Scientific Journal of Informatics
Vol. 11, No. 2, May 2024

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Optimizing Customer Segmentation in Online Retail Transactions through the Implementation of the K-Means Clustering Algorithm

Desi Adrianti Awaliyah[1*], Budi Prasetiyo[2], Rini Muzayanah[3], Apri Dwi Lestari[4]
[1, 2, 3, 4]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: The main objective of this research is optimal use of customer segmentation using the Recency, Frequency and Monetary (RFM) approach so that companies can better understand and comprehend the needs of each customer. By carrying out this segmentation, companies can communicate better and provide services tailored to each customer.
Methods: The K-means algorithm is used as the main method for customer segmentation in this research. This research uses a dataset of online retail customers. Apart from that, this research also uses the elbow method to help determine the best number of clusters to be created by the model.
Result: Based on the elbow method, the most optimal is to use 3 clusters for this case. Thus, in K-means modeling, forming 3 clusters is the best choice. Clusters produce groups of customers who have specific characteristics in each cluster. The analysis shows that quantity and unit price have a significant influence on online retail customer behavior.
Novelty: This research strengthens the trend of using the K-means algorithm for customer segmentation in online retail datasets, which has proven popular in journals from 2018 to 2022. This research creates 3 new variables that will be used by the model to understand the characteristics of customer transaction behavior. This study also emphasizes the importance of exploratory data analysis in understanding data before clustering and the use of the elbow method to determine the most appropriate number of clusters, providing a significant contribution in analyzing customer segmentation.

Keywords: Clustering, K-means, Online retail, Python, Customer segmentation
Received May 2024 / **Revised** June 2024 / **Accepted** June 2024

INTRODUCTION

In today's digital era, the online retail sector has become one of the fields that has experienced significant development. fast and dynamic. This rapid growth is also accompanied by increasing competition in the market. In this process, potential buyers compare the perceived quality of a product with its perceived sacrifice, i.e. price. An increasing increase in price is associated with a decrease in purchase probability [1]. Customer satisfaction is the key to business success, retailers need to implement a more customer-centric approach and look for innovative ways to understand customers and satisfy them [2], [3]. One way this can be done is by analyzing customer shopping behavior patterns. Once customer behavior is analyzed, business managers can provide services that match the shopping needs and shopping preferences of customers.

One method that can be implemented to analyze customer behavior is customer segmentation. Segmentation is a method in which information about customer data and behavior is interpreted and used as potential to create new business opportunities [4]. In marketing strategies, segmentation has become a major focus and is often applied [5]. In the context of online retail transactions, rich transaction data provides an opportunity to apply sophisticated customer segmentation techniques. In retail businesses, excessive merchandise purchases may result in excessive end-of-season inventory leading to low profit returns. In addition, the merchandise that retailers carry is constrained by a limited budget for each product, limited shelf space or storage space to display merchandise, and a limited number of suppliers capable of offering competitive products [6]. Therefore, it is necessary to segment customers to prepare the best strategy so that business managers can make profits and avoid losses.

[*]Corresponding author.
Email addresses: 1desiawaliyahhh@students.unnes.ac.id (Awaliyah)[*], bprasetiyo@mail.unnes.ac.id. (prasetiyo)
DOI: 10.15294/sji.v11i2.6137

Significantly, existing clustering methods can be broadly categorized into five groups, including K-means-type clustering [7]–[9], kernel-based clustering [10], graph-based clustering [11], spectral clustering [12], [13], and others. One of the popular methods for customer segmentation is using the K-Means Clustering algorithm. This algorithm allows grouping customers. By taking into account the buying patterns they exhibit, consumers are divided into different segments. Clustering is a frequently used strategy in data mining, pattern identification, significant information retrieval, and text mining where an entity, record, or document is grouped into clusters with similar or related content [14][15]. By applying K-Means Clustering to retail online transaction data, retailers can identify groups of customers who share similar characteristics, allowing them to customize marketing strategies and product offerings more effectively.

Several previous studies have discussed customer segmentation. One of them is a study conducted by Calvo-Porral and Lévy-Mangin in 2018 [16]. This study provides customer categorization in a comprehensive segmentation-based food retail business. The study implemented the Multivariate Analysis of Variance (MANOVA) technique in analyzing customer segments. The study reported that specialty food retail customers cannot be seen as a homogeneous group. There were a total of four types of clusters identified in the food retail business dataset used. Another study was conducted by Li et al. In 2021 [17]. The research discusses customer segmentation using the K-Means algorithm and also Particle Swarm Optimization. The research focused on customer segmentation on a dataset of wine customer consumption. The experiment shows the validity and rationality of the customer segmentation method.

After going through the literature stage, the K-Means algorithm has good performance in segmenting customers. However, there is no research that discusses customer segmentation using K-Means on online retail datasets. This research will study the use of K-Means clustering algorithm to improve customer segmentation in the context of online retail transactions [18]. This research will be based on the variables of ecency, frequency, and monetary (RFM) so that it can produce clusters by analyzing transactions made by customers.

METHODS

The research process will be explained in this section. Figure 1 illustrates the sequence of steps in the research method.

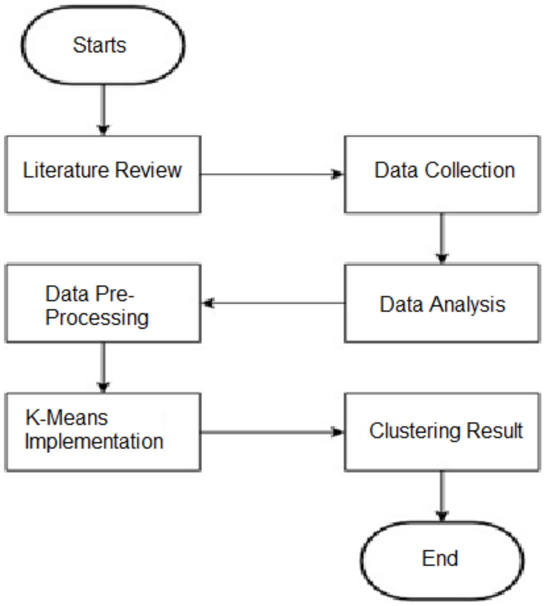

Figure 1. Research method

The steps executed in this research are in line with the sequence represented in the research methodology flow diagram.

Data collection

The dataset used in this research is an online retail customer dataset on the Kaggle dataset platform. The dataset can be accessed at https://www.kaggle.com/datasets/yasserh/customer-segmentation-dataset. Online retail transaction data, which has 54,910 entries and 8 columns, will be used in the data collection stage because it is relevant to this research. The columns consist of 8 variables, namely Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, and Country.

Preprocessing data

Data Pre-Processing is an important stage in data processing that involves cleaning, transforming, and preparing raw data for further analysis. Data processing is presented by grouping the methods into different sub-classes according to the type of data used to operate on [19]. This stage includes handling missing data, removing duplicate data, normalizing data, as well as transforming data into a format suitable for the analysis method to be applied. At the data preprocessing stage, there are several processes that will be carried out. The processes that will be undertaken include the handling of missing values, the handling of outlier data, the transformation of data, and the engineering of features.

Handling missing value

Handling missing values involves identifying empty data, then deleting variables with missing values or replacing the missing values with the mean, median, or mode of the data. Handling missing values is included in the preprocessing steps [20]. Missing values are the most common problem most commonly occurring in data missing values are randomly scattered throughout the data matrix [21]. Handling Missing Value plays a role in managing missing values in the dataset by performing certain actions. This process aims to ensure that all rows of data have values, so that the model does not need to learn empty data later. Apart from that, the presence of empty data can also prevent the model from learning optimally because it learns empty data.

Handling outlier data

Handle Outliers is the process of recognizing and handling data that is significantly different from the general pattern in a dataset. Outliers are data that stand out and show significant differences from the rest of the data, or data that do not follow the usual pattern or distribution of the data [22]. This can include removing data that is outside the normal limits, transforming the data, or using specialized statistical methods to treat outliers according to their context. The goal is to verify that the data used for analysis or modeling is more accurate and representative.

Data transformation

Data Transformation involves the process of changing the scale or shape of data in a dataset. It may involve logarithmic transformation, normalization, or standardization of data to make the data distribution more suitable to the requirements of statistical analysis or to improve the performance of modeling algorithms. Normalization techniques play a crucial role in accelerating the training process [23]. The goal is to improve the interpretation of the analysis results and enhance the quality of the models created using the transformed data.

Feature engineering

Feature engineering is a very important step, and it involves transformation, creation, extraction, and selection of features by utilizing raw data or existing features [24]. In this research, feature engineering is done by creating new variables, namely recency, frequency, and monetary variables. Recency measures how long it has been since a customer last made a purchase, with the assumption that customers who have recently transacted are more likely to respond to marketing promotions. Frequency measures how often a customer makes a purchase in a given period, indicating their level of loyalty. Monetary value measures the total money spent by the customer in the period, identifying the most valuable customers to the business.

Implementation of k-means algorithm and elbow method

K-Means algorithm, which is well-known as one of the clustering methods that can be used to identify customer segmentation results through the cluster division process [25]. The K-means algorithm is particularly prevalent in practical applications due to its simplicity and computational efficiency [26]. The objective of the K-means clustering model is to partition the data space into discrete zones and to determine the appropriate zone to which each data point should be allocated [27].

The elbow method is used to determine the most optimal number of clusters or k value from the dataset at hand. This stage is important to carry out so that the model can understand the data optimally and the mode can form the right clusters. Once the optimal k value has been identified, the k-means clustering method should be employed [28].

Clustering result

The K-Means method, as one of the well-known clustering algorithms, can be utilized to identify customer segmentation results through the process of cluster division [29]. Customers' contribution to sales is divided into segments based on their purchasing behavior as a result of this clustering.

RESULTS AND DISCUSSIONS

Study literature

This research begins with a literature study stage carried out through a literature search process that aims to identify sources that are relevant to the object of research. The literature search was conducted through various databases such as Springer, IEEE Explore, and Google Scholar, focusing on publications that are available for download and publicly accessible. In data analysis, the K-Means algorithm remains the top choice for many researchers, supported by findings from previous studies showing that the K-Means method is often used, especially in the context of data segmentation, to cluster data into more defined groups. In addition, the elbow method is also often used to establish the optimal number of clusters, allowing researchers to identify the most appropriate number of clusters for the data being analyzed.

Based on the literature review, the K-Means algorithm is proven to provide stable and reliable results in various applications, including in the context of segmentation. In addition, previous studies have also shown that the use of K-Means together with the elbow method is highly advantageous in the process of thorough data analysis and clustering. The elbow method in particular assists researchers in identifying crucial parameters, such as the optimal number of suitable clusters, thus improving the efficiency and effectiveness of data analysis.

Data collection

In this research, an online retail transaction dataset has been used which consists of 54,910 entries with 8 columns. The dataset forms the basis for the analysis and research conducted. The variables that will be utilized in this study are outlined and detailed in Table 1.

Table 1. Online retail dataset used in research

Invoice No	Stock Code	Description	Quantity	Invoice Date	Unit Price	Customer ID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2,55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2,75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 08:26	3,39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01/12/2010 08:26	7,65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01/12/2010 08:26	4,25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	01/12/2010 08:28	1,85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	01/12/2010 08:28	1,85	17850	United Kingdom
….	….	….	….	….	….	….	….

Data Preprocessing

Data transformation

The data from the pool is processed to prepare for further analysis at the data analysis and pre-processing stage. At this stage, the CustomerID variable which has a float data type is converted into a variable with a string data type. This is because the numeric value in Customer ID does not have a specific value, the variable is only a distinction between customers. The type of dataset before and after data transformation can be seen in Figure 4 and Figure 5.

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
```

Figure 4. Data type before data transformation process

```
 InvoiceNo     406829 non-null object
 StockCode     406829 non-null object
 Description   406829 non-null object
 Quantity      406829 non-null int64
 InvoiceDate   406829 non-null object
 UnitPrice     406829 non-null float64
 CustomerID    406829 non-null object
 Country       406829 non-null object
```

Figure 5. Data type after data transformation process

Handling missing value

The next stage in data exploration involves checking for the presence of missing data in the dataset, otherwise known as empty values, referring to absent or undefined values in the dataset, as seen in Figure 6.

```
InvoiceNo       0.00
StockCode       0.00
Description     0.27
Quantity        0.00
InvoiceDate     0.00
UnitPrice       0.00
CustomerID     24.93
Country         0.00
dtype: float64
```

Figure 6. Checking for missing values in the online retail dataset

From Figure 6, it is known that there are missing values in the description column and customerID column. In this research, missing values are handled by eliminating rows that have missing values. The dataset after going through data handling can be seen in Figure 7.

```
InvoiceNo      0.0
StockCode      0.0
Description    0.0
Quantity       0.0
InvoiceDate    0.0
UnitPrice      0.0
CustomerID     0.0
Country        0.0
dtype: float64
```

Figure 7. Online retail dataset after handling missing values

Feature engineering

Feature engineering is done to create recency, frequency, and monetary variables. The frequency variable is created by calculating how often customers make transactions in a certain period. With the frequency variable, the model can learn the characteristics of customers who often make transactions and customers who rarely make transactions. The value of the frequency variable can be seen in Table 2.

Table 2. Value at variable frequency

Frequency
2
128
31
73
17
…..

As opposed to the frequency variable, the recency variable is calculated by analyzing when the last time the user made a transaction. By analyzing the recency variable, the model can analyze customer characteristics, whether they are still active or inactive based on when the last time the customer made a transaction. The value of the recency variable can be seen in Table 3.

Table 2. Value at variable recency

Recency
325
1
74
18
309
…..

Furthermore, the monetary variable or in the research referred to as “ammount” is the total money spent by customers in the entire transaction. This variable can give the model an understanding of how much money a customer can spend in making a transaction. The value of the monetary variable can be seen in Table 3.

Table 2. Value at variable monetary/ammount

Ammount ($)
0.00
4310.00
1797.00
1757.00
334.40
…..

Implementation of k-means and elbow method

In this study, a model was created using the k-means analysis technique, the elbow method was used to determine the optimal number of clusters. This method is based on the assumption that Within Cluster Sum of Squares (WCSS), which is the sum of squares of the distance between each data point and the cluster center, will decrease as the number of clusters increases. This is due to the fact that each data point tends to be closer to the centroid in smaller clusters, so the WCSS will decrease as the value of k increases.

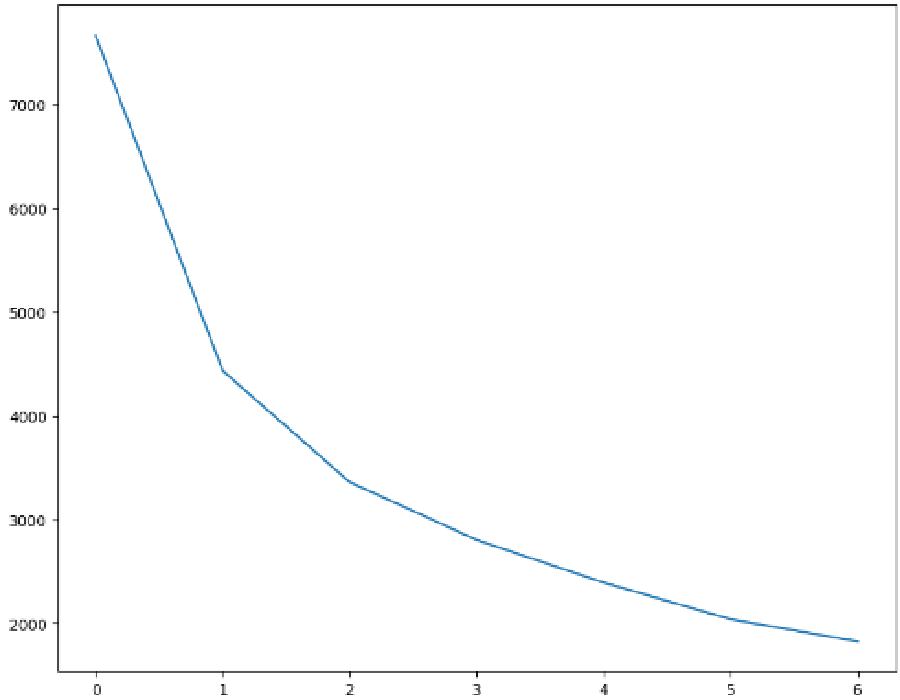

Figure 7. Graph of elbow method

From the graph, it is clear that the largest decrease in WSS occurs when the number of clusters increases from 2 to 3. After 3 clusters, the decrease in WSS starts to decrease significantly and becomes more gentle. This point, where the decrease in WSS is no longer significant with the addition of clusters, is often referred to as the “elbow” on the graph. Therefore, based on the elbow graph, the optimal number of clusters for our data is 3. This suggests that grouping the data into 3 clusters provides the best balance between model complexity and cluster quality, by minimizing within-cluster variation and avoiding overfitting with too many clusters.

Clustering result

In this study, a k value of 4 is used, indicating that the K-Means method will be used to form four different clusters. The result of this clustering will be presented in the data plot. Figure 8 displays the clustering results using the K-Means method with four clusters; the data belonging to each cluster will be marked with different symbols or colors.

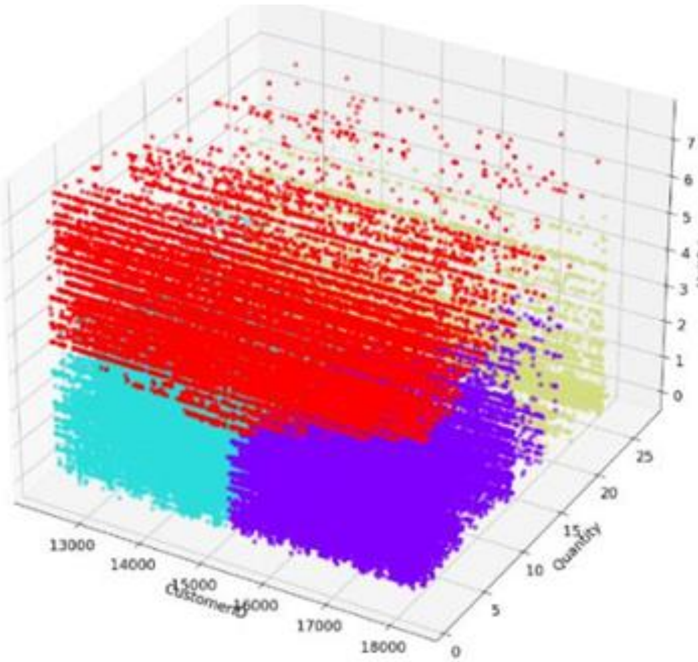

Figure 8. 3D graph of clustering results

The graph above illustrates the three most prominent clusters, which correspond to the three clusters produced by the K-Means model. Three variables are used to classify the transaction data: product quantity, unit price, and customerid. Using the resulting clustering results, we can identify the characteristics of each cluster as follows:

a) Cluster 0: Cluster 0 is a cluster of customers who have a high ammount compared to other customers. This means that customers in cluster one have high value recency and low value frequency. This cluster has a low transaction rate, but has the possibility of spending money on a large enough shopping. For example, a customer who makes a transaction once a month, but spends an ammount of up to $1000 when making a transaction.

b) Cluster 1: Cluster 1 is a cluster of customers who have a high number of transactions compared to other customers. This means that customers in cluster one have low value recency and high value frequency. This cluster has a high transaction rate, but has the possibility of spending money on shopping that is not too much. For example, a customer who makes a transaction 4 times a week, but only spends approximately $2 when making a transaction.

c) Cluster 2: Cluster 2 is a cluster of customers who rarely make transactions and do not spend too much ammount. These customers rarely make transactions and spend quite a bit of ammount when making transactions. For example, customers who make 2 transactions in 1 month, and each transaction made by customers only spend an ammount of around $2 to $5.

CONCLUSION

The purpose of this research is to segment customers based on customer characteristics in making transactions using a machine learning approach. From the K-Measn model that has been made, it is known that the model succeeded in providing 3 customer clusters. The three clusters produced have different characteristics so that the results can be used to help business processes in making decisions, for example in giving rewards to users. The existence of customer segmentation is expected to help companies in strategizing in providing services to customers. For future research, this study suggests comparing the performance of the K-Means algorithm with other segmentation algorithms. The existence of such research is expected to be able to provide Figurean about the weaknesses and shortcomings of the K-Means algorithm in segmenting data.

REFERENCES

[1] J. Berends and T. J. Gerpott, "Exploring price tolerance in online retail: A comparative analysis of price comparison website use and repeat purchases," *Electron. Commer. Res.*, 2023, doi: 10.1007/s10660-023-09791-5.

[2] J. L. Anderson, L. D. Jolly, and A. E. Fairhurst, "Customer relationship management in retailing: A content analysis of retail trade journals," *J. Retail. Consum. Serv.*, vol. 14, no. 6, pp. 394–399, Nov. 2007, doi: 10.1016/j.jretconser.2007.02.009.

[3] M. J. A. Berry and G. Linoff, "Data mining techniques : for marketing, sales, and customer support." J. Wiley, New York [etc. [Online]. Available: http://www.wiley.com/compbooks/

[4] M. Alves Gomes and T. Meisen, *A review on customer segmentation methods for personalized customer targeting in e-commerce use cases*, vol. 21, no. 3. Springer Berlin Heidelberg, 2023. doi: 10.1007/s10257-023-00640-4.

[5] A. Saxena, A. Agarwal, B. K. Pandey, and D. Pandey, "Examination of the Criticality of Customer Segmentation Using Unsupervised Learning Methods," *Circ. Econ. Sustain.*, no. 0123456789, 2024, doi: 10.1007/s43615-023-00336-4.

[6] A. G. Kök, M. L. Fisher, and R. Vaidyanathan, "Assortment Planning: Review of Literature and Industry Practice," 2008, pp. 99–153. doi: 10.1007/978-0-387-78902-6_6.

[7] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979, doi: 10.2307/2346830.

[8] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, in SODA '07. USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[9] H. Ismkhan and M. Izadi, "K-means-G*: Accelerating k-means clustering algorithm utilizing primitive geometric concepts," *Inf. Sci. (Ny).*, vol. 618, pp. 298–316, Dec. 2022, doi: 10.1016/j.ins.2022.11.001.

[10] G.-Y. Zhang, Y.-R. Zhou, X.-Y. He, C.-D. Wang, and D. Huang, "One-step Kernel Multi-view Subspace Clustering," *Knowledge-Based Syst.*, vol. 189, p. 105126, Feb. 2020, doi: 10.1016/j.knosys.2019.105126.

[11] D. Huang, J. Lai, and C.-D. Wang, "Ensemble clustering using factor graph," *Pattern Recognit.*, vol. 50, pp. 131–142, Feb. 2016, doi: 10.1016/j.patcog.2015.08.015.

[12] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwoh, "Ultra-Scalable Spectral Clustering and Ensemble Clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 6, pp. 1212–1226, Jun. 2020, doi: 10.1109/TKDE.2019.2903410.

[13] X. Yang, W. Yu, R. Wang, G. Zhang, and F. Nie, "Fast spectral clustering learning with hierarchical bipartite graph for large-scale data," *Pattern Recognit. Lett.*, vol. 130, pp. 345–352, Feb. 2020, doi: 10.1016/j.patrec.2018.06.024.

[14] Lina, "No TitleÉ?__," *Ekp*, vol. 13, no. 3, pp. 1576–1580, 2015.

[15] A. Qtaish, M. Braik, D. Albashish, M. T. Alshammari, A. Alreshidi, and E. J. Alreshidi, "Optimization of K-means clustering method using hybrid capuchin search algorithm," *J. Supercomput.*, vol. 80, no. 2, pp. 1728–1787, 2024, doi: 10.1007/s11227-023-05540-5.

[16] C. Calvo-Porral and J.-P. Lévy-Mangin, "From 'foodies' to 'cherry-pickers': A clustered-based segmentation of specialty food retail customers," *J. Retail. Consum. Serv.*, vol. 43, pp. 278–284, Jul. 2018, doi: 10.1016/j.jretconser.2018.04.010.

[17] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," *Appl. Soft Comput.*, vol. 113, p. 107924, Dec. 2021, doi: 10.1016/j.asoc.2021.107924.

[18] B. Wan, W. Huang, B. Pierre, Y. Cheng, and S. Zhou, "K-Means algorithm based on multi-feature-induced order," *Granul. Comput.*, vol. 9, no. 2, 2024, doi: 10.1007/s41066-024-00470-w.

[19] S. Molnár and L. Tamás, *Variational autoencoders for 3D data processing*, vol. 57, no. 2. Springer Netherlands, 2024. doi: 10.1007/s10462-023-10687-x.

[20] I. Pratama, A. E. Permanasari, I. Ardiyanto, and R. Indrayani, "A review of missing values handling methods on time-series data," *2016 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2016 - Proc.*, pp. 1–6, 2017, doi: 10.1109/ICITSI.2016.7858189.

[21] A. Vesin *et al.*, "Reporting and handling missing values in clinical studies in intensive care units," *Intensive Care Med.*, vol. 39, no. 8, pp. 1396–1404, 2013, doi: 10.1007/s00134-013-2949-1.

[22] P. Chauhan and M. Shukla, "A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm," *Conf. Proceeding - 2015 Int. Conf. Adv. Comput. Eng. Appl. ICACEA 2015*, pp. 580–585, 2015, doi: 10.1109/ICACEA.2015.7164758.

[23] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, “Normalization Techniques in Training DNNs: Methodology, Analysis and Application,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10173–10196, 2023, doi: 10.1109/TPAMI.2023.3250241.

[24] N. Aggarwal, B. S. Saini, and S. Gupta, “Feature engineering-based analysis of DaTSCAN-SPECT imaging-derived features in the detection of SWEDD and Parkinson’s disease,” *Comput. Electr. Eng.*, vol. 117, p. 109241, Jul. 2024, doi: 10.1016/j.compeleceng.2024.109241.

[25] “s41870-022-01012-w.pdf.crdownload.”

[26] W. Wu, W. Wang, X. Jia, and X. Feng, “Transformer Autoencoder for K-means Efficient clustering,” *Eng. Appl. Artif. Intell.*, vol. 133, p. 108612, Jul. 2024, doi: 10.1016/j.engappai.2024.108612.

[27] D. Parnes and A. Gormus, “Prescreening bank failures with K-means clustering: Pros and cons,” *Int. Rev. Financ. Anal.*, vol. 93, p. 103222, May 2024, doi: 10.1016/j.irfa.2024.103222.

[28] A. R. Lashiyanti, I. R. Munthe, and F. A. Nasution, “Optimisasi Klasterisasi Nilai Ujian Nasional dengan Pendekatan Algoritma K-Means, Elbow, dan Silhouette,” *J. Ilmu Komput. dan Sist. Inf.*, vol. 6, no. 1 SE-, pp. 14–20, Mar. 2023, [Online]. Available: http://ejournal.sisfokomtek.org/index.php/jikom/article/view/1550

[29] Z. Salman and A. Alomary, “Performance of the K - means and fuzzy C - means algorithms in big data analytics,” *Int. J. Inf. Technol.*, vol. 16, no. 1, pp. 465–470, 2024, doi: 10.1007/s41870-023-01436-y.