting cluster characteristics.

This study aims to implement the K-Medoids and DBSCAN algorithms in the customer segmentation process based on order data obtained from a logistics company (Agustin et al., 2025), in order to identify customer behavior patterns and transaction characteristics that have certain similarities. Through the application of both algorithms, this study attempts to evaluate the performance of each method using the Silhouette Score metric as an indicator to assess the level of cohesion and separation between the formed clusters, so that it can be known which algorithm is more optimal in grouping customer data with complex structures and distributions. In addition, the clustering results obtained are interpreted in depth to produce strategic insights that can be used as a basis for decision making in formulating business strategies, especially in improving operational efficiency, service effectiveness, and customer satisfaction in the increasingly competitive and dynamic logistics sector.

METHODOLOGY

This research uses a descriptive quantitative approach with a data mining method based on unsupervised learning (Unus, 2024).

Data source

The research data was collected from a logistics company operating in Indonesia, with a total of 12,000 order data entries over the past year. Each entry contains customer information such as the order type (Job Order Type), cost (Cost Amount), place of receipt (Place of Receipt), and customer identity (Client).

Research Stages

a. Data Preprocessing:

Includes data cleansing to remove empty and duplicate values, feature selection to select important attributes, label encoding to convert categorical data to numeric, and data normalization using StandardScaler (Gori et al., 2024).

b. Model Application:

K-Medoids is applied to determine the optimal cluster based on the Elbow and Silhouette Score values (Rahmawati et al., 2024).

DBSCAN was applied with parameters eps = 0.15 and min_samples adjusted to the data density.

c. Model Evaluation:

Using the Silhouette Coefficient to assess the extent to which objects within a cluster have internal similarities and differences between clusters.

d. Interpretation and Visualization:

Visualization of cluster results is done using a scatter plot to describe the distribution of customers based on the segments formed (Tabianan et al., 2022).

RESULTS AND DISCUSSION

Clustering Results

The clustering process using K-Medoids resulted in nine relatively evenly distributed customer clusters. The cluster with the highest average cost (around 2,326.84) and a large number of clients (average 235) was identified as a premium customer that contributes significantly to revenue. Meanwhile, the cluster with low costs and low order frequency represents individual customers or small businesses.

The DBSCAN model produces a number of clusters of varying shapes and sizes, as well as some noise points (Yu et al., 2019). This demonstrates DBSCAN's ability to capture irregular variations in data, such as customers with inconsistent ordering patterns or low transaction frequency.

The clustering stage begins after the data preprocessing process is complete, which includes cleaning for missing values, removing duplicates, encoding categorical attributes, and normalizing all numeric variables using StandardScaler (Desai et al., 2025). The cleaned and standardized data is then used to build two clustering models: K-Medoids and DBSCAN.

In the K-Medoids model, the optimal number of clusters is determined using two approaches, namely the Elbow Method and the Silhouette Coefficient (Firmansyah et al., 2025). Based on the test results in the range of k = 2 to k = 11, the optimal point was found at k = 9, where the Silhouette Score value reached 0.3559. This value indicates that objects in the cluster have a fairly close distance from each other (high cohesion), and are quite far from other clusters (good separation).

The cluster distribution visualization results show that each cluster has different characteristics based on the Cost Amount and Job Order Type variables. For example, cluster 0 contains customers with a high number of

orders and large transaction values, while clusters 5 and 8 represent customers with fluctuating transaction activity. This shows that the K-Medoids method is able to group customers with both relatively stable and varying order patterns, making it suitable for customer segmentation analysis with mixed numeric data.

Meanwhile, the DBSCAN model was implemented with parameters eps = 0.15 and min_samples = 5, which were obtained from the k-distance graph analysis. The implementation results showed that DBSCAN was able to form a number of clusters with various sizes and shapes and identified several data points that did not belong to any cluster (called noise points). This proves the superiority of DBSCAN in handling data that is not homogeneous and has a non-linear distribution.

Overall, the clustering results of both models indicate that the customer data patterns of logistics companies are highly complex, with significant variations in order frequency, transaction value, and delivery location (Zheng et al., 2023). The differences in cluster structure between K-Medoids and DBSCAN provide an important basis for selecting the appropriate algorithm for a specific business context and analysis objective.

Model Evaluation

The Silhouette Score for K-Medoids is 0.3559, while DBSCAN obtains 0.3233. This value indicates that K-Medoids has more compact and well-separated clusters, while DBSCAN tends to produce overlapping clusters but is more flexible in handling complex data distributions.

Additionally, DBSCAN's computation time is faster because the algorithm only calculates distances within a certain radius (eps), while K-Medoids requires a full iteration through the data. However, overall, K-Medoids is superior in producing cluster structures that are easily interpreted in business terms.

Model evaluation was conducted to assess the quality and effectiveness of the clustering results produced by each algorithm. This study used the Silhouette Score metric, which measures the extent to which objects within a cluster are similar to their own cluster compared to other clusters. A value close to +1 indicates that the object is placed in the correct cluster, while a value close to 0 or negative indicates overlap between clusters.

The evaluation results show that:

The K-Medoids algorithm obtained a Silhouette Score value of 0.3559, indicating that the cluster formed was quite compact and had clear separation between customer groups.

The DBSCAN algorithm obtained a Silhouette Score of 0.3233, which is slightly lower, indicating some overlap between clusters although this model remains effective for identifying complex data densities.

From these results, it can be concluded that K-Medoids has better performance for data with a relatively regular structure, while DBSCAN excels in detecting anomalies and patterns in data with varying densities. In the

context of the logistics business, K-Medoids can be used to group customers based on economic value and stable purchasing behavior, while DBSCAN is suitable for detecting customers with erratic or inconsistent behavior.

Cluster Characteristics Analysis

Each cluster has unique characteristics:

- Cluster 0 (K-Medoids): Large customers with high transaction values and distributed delivery locations. High potential for priority service strategies or loyalty programs.
- Cluster 1: Customers with moderate activity and moderate costs, can be upgraded to the premium category through promotions and incentives.
- Cluster 2: New customers or small transactions, suitable for retention marketing strategies.

Meanwhile, in DBSCAN, large clusters with high density indicate active customers, while noise points indicate unstable or inconsistent customers in ordering.

Further analysis was conducted on the characteristics of each cluster to understand customer behavior in each segment (Abbasimehr & Shabani, 2021). Based on the results of the K-Medoids model, nine customer clusters were obtained with the following profiles:

- Cluster 0: Large corporate customers with high transaction value (average Cost Amount 2,326.84) and high order frequency. This cluster contributes significantly to the company's revenue and can be prioritized for premium services.
- Cluster 1: Customers with medium activity levels (average cost of 696.05) and moderate order stability. This segment has the potential to be upgraded to core customers through loyalty programs.
- Clusters 2 and 3: Small and new customers with low transaction values and irregular orders. A more aggressive promotional approach is needed to increase purchase frequency.
- Clusters 5 and 6: Customers with high order variety and high but unstable cost values. This segmentation describes project or seasonal clients.
- Cluster 8: Low-cost but high-frequency customers, generally from the small retail and local distribution sectors.

Meanwhile, the DBSCAN clustering results show a more flexible structure, where the main cluster is formed in high-density data areas, while customers that do not meet the density criteria are considered noise. DBSCAN cluster 0 describes customers with high activity but low costs, while cluster 2 describes large customers with high-value orders but are limited to certain regions.

This comparison shows that DBSCAN is able to identify customers that do not fit common patterns (e.g., customers with high costs but low frequency), which are often overlooked by distance-based algorithms such as K-Medoids.

Business Visualization and Interpretation Analysis

Visualizing clustering results using a two-dimensional scatter plot strengthens our understanding of customer distribution within the data space (Ali et al., 2019). In the K-Medoids model, data points appear to form relatively symmetrical and clearly separated cluster boundaries, while in the DBSCAN model, the boundaries between clusters are more dynamic and irregular. This demonstrates the fundamental difference between partition-based and density-based approaches (Anjara, 2025).

In business terms, these results provide strategic information for logistics companies to take a different approach to each customer segment:

- Customers in high-value clusters can be focused on customer retention strategies and personalized service offerings.
- Customers in low-cost but high-frequency clusters can be targeted for operational efficiency, for example through distribution channel optimization or service bundling.
- Customers detected as noise can be an indicator of data inconsistencies, or an opportunity for a new customer acquisition strategy through a direct sales approach.

Business Implications

The research findings provide insights for logistics company management to develop more data-driven strategies. By understanding customer segments, companies can:

- Customize services based on customer value.
- Improve the efficiency of operational resource allocation.
- Developing a more personalized logistics service recommendation system.

The implementation of these clustering results can also be integrated with a CRM (Customer Relationship Management) system to strengthen long-term relationships with customers (Melovic et al., 2022).

Comparison with Previous Research

The results of this study align with the findings of Sulistyawati and Sadikin (2021) that K-Medoids provides stable performance in customer segmentation, and support the study by Wahyuningtyas et al. (2023) that found DBSCAN superior in detecting noisy data. However, the contribution of this study is its direct application to complex and large-scale logistics industry data, providing a practical approach for data mining applications in the transportation and distribution sector.

The findings of this study corroborate those of a previous study by Sulistyawati & Sadikin (2021), which found that K-Medoids provides more stable customer segmentation than other algorithms such as K-Means. Furthermore, the DBSCAN results in this study are consistent with the study by Wahyuningtyas et al. (2023), which emphasized DBSCAN's superiority in handling irregular data distributions and detecting noise.

However, the main contribution of this research lies in its application to the context of the Indonesian logistics industry, which has unique characteristics such as large order data, high delivery location variation, and fluctuating order frequency (Judijanto et al., 2024). By combining the two algorithms, companies can gain two analytical perspectives:

1. Structured segmentation that describes customer value (via K-Medoids).
2. Identify anomalous patterns and non-routine customers (via DBSCAN).

Thus, the results of this study are not only theoretical, but also practical for data-based decision making in the fields of logistics and customer management (Lu et al., 2019).

CONCLUSION

This study proves that the K-Medoids and DBSCAN algorithms can be used effectively in analyzing customer segmentation based on order data in logistics companies. The analysis results show that the K-Medoids algorithm produces clustering with the highest Silhouette Score value of 0.3559, which indicates a good level of cohesion and separation between clusters, thus being able to group customers based on order patterns more accurately. Meanwhile, the DBSCAN algorithm shows advantages in terms of flexibility and resilience to noisy data, although its cluster quality is slightly lower than that of K-Medoids. Interpretation of the clustering results shows that each cluster represents a customer segment with different characteristics and business potential, providing strategic insights for companies in understanding customer behavior. Practically, these findings can be used as a basis for developing customer recommendation systems and more targeted data-driven marketing strategies in the logistics sector. For further research, it is recommended to use multi-year datasets for more representative analysis results, apply hybrid clustering methods to improve segmentation accuracy, and integrate with predictive machine learning algorithms to estimate future customer behavior.

BIBLIOGRAPHY

Abbasimehr, H., & Shabani, M. (2021). A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. *Kybernetes*, *50*(2), 221–242. https://doi.org/10.1108/K-09-2018-0506

Agustin, E. W., Uthami, K., Ulfa, A. I., Efrizoni, L., & Rahmaddeni, R. (2025). Optimization of Customer Segmentation in the Retail Industry Using the K-Medoid Algorithm. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *5*(3), 766–775. https://doi.org/10.57152/malcom.v5i3.1977

Ali, M., Jones, M. W., Xie, X., & Williams, M. (2019). TimeCluster: Dimension reduction applied to temporal data for visual analytics. *The Visual Computer*, *35*(6–8), 1013–1026. https://doi.org/10.1007/s00371-019-01673-y

Anjara, F. (2025). Analisis Visualisasi Data untuk Mendukung Pengambilan Keputusan Bisnis pada UMKM Sajisaja Menggunakan Tableau. *Ekopedia: Jurnal Ilmiah Ekonomi*, *1*(3), 1532–1539. https://doi.org/10.63822/c07wwq45

Aulia, R., Julianti, N., Putri, S. F., Efrizoni, L., & Deni, R. (2025). Optimalisasi Pengelompokan Gangguan Kecemasan dalam Mendukung Tujuan Pembangunan Berkelanjutan Menggunakan Algoritma K-Means dan K-Medoids. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *12*(2). https://doi.org/10.35957/jatisi.v12i2.11495

Azzahra, Y. A., & Akbar, Y. (2024). Komparasi Penerapan Algoritma C4.5 dan Naïve Bayes untuk Ketepatan Waktu Pengiriman Barang Pada PT. Rtrans Logistik Artamandiri. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *5*(3), 2768–2780. https://doi.org/10.35870/jimik.v5i3.1003

Dery Shadani & Muhammad Irwan Padli Nasution. (2024). Optimalisasi Pengelolaan Informasi Data Untuk Peningkatan Kualitas Layanan Di Era Digital. *Journal Of Informatics And Busisnes*, *2*(1), 47–51. https://doi.org/10.47233/jibs.v2i1.995

Desai, P., Karthik, P., Loganathan, D., Preethi, S., & Bharani, B. R. (2025). Different data cleaning techniques and normalization techniques with focus on current normalization techniques: A study. In T. Sengodan, S. Misra, & M. M, *Advances in Electrical and Computer Technologies* (1st ed., pp. 332–349). CRC Press. https://doi.org/10.1201/9781003515470-46

Firmansyah, M. I., Kustiyahningsih, Y., Rahmanita, E., Abidin, M. S., & Satoto, B. D. (2025). Optimization of MSMEs Clustering in Sampang District Using K-Medoids Method and Silhouette Coefficient Method. *Teknika*, *14*(1), 1–8. https://doi.org/10.34148/teknika.v14i1.1116

Gori, T., Sunyoto, A., & Al Fatta, H. (2024). Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(1), 215–224. https://doi.org/10.25126/jtiik.20241118074

Judijanto, L., Asniar, N., Kushariyadi, K., Utami, E. Y., & Telaumbanua, E. (2024). Application of Integrated Logistics Networks in Improving the Efficiency of Distribution and Delivery of Goods in Indonesia a Literature Review. *Sciences Du Nord Economics and Business*, *1*(01), 01–10. https://doi.org/10.58812/sneb.v1i1.6

Lu, J., Yan, Z., Han, J., & Zhang, G. (2019). Data-Driven Decision-Making (D^3 M): Framework, Methodology, and Directions. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *3*(4), 286–296. https://doi.org/10.1109/TETCI.2019.2915813

Masbullah Masbullah & Salmi Yuniar Bahri. (2024). Manajemen Strategi Segmenting, Targeting, Positioning dalam Pemasaran Internasional. *Sammajiva: Jurnal Penelitian Bisnis Dan Manajemen*, *2*(4), 108–118. https://doi.org/10.47861/sammajiva.v2i4.1489

Melovic, B., Rondovic, B., Mitrovic-Veljkovic, S., Ocovaj, S. B., & Dabic, M. (2022). Electronic Customer Relationship Management Assimilation in

Southeastern European Companies—Cluster Analysis. *IEEE Transactions on Engineering Management*, *69*(4), 1081–1100. https://doi.org/10.1109/TEM.2020.2972532

Rahmawati, T., Wilandari, Y., & Kartikasari, P. (2024). Analisis Perbandingan Silhouette Coefficient dan Metode Elbow pada Pengelompokan Provinsi di Indonesia Berdasarkan Indikator IPM dengan K-Medoids. *Jurnal Gaussian*, *13*(1), 13–24. https://doi.org/10.14710/j.gauss.13.1.13-24

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

Unus, E. G. (2024). Teknologi Data Mining Berbasis Metode Clustering Sebagai Ujung Tombak Perkembangan UMKM Di Indonesia Dalam Era Revolusi Industri 4.0. *Jurnal Repositor*, *3*(3). https://doi.org/10.22219/repositor.v3i3.31068

Yosia, & Siregar, B. (2024). Comparative Analysis of K-Means and K-Medoids Algorithms for Product Sales Clustering and Customer. *Journal of Mathematics, Computations and Statistics*, *7*(2), 360–370. https://doi.org/10.35580/jmathcos.v7i2.4053

Yu, H., Chen, L., Yao, J., & Wang, X. (2019). A three-way clustering method based on an improved DBSCAN algorithm. *Physica A: Statistical Mechanics and Its Applications*, *535*, 122289. https://doi.org/10.1016/j.physa.2019.122289

Zheng, K., Huo, X., Jasimuddin, S., Zhang, J. Z., & Battaïa, O. (2023). Logistics distribution optimization: Fuzzy clustering analysis of e-commerce customers' demands. *Computers in Industry*, *151*, 103960. https://doi.org/10.1016/j.compind.2023.103960