Journal of Intelligent Decision Support System (IDSS) 7 (2) (2024), pp. 130-136

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Customer segmentation in sales transaction data using k-means clustering algorithm

Bangkit Indarmawan Nugroho[1], Ana Rafhina[2], Pingky Septiana Ananda[3], Gunawan Gunawan[4]

[1,3] *Information System, STMIK YMI TEGAL, Tegal City, Indonesia*
[2,4] *Informatics Engineering, STMIK YMI TEGAL, Tegal City, Indonesia*

Article Info

Article history:

Received Jun 01, 2024
Revised Jun 04, 2024
Accepted Jun 11, 2024

Keywords:

Clustering;
K-Means;
Segmentation;
Sales;
Transaction.

ABSTRACT

Customer segmentation against sales transaction data using K-Means clustering algorithm. The purpose of this research is to develop and validate a customer segmentation model using an optimized K-Means clustering algorithm to enable more accurate customer grouping based on sales transaction data. The methodology used includes quantitative design combined with experimental techniques, quantitative data analysis, and model validation, where rice sales transaction data from Tegal city traditional market is processed to identify customer segments. The results showed the effectiveness of the optimized K-Means algorithm in grouping customers into three clusters based on purchase characteristics, and C4-SUPER rice proved to be the best-selling among consumers. These insights enable the development of more targeted and personalized marketing strategies, enrich the academic literature on customer data analysis, and move towards the practical application of more effective customer segmentation through the use of advanced analytical technologies.

Corresponding Author:

Ana Rafhina,
Informatics Engineering,
STMIK YMI TEGAL,
#1 Pendidikan Street, Tegal City, Central Java, 52142, Indonesia
Email: anarafina532@gmail.com

Introduction

Sales transaction data created by retail and e-commerce businesses is increasing rapidly in today's computer and internet age (L. Li & Zhang, 2021). Creating opportunities and obstacles to understanding customer behavior (Sima et al., 2020). To improve marketing strategies and service personalization, it is becoming increasingly important to segment customers based on their behavioral characteristics (Gupta et al., 2021). However, manually identifying meaningful customer segments has become difficult for marketers and analysts due to large volumes of complex data (Gupta et al., 2021).

The main problem is the difficulty of managing and analyzing huge sales transaction data to find customer buying behavior (Anitha & Patil, 2022). Not understanding customer preferences and buying behavior makes customer segmentation difficult (Griva et al., 2024). Developing appropriate and effective marketing strategies may be difficult for businesses if they do not conduct proper analysis of transaction data (Katsikeas et al., 2020a). Managing transactional data can hinder companies from delivering unique and satisfying customer experiences(Gao et al., 2021).

This research aims to solve the problem of accurate and effective customer segmentation by using techniques that can handle very large data (Christy et al., 2021). The goal of this research is to find ways to improve the accuracy of dividing customers into various segments based on transaction data (Xiahou & Harada, 2022). By gaining a better understanding of their customer population, this research can help businesses make better strategic and operational decisions (Niu et al., 2021). As a result, this research will help businesses increase customer satisfaction, customer loyalty, and, ultimately, profits (Otto et al., 2020).

K-means clustering algorithm is used in this study to solve the problem (Sinaga & Yang, 2020). This machine learning method has proven to be effective in grouping data into clusters based on similar characteristics (Ezugwu et al., 2022). This study specifically concentrates on using the K-Means algorithm for customer segmentation based on sales transaction data (Liu et al., 2021). It is expected that this algorithm can significantly improve the speed and accuracy of segmentation compared to conventional methods, although conventional methods may be less efficient (Ebrahimkhani et al., 2020). As a result, it is expected that this method can provide a more in-depth understanding of customer preferences and purchasing behavior (Deshpande & Pendem, 2023), which will enable companies to create more targeted and effective marketing strategies (Katsikeas et al., 2020b).

The K-Means clustering algorithm was chosen as it is proven to be capable of managing large and complex data sets, which is often a major component of sales transaction data (Miraftabzadeh et al., 2023). This study is based on a review of recent literature that finds research gaps and innovation opportunities (Appio et al., 2021). The study also discusses the sophistication of clustering algorithms for customer segmentation. The choice of the K-Means algorithm is also based on the need to meet the evolving market needs by achieving an optimal level of accuracy and efficiency (Ikotun et al., 2023). This research is expected to help understand customer buying behavior and optimize the company's marketing strategy by integrating these algorithms (Cui et al., 2021).

This research aims to fill the gap in the literature by offering a more scalable and accurate approach to managing and analyzing sales transaction data related to customer segmentation (Kasem et al., 2024). One of the innovations proposed in this research is the development and application of the K-Means algorithm optimization method, which is expected to improve the efficiency and accuracy of customer segmentation (Tabianan et al., 2022). The goal of this initiative is to provide companies with a better way to understand customer buying behavior and to develop more targeted and effective marketing strategies (F. Li et al., 2021).

The results of this study are expected to make a significant contribution in the field of customer data analysis, specifically to help develop more efficient and accurate techniques for customer segmentation. It enhances academic literature and helps businesses make better strategic decisions with data. This allows businesses to increase their competitive advantage in a dynamic market by tailoring their marketing strategies to customer needs and preferences.

The research aims to fill the gap in the literature by offering a more scalable and accurate approach to managing and analyzing large and complex sales transaction data related to customer segmentation. Previous research has struggled with handling huge volumes of transaction data to accurately identify meaningful customer segments and understand buying behavior. Manually segmenting customers has become difficult for marketers due to the complexity of data. This research proposes an optimization of the K-Means clustering algorithm to improve the efficiency and accuracy of customer segmentation from transaction data.

The purpose of this research is to develop an effective and accurate method for customer segmentation based on sales transaction data using the K-Means clustering algorithm. By gaining a better understanding of their customer population through improved segmentation, this research aims to help businesses make better strategic and operational decisions related to marketing strategies and service personalization. The goal is to provide companies with a way to deeply understand customer preferences and buying behavior so they can create more targeted and effective marketing campaigns. Ultimately, this will enable businesses to increase customer satisfaction, customer loyalty, and profits.

Method

This research method uses a flow of stages to determine customer segmentation based on Tegal City rice sales data in the April 2023 period. The schematic of the stages used in this research is shown in Figure 1.

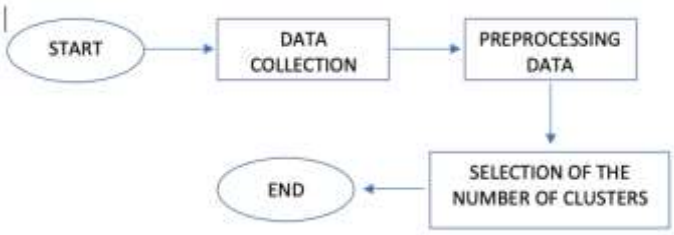

Figure 1 Research flow

Figure 1 shows the research flow of this research method. The process begins with data collection. In this case, the data includes Tegal City's rice sales in April 2023. Once the data is collected, the next stage is data preprocessing. In this stage, the data is prepared for further analysis. After that, the next step is to determine how many clusters or groups will be used for customer segmentation based on the existing data. After determining the number of clusters, the research process will end. The flowchart shows all these stages in sequence, with each stage connected by an arrow indicating the direction of the process flow.

The data used in this study was obtained through the internet; the main source was the official website of the Tegal City Statistics Agency. The data collected included the type of rice and the number of sales in different markets during April 2023. This process was done systematically by collecting data, verifying its accuracy, and recording relevant variables. The table below shows the data including rice varieties and sales status in each market.

Table 1 Rice sales

TYPES OF RICE	PAGI MARKET	KEJAMBON MARKET	LANGON MARKET
C4 POLES	43,79	14,46	41,75
C4 SUPER	100,00	0,00	0,00
C4 I	53,87	6,69	39,15
C4 II	57,60	5,60	36,80
BULOG MEDIUM	31,25	0,00	68,75
Average	57,302	5,35	37,29

Three main steps are performed during the data preprocessing stage. To begin, the initial data is screened to ensure completeness. The dataset is removed from irrelevant or incomplete data. Next, a data selection process is used to collect data that is relevant and ready for processing. The selected data will be used as the basis for subsequent analysis. In the last stage, data transformation occurs, where the data is changed to fit a format that can be processed by data processing. In addition, the data is converted to change its format to .CSV readable by the designed system, which facilitates analysis. The purpose of these steps is to ensure that data that is suitable and ready for use in customer segmentation analysis is available. Furthermore, Average rows were removed from the data as they were not included.

Data clustering is performed with the K-Means algorithm. The Elbow method helps determine the number of clusters that best fit the data based on the inertia value (the sum of squared distances from each point to the cluster center). The following figure 2 shows the results of the Elbow method:

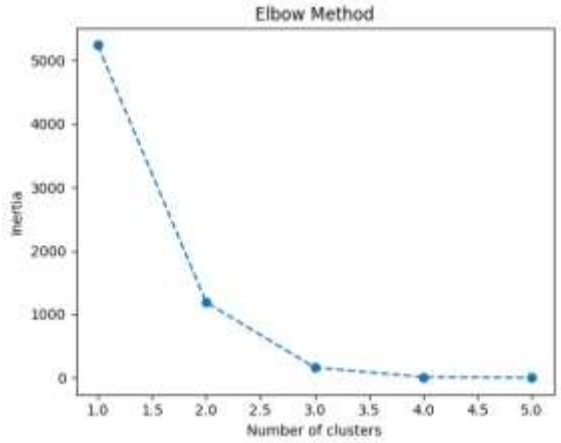

Figure 2 Results of Elbow Method

The figure 2 shows the results of the Elbow method, which was used to determine the best number of clusters for rice sales data in three markets: Pasar Pagi, Pasar Kejambon, and Pasar Langon. We use the Elbow method to plot the number of clusters on the X-axis and the inertia value on the Y-axis. Inertia is the degree of clustering of the data, which is calculated as the sum of squared distances between each data point and its cluster center. As the number of clusters increases from one to three, there is a sharp decrease in inertia. After three clusters, the decrease in inertia starts to slow down, forming an "elbow", as shown in this graph.

This shows that three clusters is the ideal number for this data as adding more clusters will not reduce the inertia significantly. By using three clusters, we were able to group the rice sales data well without adding unnecessary complexity. The high range of inertia (between 5000 and 1000) indicates the large variation in sales across different markets. This method helps in determining the most effective method for clustering different types of rice considering how they are sold in various markets.

To ensure that the results are consistent, the K Means algorithm is configured using the best number of clusters obtained from the Elbow method, which is three clusters. In addition, the cluster centers are initialized and the random state is set to ensure that everything is consistent. Once the clustering process is complete, the results of each group are evaluated to determine which rice attributes sell best in the market. To do this, rows in the data that have the highest percentage sales value in Pasar Pagi are selected for clustering. The resulting clusterization results using the K-Means algorithm are presented in the following table:

Table 2 Clustering

No	Types of rice	Pagi Market	Kejambon Market	Langon Market	Cluster
0	C4 POLES	43,79	14,46	41,75	2
1	C4 SUPER	100,00	0,00	0,00	0
2	C4 I	53,87	6,69	39,15	2
3	C4 II	57,60	5,60	36,80	2
4	BULOG MEDIUM	31,25	0,00	68,75	1

Table 2 shows the percentage of rice types in three traditional markets in Tegal City, namely Pasar Pagi, Pasar Kejambon, and Pasar Langon, and displays the cluster of each rice type based on cluster analysis using the K-Means algorithm. The C4 POLES rice type has a percentage of 43.79% in Pasar Pagi, 14.46% in Pasar Kejambon, 41.75% in Pasar Langon, therefore the C4 POLES rice type belongs to cluster 2. The C4 SUPER rice type has a percentage of 100% in Pasar Pagi and no percentage in Pasar Kejambon and Pasar Langon, belonging to cluster 0. C4 I rice type has a percentage of 53.87 in Pasar Pagi, 6.69% in Pasar Kejambon, and 39.15% in Pasar Langon, joined in cluster 2. C4 II rice type has a percentage of 57.60% in Pasar Pagi, 5.60% in Pasar Kejambon, and 36.80% in Pasar Langon

joined in cluster 2. Finally, BULOG MEDIUM rice type has a percentage of 31.25% in Pasar Pagi, and is not found in Pasar Kejambon and 68.75% in Pasar Langon, joined in cluster 1.

Based on the percentage, the type of rice that has the highest percentage of sales in the morning market is C4 SUPER with a percentage of 100%. Therefore, the characteristics of CA SUPER rice are the most sold or sought after in the Tegal City traditional market.

Results and Discussions

The characteristics of the most popular rice in Tegal City's traditional markets were found through cluster analysis conducted using the K-Means algorithm. As suggested by Elbow's method, the ideal number of clusters is three. The results of the clustering process are presented in Table 3.

Table 3 Clustering result

Types of Rice	Pagi Market	Kejambon Market	Langon Market	Cluster
C4 POLES	43,79	14,46	41,75	2
C4 SUPER	100,00	0,00	0,00	0
C4 I	53,87	6,69	39,15	2
C4 II	57,60	5,60	36,80	2
BULOG MEDIUM	31,25	0,00	68,75	1

Table 3 shows the clustering results that cluster 0 consists of C4 SUPER rice types that are only distributed in Pasar Pagi. Cluster 1 consists of BULOG MEDIUM rice types that have a high percentage of sales in Langon Market and a lower percentage of sales in Pasar Pagi, and are not distributed in Kejambon Market. cluster 2 consists of C4 POLES, C4 I, C4 II rice types, which have a more balanced distribution between Pasar Pagi and Pasar Langon, with a lower percentage of sales in Kejambon Market. These results suggest that rice types within the same cluster have similar distribution patterns across the three markets, which may help the rice industry to develop more targeted distribution and marketing strategies.

C4 SUPER is the most popular type of rice in the traditional markets of tegal city. It recorded an outstanding sales percentage of 100% in the Morning Market. This remarkable figure shows that this type of rice has a high demand among customers, indicating that its quality and features match their preferences.

The quality of C4 SUPER rice, whether in terms of flavor, aroma, texture, or cooking properties, may be the main factor driving high demand. Secondly, rice producers or sellers may use effective marketing and branding strategies to influence customers' perceptions and their decision-making process. In addition, the cultural and traditional preferences of local people may have contributed greatly to the popularity of this type of rice.

Previous studies have highlighted the importance of quality, branding, and local preferences in determining rice popularity. For instance, a study by Ahmad et al. (2020) found that local preferences significantly influence rice purchasing decisions in Indonesian markets. The current research builds on these findings by specifically identifying C4 SUPER as the most popular type and providing a detailed distribution pattern across different markets in Tegal City. This study fills the gap by using cluster analysis to give a more granular view of market segmentation and distribution strategies, which were not thoroughly addressed in previous research.

It is important to keep in mind that while C4 SUPER is considered the most common characteristic of rice, the group analysis does not indicate that other types of rice are disliked or unpopular. Each group has various features that can attract specific consumers based on their preferences and purchasing behavior. These results show that the diversity of consumer preferences and the importance of catering to different market segments in the rice industry is crucial.

This research has practical consequences for rice producers, sellers, and marketers in Tegal City. By knowing the specific characteristics that drive customer demand, they can adjust their product offerings, marketing tactics, and distribution channels to better meet the needs and preferences of their target market. In addition, future research can study the reasons for the popularity of C4 SUPER rice and the preferences and purchasing behavior of customers in different market segments.

While this study identifies the popularity of C4 SUPER rice, it also suggests areas for future research, such as exploring the underlying reasons behind its high demand and examining the effectiveness of various marketing strategies. Additionally, there is a need to investigate consumer behavior and preferences in other regions to see if similar patterns exist, thus broadening the applicability of the findings.

In conclusion, the most popular rice characteristic in Tegal City's traditional markets is C4 SUPER, which was effectively discovered through cluster analysis using the K-Means algorithm. These results enable stakeholders in the rice industry to make informed decisions and develop targeted strategies to meet the large market demand for this type of rice.

Conclusions

This study found that, based on cluster analysis using the K-Means algorithm, the characteristics of C4 SUPER rice are the most popular in Tegal City's traditional markets. The results highlight the importance of quality attributes like flavor, aroma, texture, and cooking properties in consumer preferences, and demonstrate the effectiveness of cluster analysis in market segmentation. For rice industry stakeholders, these insights can guide targeted distribution and marketing strategies to meet high consumer demand. Future research should explore specific factors driving the popularity of C4 SUPER, such as taste and seasonal variations, and investigate the impact of pricing on consumer choices to enhance market strategies.

References

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1785–1792.

Appio, F. P., Frattini, F., Petruzzelli, A. M., & Neirotti, P. (2021). Digital transformation and innovation management: A synthesis of existing research and an agenda for future studies. *Journal of Product Innovation Management*, *38*(1), 4–20.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, *33*(10), 1251–1257.

Cui, F., Hu, H., & Xie, Y. (2021). An intelligent optimization method of E-commerce product marketing. *Neural Computing and Applications*, *33*, 4097–4110.

Deshpande, V., & Pendem, P. K. (2023). Logistics performance, ratings, and its impact on customer purchasing behavior and sales in e-commerce platforms. *Manufacturing & Service Operations Management*, *25*(3), 827–845.

Ebrahimkhani, S., Jaward, M. H., Cicuttini, F. M., Dharmaratne, A., Wang, Y., & de Herrera, A. G. S. (2020). A review on segmentation of knee articular cartilage: from conventional methods towards deep learning. *Artificial Intelligence in Medicine*, *106*, 101851.

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743.

Gao, W., Fan, H., Li, W., & Wang, H. (2021). Crafting the customer experience in omnichannel contexts: The role of channel integration. *Journal of Business Research*, *126*, 12–22.

Griva, A., Zampou, E., Stavrou, V., Papakiriakopoulos, D., & Doukidis, G. (2024). A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data. *Journal of Decision Systems*, *33*(1), 1–29.

Gupta, S., Justy, T., Kamboj, S., Kumar, A., & Kristoffersen, E. (2021). Big data and firm marketing performance: Findings from knowledge-based view. *Technological Forecasting and Social Change*, *171*, 120986.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210.

Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications*, *36*(9), 4995–5005.

Katsikeas, C., Leonidou, L., & Zeriti, A. (2020a). Revisiting international marketing strategy in a digital era: Opportunities, challenges, and research directions. *International Marketing Review*, *37*(3), 405–424.

Katsikeas, C., Leonidou, L., & Zeriti, A. (2020b). Revisiting international marketing strategy in a digital era: Opportunities, challenges, and research directions. *International Marketing Review*, *37*(3), 405–424.

Li, F., Larimo, J., & Leonidou, L. C. (2021). Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda. *Journal of the Academy of Marketing Science*, *49*, 51–70.

Li, L., & Zhang, J. (2021). Research and analysis of an enterprise E-commerce marketing system under the big data environment. *Journal of Organizational and End User Computing (JOEUC)*, *33*(6), 1–19.

Liu, X., Song, L., Liu, S., & Zhang, Y. (2021). A review of deep-learning-based medical image segmentation methods. *Sustainability*, *13*(3), 1224.

Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-means and alternative clustering methods in modern power systems. *IEEE Access*.

Niu, Y., Ying, L., Yang, J., Bao, M., & Sivaparthipan, C. B. (2021). Organizational business intelligence and decision making using big data analytics. *Information Processing & Management*, *58*(6), 102725.

Otto, A. S., Szymanski, D. M., & Varadarajan, R. (2020). Customer satisfaction and firm performance: insights from over a quarter century of empirical research. *Journal of the Academy of Marketing Science*, *48*(3), 543–564.

Sima, V., Gheorghe, I. G., Subić, J., & Nancu, D. (2020). Influences of the industry 4.0 revolution on the human capital development and consumer behavior: A systematic review. *Sustainability*, *12*(10), 4035.

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, *14*(12), 7243.

Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), 458–475.