Comparison of Elbow and Silhouette Methods in Optimizing K-Prototype Clustering for Customer Transactions

Dendy Arizki Kuswardana[a], Dwi Arman Prasetya[b*], Trimono[c], I Gede Susrama Mas Diyasa[d]

[a,c]*Department of Data Science, UPN "Veteran" Jawa Timur, Surabaya, Indonesia*
[b,d]*Department Master of Information Technology, UPN "Veteran" Jawa Timur, Surabaya, Indonesia*
email: [a]21083010006@student.upnjatim.ac.id, [c]trimono.stat@upnjatim.ac.id, [d]igsusrama.if@upnjatim.ac.id
**corresponding outhor: [b]arman.prasetya.sada@upnjatim.ac.id*

DOI: https://doi.org/10.21107/edutic.v12i1.29744
Diterima: 17 Maret 2025 | Direvisi: 25 April 2025 | Diterbitkan : 04 Mei 2025

Abstract

This study presents a comparative analysis of the Elbow and Silhouette methods to identify the ideal number of clusters in the application of the K-Prototypes algorithm for customer clustering using purchase transaction data. The K-Prototypes algorithm is used because of its ability to handle numeric and categorical data simultaneously. Customer purchase transaction data from the Point of Sale (POS) system is analyzed through preprocessing, feature transformation, and attribute segmentation stages before being clustered using the K-Prototypes algorithm. To identify the optimal number of clusters, this study uses two methods: the Elbow and Silhouette methods. The results show that the Elbow method produces 2 clusters with a model evaluation score of 0.6186, while the Silhouette method produces 2 clusters with the same score of 0.6186. In terms of computational efficiency, the Elbow method also shows faster processing time results, highlighting the importance of choosing the right method to identify the ideal number of clusters, ensuring alignment with the specific objectives of the analysis, whether considering superior cluster differences or prioritizing more efficient model configurations.

Keywords: *Elbow, Silhouette, K-Prototype, Clustering, Customer Transactions*

INTRODUCTION

The development of the culinary industry in today's digital era encourages business actors to not only focus on product innovation but also adopt data-driven approaches to better understand consumer preferences (Maurya, 2024). As customer needs and preferences continue to evolve, businesses in this sector are required to constantly adapt and innovate in order to remain competitive and relevant in the market. Amid intense competition, the comprehensive utilization of transaction data becomes crucial to enable more precise marketing and service strategies (Hindrayani & Timur, 2020). An extensively utilized approach involves segmenting customers using unsupervised learning techniques, with clustering being a primary method.

Specifically, the growth of the culinary sector in Surabaya has shown a significant trend, with the number of business operators increasing from 6.32% in 2013 to 6.64% in 2015 (Ardian & Syairudin, 2018). The Japanese culinary sector has emerged as one of the fastest-growing segments, with an estimated annual growth rate of 10% to 15%, indicating a strong public interest in the distinctive flavors of Japanese cuisine. This trend has consequently spurred the rise of numerous new ventures within the Japanese food industry (Amalijah & Fredy, 2023). Nevertheless, the growing presence of competitors has heightened market competition. As a business entity, XYZ eatery is confronted with considerable challenges in retaining customer loyalty. The dependence on traditional, non-data-driven approaches has proven insufficient, thereby necessitating the adoption of innovative, data-driven strategies to

achieve a deeper and more accurate understanding of consumer purchasing behavior. (S. Dhivya Devi & Usman Ak, 2024).

In addressing these challenges, XYZ eatery seeks to develop data-driven strategies to reinforce its competitive position in an increasingly dynamic market. Through comprehensive analysis of customer purchasing behavior, the business aims to enhance its services to be more personalized and tailored to the needs of its customers. (Ijegwa David Acheme & Esosa Enoyoze, 2024). The appropriate approach in addressing the dynamics of market needs is through customer grouping analysis, which allows for a deeper understanding of the preferences of each customer group (Prasetya et al., 2025). Through this approach, it is possible to design more targeted marketing strategies, offer relevant products for each customer group, and build sustainable customer loyalty (Idhom et al., n.d.). This approach is expected to strengthen competitiveness, particularly amid the rapid growth of the Japanese culinary industry in Surabaya.

The use of the K-Prototype algorithm in customer grouping is becoming increasingly relevant due to its ability to handle mixed data, namely numerical and categorical data, which are commonly found in customer purchase transactions (Girsang, 2020). However, the effectiveness of this clustering process is highly influenced by the determination of the appropriate number of clusters. Selecting an suboptimal number of clusters may lead to inaccurate grouping, thus reducing the strategic value of the analysis (Sipayung et al., 2015). Thus, it is imperative to employ a systematic and empirically measurable approach to correctly identify the ideal number of clusters before implementing the K-Prototype algorithm.

Identifying the most appropriate number of clusters represents a critical stage in the clustering process, as it plays a key role in ensuring the accuracy and meaningfulness of the resulting data groupings. A poor choice in the number of clusters can lead to misleading interpretations and reduce the effectiveness of the overall analysis (Arunachalam et al., 2025). The two commonly used methods are the Elbow Method and Silhouette, each with a different approach. The Elbow Method focuses on the reduction of the total intra-cluster distance, while Silhouette evaluates how well an object is clustered compared to other clusters (Punhani et al., 2022). Although widely used, research that directly addresses the comparison of the effectiveness of these two methods on mixed data with the K-Prototype algorithm is still relatively rare, especially in the context of customer purchase transactions. This is important because an inappropriate method selection can result in unrepresentative clustering and reduce the strategic value of data analysis (Prasetya et al., 2020). This study aims to assess and compare the effectiveness of the Elbow and Silhouette methods in determining the optimal number of clusters for customer groupingusing the K-Prototype algorithm. The data used includes numerical attributes such as transaction amounts and product quantities, as well as categorical attributes such as payment methods, order types, and transaction times This study combines the use of Euclidean distance for numerical data and Hamming distance for categorical data, thereby enhancing the accuracy of clustering based on the specific characteristics of the data. The findings from this comparison aim to offer a clearer, more objective assessment of which method is best suited for clustering mixed data, while also serving as a guide for the creation of more focused and data-driven business strategies (Wara, 2019).

METHOD

This study implements a systematic process prior to clustering using the K-Prototypes algorithm, as illustrated in Fig. 1. Customer purchase transaction data from the period of January to September 2024 is grouped based on purchasing patterns using K-Prototypes, which is capable of handling both numerical and categorical data simultaneously. This algorithm combines the approaches of K-Means and K-Modes with a balancing parameter (γ). The process involves gathering data,

performing preprocessing, conducting feature engineering, calculating distances, determining the ideal number of clusters through the comparison of elbow and silhouette techniques, followed by modeling and evaluating the model.

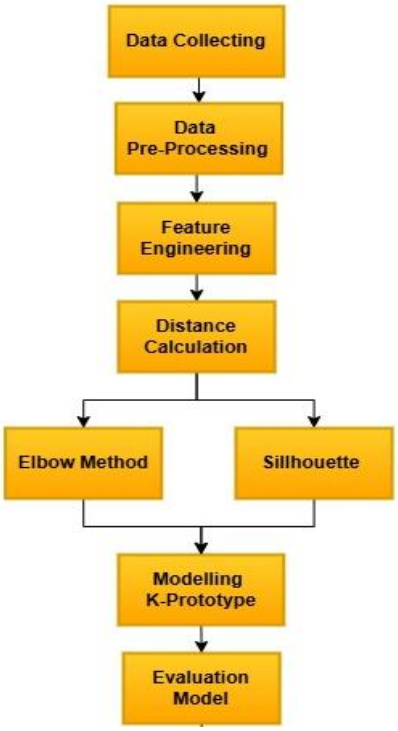

Fig 1. Research Flowchart

The customer purchase history data was retrieved from the Point of Sale (POS) system in Excel format, which automatically records each transaction. Subsequently, the data underwent a pre-processing phase to ensure the accuracy of the clustering process, involving procedures such as handling missing values, removing duplicates, validating data types, and performing feature transformations, including item counting and temporal feature extraction (Riyantoko et al., 2022). The prepared data is divided into numerical attributes (such as Total Sales and Number of Products) and categorical attributes (such as Order Type, Payment Method, and Time). To calculate the distance between data points, Euclidean Distance is used for numerical attributes, while Hamming Distance is applied to categorical attributes. The next step is to identify the ideal number of clusters by utilizing two comparison approaches: the Elbow method and the Silhouette method. These techniques evaluate how accurately an object is grouped within its designated cluster, in comparison to other clusters. The modeling phase utilizes the K-Prototypes algorithm to group customer purchase transactions based on the separated characteristics. The clustering model's results are evaluated using the Silhouette Score, which indicates the quality of the clustering, with higher scores reflecting more optimal groupings.

RESULTS AND DISCUSSION

This study offers a thorough comparison Elbow and Silhouette methods are used to identify the ideal number of clusters when analyzing customer purchase transactions through the K-Prototype algorithm. As illustrated in Figure 2, the Elbow method suggests that the optimal clustering solution is achieved with two clusters. This is evidenced by the most distinct 'elbow' point, where the cost function begins to level off following a steep decline from the first to the third cluster. Beyond this point, the rate

of decrease becomes negligible, thereby supporting the selection of two clusters as the optimal configuration.

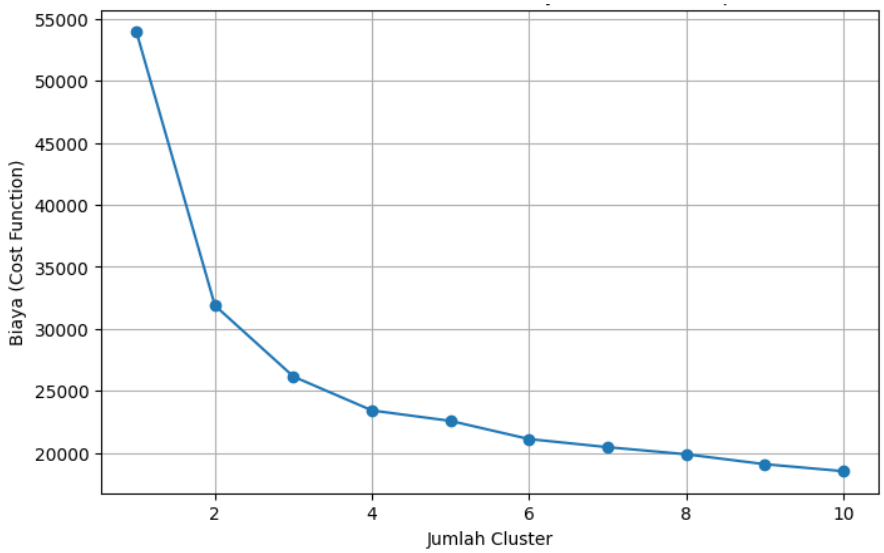

Fig 2. Cluster Determination Using Elbow

As in Fig. 3, The optimal number of clusters is determined by finding the point where the Silhouette score reaches its maximum value achieves its maximum value. This typically occurs when the data is divided into two clusters, providing the most meaningful and distinct grouping. At this point, the silhouette coefficient is approximately 0.45, indicating a clear separation between clusters and strong cohesion within each cluster. A high silhouette score reflects that the data points within a cluster exhibit significant similarity to each other while being clearly distinguishable from the points in other clusters. Therefore, the selection of two clusters is considered the most optimal based on the Silhouette analysis.

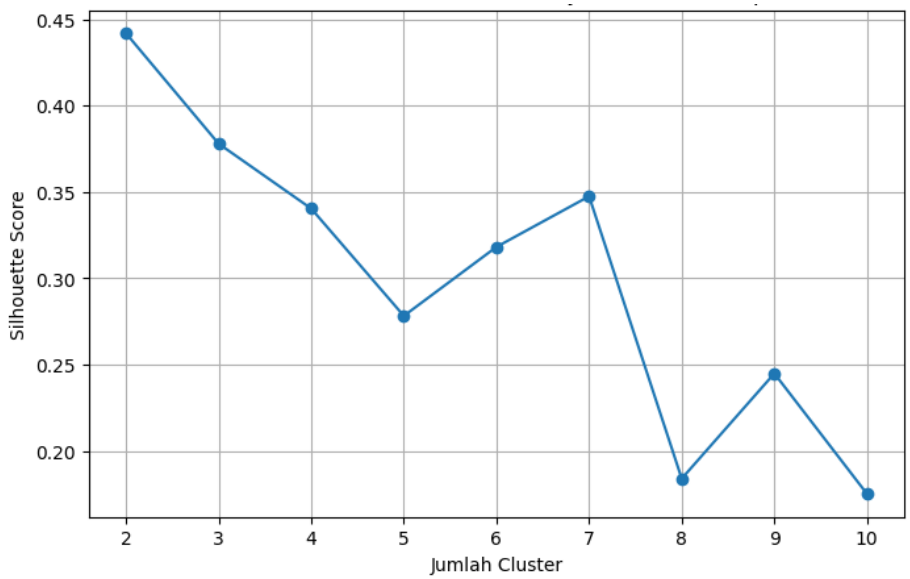

Fig 3. Cluster Determination Using Silhouette

The experimental results in Table evaluate two techniques for identifying the ideal number of clusters: the Elbow Method and the Silhouette Method. These findings reveal a significant difference in

their effectiveness and applicability for clustering tasks. The Elbow Method suggests the use of two clusters, yielding a Silhouette Score of 0.6186, whereas the Silhouette Method recommends two clusters with a slightly lower score of 0.6186. This divergence underscores the methodological differences in how each approach assesses clustering quality. From a computational efficiency perspective, the Elbow Method proves to be more time-efficient, requiring approximately 4 minutes and 42 seconds, as opposed to the Silhouette Method, which takes about 5 minutes and 7 seconds. Although the time disparity is relatively minor, it still indicates that the Elbow Method entails lower computational complexity. Moreover, the higher Silhouette Score associated with the Elbow Method implies that the formation of two clusters offers clearer inter-cluster separation compared to the two cluster configuration.

Table 1. Method Performance Results

Method	Optimal Number of Clusters	Silhouette Score Evaluation	Execution Time
Elbow	2	0.6186	4m 42s
Silhouette	2	0.6186	5m 7s

Based on the experimental results, that choosing an appropriate method to determine the optimal number of clusters requires careful consideration, considering factors such as data characteristics and analysis objectives, to ensure meaningful and accurate clustering outcomes. If the main focus is on the quality of data separation and the clarity of the division between clusters, then the Elbow Method approach with 2 cluster recommendations can be considered a superior choice, especially since the elbow method only requires a faster execution time with the same accurate model evaluation results. On the other hand, if the analysis goal is more oriented towards simplifying the model structure or operational needs that are lighter, then the 2 cluster option obtained through the Silhouette method remains relevant, given that the evaluation values are still within an acceptable range.

CONCLUSIONS

The aim of this study assesses the Elbow and Silhouette Methods to identify the ideal number of clusters when applying the K-Prototype algorithm to customer purchase transaction data, aiming to assess the effectiveness of each method in identifying the most suitable clustering solution. The results indicate that the Elbow Method recommends two clusters as the optimal number, characterized by a significant reduction in the cost function and supported by a Silhouette value of 0. 6186. In contrast, the Silhouette Method suggests the use of two clusters, with a Silhouette value of 0.6186 representing its highest point.

From the perspective of time efficiency, the Elbow Method shows a faster computational process compared to the Silhouette Method. Although there is no difference in the Silhouette Score evaluation results, these results highlight the importance of adjusting the process of determining the ideal number of clusters to meet specific analysis objectives. In conclusion, these results underline the importance of considering factors such as cluster quality, efficiency, and practical requirements in selecting the most appropriate customer clustering strategy.

REFERENCES

Amalijah, E., & Fredy, M. (2023). Pemetaan Restoran Jepang dan Kuliner Milenial di Surabaya. *Jurnal Sakura : Sastra, Bahasa, Kebudayaan dan Pranata Jepang*, *5*(1), 169. https://doi.org/10.24843/JS.2023.v05.i01.p10

Ardian, S., & Syairudin, B. (2018). Development strategy of culinary business employing the Blue Ocean Strategy (BOS). *IPTEK Journal of Proceedings Series*, *0*(3), 153. https://doi.org/10.12962/j23546026.y2018i3.3722

Arunachalam, M., Sekar, S., Erdmann, A. M., Sajith Variyar, V. V., & Sivanpillai, R. (2025). Comparative Analysis of Machine Learning Algorithms and Statistical Techniques for Data Analysis in Crop Growth Monitoring with NDVI. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVIII-M-5–2024*, 15–20. https://doi.org/10.5194/isprs-archives-XLVIII-M-5-2024-15-2025

Girsang, A. S. (2020). Clustering Hostels Data for Customer Preferences using K-Prototype Algorithm. *International Journal of Emerging Trends in Engineering Research*, *8*(6), 2650–2653. https://doi.org/10.30534/ijeter/2020/70862020

Hindrayani, K. M., & Timur, J. (2020). *Business Intelligence For Educational Institution: A Literature Review*. *2*(1). https://doi.org/10.33005/ijconsist.v2i1.32

Idhom, M., Priananda, A. M., Raynaldi, A., Nur, R., Pamungkas, S. A., & Wardana, A. C. (n.d.). *UPAYA REBRANDING SEBAGAI BENTUK KEPEDULIAN TERHADAP UMKM*. *2*(4). https://doi.org/10.56855/jcos.v2i4.1112

Ijegwa David Acheme & Esosa Enoyoze. (2024). Customer personality analysis and clustering for targeted marketing. *International Journal of Science and Research Archive*, *12*(1), 3048–3057. https://doi.org/10.30574/ijsra.2024.12.1.1003

Maurya, N. K. (2024). Decoding Consumer Dynamics: A Deep Dive into Food Industry Surveys and Trends. *Nutrition and Food Processing*, *07*(14), 01–06. https://doi.org/10.31579/2637-8914/275

Prasetya, D. A., Nguyen, P. T., Faizullin, R., Iswanto, I., & Armay, F. (2020). Resolving the Shortest Path Problem using the Haversine Algorithm. *Journal of Critical Reviews*, *7*(1). http://10.22159/jcr.07.01.11

Prasetya, D. A., Sari, A. P., Idhom, M., & Lisanthoni, A. (2025). Optimizing Clustering Analysis to Identify High-Potential Markets for Indonesian Tuber Exports. *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *7*(1), 113–122. https://doi.org/10.35882/ijeeemi.v7i1.55

Punhani, A., Faujdar, N., Mishra, K. K., & Subramanian, M. (2022). Binning-Based Silhouette Approach to Find the Optimal Cluster Using K-Means. *IEEE Access*, *10*, 115025–115032. https://doi.org/10.1109/ACCESS.2022.3215568

Riyantoko, P. A., Fahrudin, T. M., Prasetya, D. A., Trimono, T., & Timur, T. D. (2022). Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data Spam dan Non-Spam. *PROSIDING SEMINAR NASIONAL SAINS DATA*, *2*(1), 103–111. https://doi.org/10.33005/senada.v2i1.53

S. Dhivya Devi, A. V. B., G. Lakshmi, & Usman Ak, S. B., Syed Shujauddin Sameer,. (2024). Data-Driven Decision-Making: Leveraging Analytics for Performance Improvement. *Journal of Informatics Education and Research*, *4*(3). https://doi.org/10.52783/jier.v4i3.1298

Sipayung, E. M., Fiarni, C., & Tanudjaya, R. (2015). DECISION SUPPORT SYSTEM FOR POTENTIAL SALES AREA OF PRODUCT MARKETING USING CLASSIFICATION AND CLUSTERING METHODS. *Proceeding 8 Th International Seminar on Industrial Engineering and Management*, 33–39.

Wara, S. S. M. (2019). *ANALISIS RESPONS WARGANET TERHADAP DEBAT CALON PRESIDEN 2019 DI TWITTER DENGAN METODE CLUSTERED SUPPORT VECTOR MACHINES* [INSTITUT TEKNOLOGI SEPULUH NOPEMBER]. https://repository.its.ac.id/64282/1/06211540000101_Undergraduate_Thesis.pdf