Jurnal ELTIKOM :
Jurnal Teknik Elektro, Teknologi Informasi dan Komputer

Vol. 7, Issue 1, June 2023, page. 21-28
ISSN 2598-3245 (Print), ISSN 2598-3288 (Online)
DOI: http://doi.org/10.31961/eltikom.v7i1.648
Available online at http://eltikom.poliban.ac.id

CUSTOMER SEGMENTATION BASED ON LOYALTY LEVEL USING K-MEANS AND LRFM FEATURE SELECTION IN RETAIL ONLINE STORE

Tiara Lailatul Nikmah[1*], Nur Hazimah Syani Harahap[1], Gina Cahya Utami[2], Muhammad Mirza Razzaq[3]

[1)] Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia
[2)] Department of Informatics, Universitas Amikom Purwokerto, Purwokerto, Indonesia
[3)] Department of Informatics Engineering, Universitas Dian Nuswantoro, Semarang, Indonesia
e-mail: tiaralaila21@gmail.com, nurhazimahsyani02@gmail.com, ginacu12@gmail.com, mirzaraz-zaq006@gmail.com

Received: November 22, 2022 – Revised: February 23, 2023 – Accepted: February 28, 2023

ABSTRACT

Customer experience is a key component in increasing sales numbers. Customers are important assets that must be kept up for a corporation or firm. Prioritizing customer service is one way to protect client loyalty. To ensure that service priority is right on target, this research was conducted on groups of consumers who are anticipated to have high business prospects. The 2011 retail online shop sales dataset with 379,980 records and eight characteristics was used. The length, recency, frequency, and monetary (LRFM) feature selection approach was used in the study process to select features for further segmentation using the K-Means data mining method to define consumer types. Following the completion of the research, clients were divided into four categories: Premium Loyalty, Inertia Loyalty, Latent Loyalty, and No Loyalty. The correct clustering results are displayed in the validation test using the Silhouette Score Index technique, which yielded a score value of 0.943898. Based on the outcomes of this segmentation, business actors may prioritize providing clients with the proper service.

Keywords*: customer segmentation, k-means, LRFM, online retail.*

I. INTRODUCTION

ADVANCES in information technology today make it easier for people to meet their needs. Convenience is obtained in various fields, one of which is shopping. People use online shopping services, or also known as e-commerce. E-commerce is another term for an online store, a shopping method involving social networking sites for buying and selling transactions. This activity eliminates the need for consumers and sellers to physically visit stores to view, buy, and sell goods. They can browse items online, make purchases, send payments, and items will be delivered to their homes using courier services without having to leave the comfort of their homes [1].

There are many types of online stores on the internet, both small and large. Although there are many new online stores, not everyone is interested. Since customers do not see the store as an attractive place to make purchases or sales, many online stores fail to make a profit [2]. A big factor in increasing sales figures is customer experience. Consumers are more likely to believe in using online purchasing services when they have more experience [3].

E-commerce companies invest heavily in early detection [4]. Understanding consumer behavior is very important for business actors. Every transaction that occurs on the internet between sellers and buyers can be recorded as historical data and can be used to predict customer behavior [5]. Machine learning algorithms can be used to implement this strategy in online retail stores [6]. Sellers should be able to analyze the data to group customers by loyalty level so that information can be obtained to estimate the priority level of customer service. In addition, sellers can develop the best service strategy from these segmentations to maintain customer loyalty.

The K-means clustering algorithm is a method of vector quantization used to partition observations into clusters [7]. It is an unsupervised learning algorithm used to solve clustering problems in machine learning or data science [8]. K-Means algorithm had weaknesses in previous studies, particularly in estimating the number of clusters [9]. The algorithm finds the best centroids by alternating between assigning data points to clusters based on the current centroids and choosing new centroids. The Elbow method is useful for determining the best number of clusters [10]. It works by plotting the explained variance against the number of clusters and finding an "elbow" point, where the rate of decrease sharply shifts[11], [12]. The elbow method uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters. Evaluating the SSE value is used to test the consistency of the corresponding number of clusters. SSE is a technique for evaluating clusters based on the highest error. In accordance with the concept of the Elbow technique applied to customer groupings using K-Means, SSE generates the maximum distance from a number of k value tests [13]. The graph will eventually experience a considerable decrease with a curve known as the angle criterion. The optimal value k, then becomes its best number of clusters [14].

LRFM (Length, Recency, Frequency, Monetary Value) [5] analysis is a marketing technique used to quantitatively rank and group customers based on the length, recency, frequency, and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns [15], [16]. It evaluates which customers are of highest and lowest value to an organization based on purchase length, recency, frequency, and monetary total of their recent transactions [17]. This study used customer segmentation and the K-Means algorithm based on the LRFM model for feature selection to classify potential customer values. Consumers were segmented using the K-Means method based on payment information obtained from the LRFM model to determine the level of potential customers. The cluster validation test results were carried out using the Silhouette Coefficient Index to determine the accuracy of the cluster.

The rest of the study is arranged as follows. In the second part, the research method describes the proposed method framework, data preprocessing, LRFM feature selection, elbow method, k-means clustering, and evaluation. In the third part, the result and discussion explain the results of the evaluation and our interpretation of the results. Then, in the fourth part, the conclusion explains the conclusion of the study and a summary of our research contributions.

II. RESEARCH METHOD

Research methods are procedures that are followed as guidelines when conducting research so that the results can be explained scientifically. The methods used in this study can be seen in Figure 1.

A. Preprocessing Data

In the flowchart described above, the first stage carried out is data preprocessing. Before starting the Data Mining process, data must be cleaned by removing duplicate data, identifying missing values or inconsistent data and correcting errors in the data [18].

B. Feature Selection LRFM

After the completion of data processing, feature selection was performed using L, R, F, and M attributes. The LRFM model was developed from the RFM model created by Arthur Hughes [19]. The RFM model is commonly used for segmenting customer behavior, and it assesses the profitability or potential customers based on their Recency, Frequency, and Monetary values. The LRFM model, which is a development of the RFM model created by Arthur Hughes, is also used for feature selection after data processing. In addition to the three existing variables, namely Recency, Frequency, and Monetary, Chang and Tsay introduced a new variable called Length [9].

The LRFM model has oversight because it is more suitable for analyzing the level of potential customers. The segmentation method with LRFM gives value according to the value of potential customers [20]. Variables in the clustering process will be selected according to the values of L, R, F and M. Then the data is carried out normalization. Normalization of data is an important step in LRFM (Length,

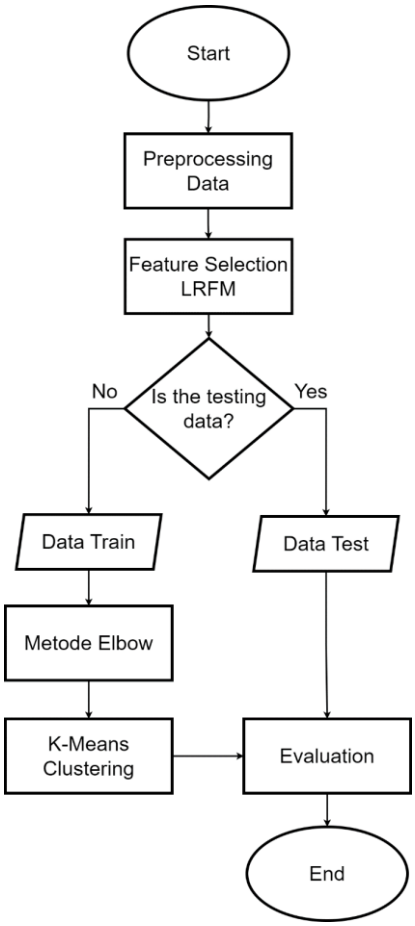

Figure 1. K-Means Method on Customer Segmentation

Recency, Frequency, Monetary Value) analysis [17-21]. Data normalization aids in comparing customers across different segments and identifying patterns in customer behavior, and data should be standardized to the same scale as the Standard Scaler method at this stage due to the vast difference in data between L, R, F, and M.

The segmentation or clustering stage is performed with the K-Means Clustering algorithm for k=1 to k=n. K-Means is one of the clustering techniques often used to group data based on comparable qualities [22]. These data groups are called clusters. Each input is trained using K-Means to group to the nearest cluster. Based on its input data, K-Means uses weights and all their neighbors to dynamically update their destination functions [23].

The stage in the first K-Means method is to determine the number of clusters. The number of clusters is determined using the Elbow method. The Elbow method is used to determine the optimal number of clusters from a dataset by minimizing the total number of variations or squared distances within the clusters. This is achieved through a visual method starting with a value of k=2 and incrementing it by 1 at each step. When the value of k = 3, if there is a significant change that is inversely proportional to the previous value, the number of clusters before the change is considered the most suitable [24]. In its use, the Elbow method uses the evaluation of SSE values. SSE is a way of validating a cluster through the sum of squares of each cluster member towards its center [25]. The greater the distance that forms the elbow point, the greater the number of clusters that is optimal [26]. The SSE formula is as follows in Equation 1.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x) \quad (1)$$

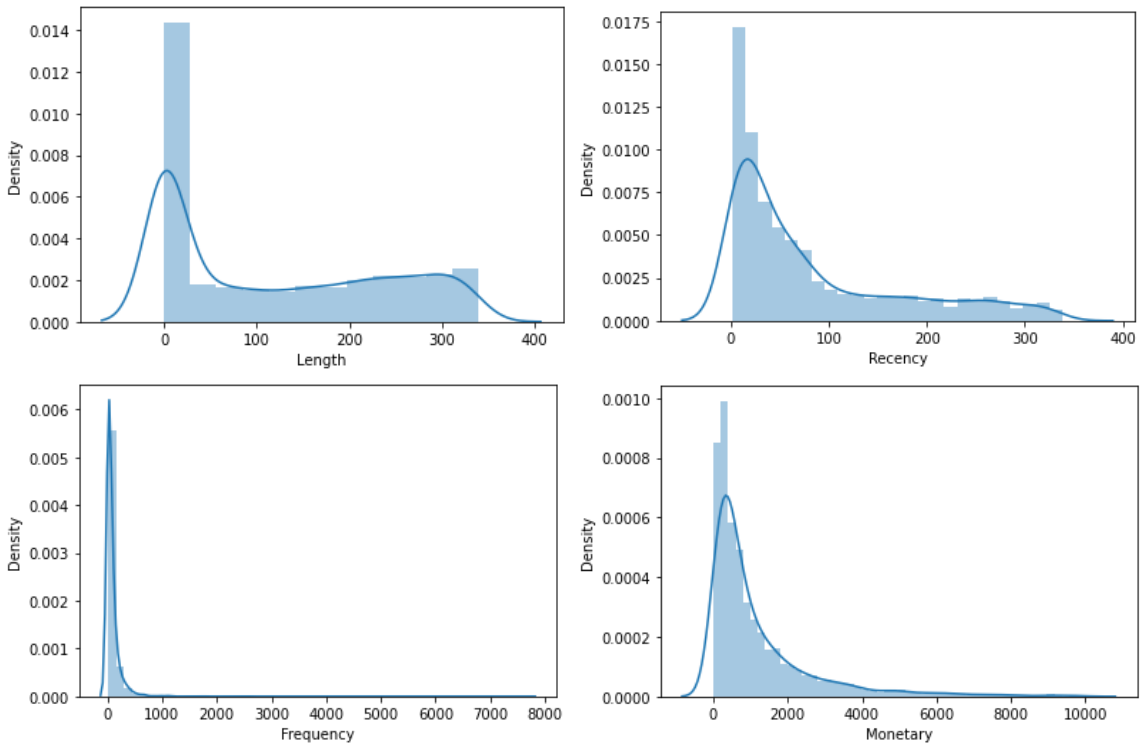

Figure 2. Length, Recency, Frequency, and Monetary (LRFM) distribution chart

where k is the number of clusters formed, C_i is cluster i, M_i is center of cluster i, and $dist^2(m_i,x)$ is distance between data point x and center of cluster m_i.

The next stage involves the arbitrary selection of the initial centroid based on the quantity of clusters [27]. Furthermore, the Euclidean distance formula at Equation (2) was used to calculate the distance of the data to the centroid.

Following the calculation of the distance for each record using the Euclidean distance formula, the centroid is updated by calculating the average value of the values in each cluster. If there is still data that moves clusters or changes in centroid values, the process returns to stage 3 [28].

The last step in the customer segmentation method is evaluation. Evaluation is a step to measure the performance of the methods used. This step uses the Silhouette Coefficient Index score matrix. The Silhouette Coefficient Index is an analytical method for obtaining validation values in the clustering method [29].

III. Result and Discussion

A. Preprocessing Data

The research utilized a dataset from Kaggle's online retail platform within the time frame of January 4th, 2011 to December 10th, 2011, consisting of 379980 records with 8 different attributes. These attributes include Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country.

Data cleaning, reducing missing values, and removing duplicate data are a few of the phases that must be completed before clustering. The range of values for each variable is then standardized as part of the data standardization process using a standard scaler. The Google Collab IDE is used for the preprocessing step before the clustering step.

B. Feature Selection LRFM

Using the K-Means technique, particular data properties are retrieved as part of the consumer segmentation process. Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country are some of these properties. In this study, the LRFM approach was used to choose qualities or features.

TABLE 1
SSE RESULTS ON THE ELBOW METHOD

Cluster	SSE
1	21219,99
2	16166,19
3	12264,03
4	9898,52
5	8349,98
6	6931,56
7	5905,93
8	5155,66
9	4554,50

TABLE 2
LRFM SEGMENTATION RESULTS

Customer ID	Length	Recency	Frequency	Monetary	Cluster
12346	0	325	2	1	3
12347	315	2	151	3598,21	2
12348	243	75	14	904,44	2
..	..	..	..	..	..
18283	334	3	756	2094,88	2
18287	159	42	70	1837,28	2

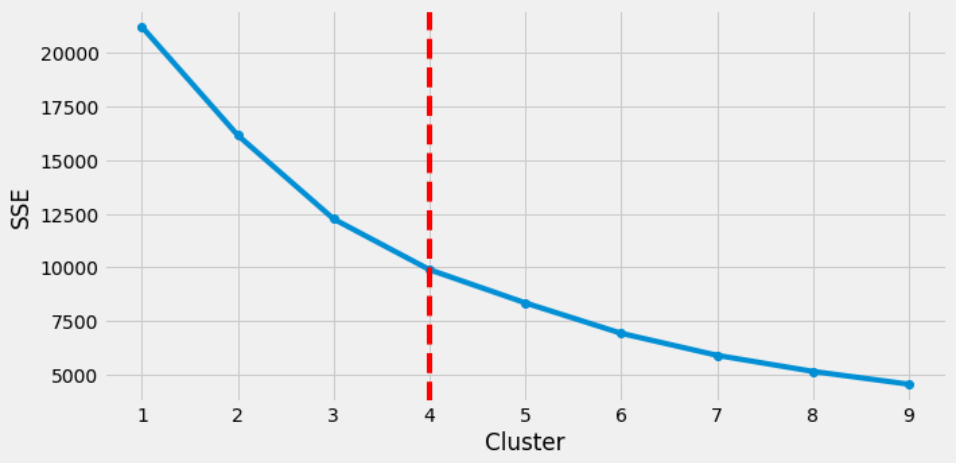

Figure 3. Elbow Chart Results

Based on Figure 2, L is the length, which in this context refers to the timeline from the first to the final date on which a client transacted with the business throughout the study period. Data were collected for this study between January 2011 and December 2011. R is a Recency where in this case it is intended to span from the last date of the transaction made by the customer to the last date of the transaction at the company, which is December 10, 2011. F is a Frequency where in this case it is intended how often customers make transactions within the period of January 2011 to December 2011. Additionally, M is the Monetary where it refers to the amount of money that clients spent between January 2011 and December 2011. Then, the data is divided into 2 parts, namely 80% training data and 20% test data. This data share uses train_test_split functions from Python's Sklearn library.

C. Customer Segmentation with K-Means Clustering

After data division, the data set is used to train the K-Means model. Nine clusters have been developed as a result of the K-Means segmentation method. The performance of each cluster in this test is adapted to fit the range of possible values from the Elbow method. The test results with the elbow approach are shown in Figure 3.

The results of the SSE calculation used to test the consistency of the corresponding number of clusters in the Elbow method can be seen in Table 1. According to Figure 1, clusters 1, 2, 3, and 4 show the greatest and sharpest graph decreases, which indicate a fall, while subsequent spots show a more gradual decline. Then, k is utilized with a value of 4. Based on the LRFM property that we have supplied, segmentation is carried out. The outcomes of the clustering are displayed in Table 2.

TABLE 3
CLUSTER CATEGORIZATION RESULTS

Customer ID	Length	Recency	Frequency	Monetary	Cluster	Label
12346	0	325	2	1	3	No Loyalty
12347	315	2	151	3598,21	2	Inertia Loyalty
12348	243	75	14	904,44	2	Inertia Loyalty
..	..	..	..	..	..	..
18283	334	3	756	2094,88	2	Inertia Loyalty
18287	159	42	70	1837,28	2	Inertia Loyalty

TABLE 4
SILHOUETTE COEFFICIENT INDEX RESULTS ON K-MEANS CLUSTERING

Cluster	Silhouette Coefficients Index
1	0,978971
2	0,939537
3	0,913187
4	0,632766
5	0,473583
6	0,480639
7	0,478278
8	0,486042
9	0,405870

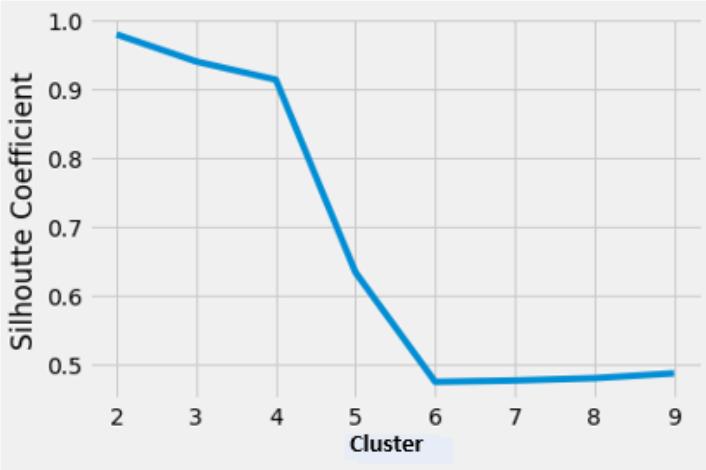

Figure 4. Silhouette Coefficient Index Graph

Based on Table 1, the subscriptions are grouped into 4 categories [30]. First, premium loyalty, the level of loyalty to customers if the level of attachment is so high that it can run in harmony with repurchase activity. Second, inertia loyalty, the level of loyalty to customers if there is a low attachment to a high repurchase. Third, latent loyalty, the level of loyalty is hidden in customers if loyalty or attachment is relatively high but the repurchase rate is low. Fourth, no loyalty, the level of customer loyalty if the level of attachment is low and the purchasing activity is also low. The results of cluster categorization can be seen in Table 3.

D. Evaluation

The result of calculating the value of the Silhouette Index (*SI*) between -1 to 1. If *SI* = 1, object *i* is already in the right cluster. If the value of *SI* = 0, then object *i* is between two clusters so that the object is not clear. It should be entered into cluster A or cluster B. However, if *SI* = -1, it means that the resulting cluster structure is overlapping, so object *i* is more appropriately inserted into another cluster [31]. Table 4 shows the *SI* calculation results and Figure 2 shows the *SI* graph.

Based on the evaluation results using the Silhouette Score index in Figure 4, a value of 0.943898 was obtained. This value is the expected value. These results show good results because the Silhouette Score index value is getting closer to 1 which indicates that the clustering is getting better and the objects in the cluster are getting more precise. Therefore, the results of this study are in accordance with the expected results.

IV. CONCLUSION

The rapid growth of the e-commerce industry has increased competition among online businesses, forcing owners to improve the quality of their customer service in order to stay in business. The implementation of service priorities to customers must be based on their level of loyalty, requiring the development of a model to analyze potential customers and group them accordingly. To achieve this, this study utilized the K-Means algorithm and LRFM method, segmenting payment behavior and using LRFM for feature selection to measure potential customers' loyalty level. The segmentation divided customers into four groups: Premium Loyalty, Inertia Loyalty, Latent Loyalty, and No Loyalty. The Silhouette Score Index test resulted in a score of 0.943898, enabling business actors to prioritize the right service for their customers based on the segmentation results.

REFERENCES

[1] S. M. Taj and A. Kumaravel, "Intentions of online shoppers prediction by fuzzy petri nets construction," Int. J. Sci. Technol. Res., vol. 9, no. 2, pp. 1761–1768, 2020.

[2] Y. Christian, "Comparison of machine learning algorithms using weka and sci-kit learn in classifying online shopper intention," J. Informatics Telecommun. Eng., vol. 3, no. 1, pp. 58–66, 2019.

[3] K. C. Koththagoda and H. Herath, "Factors influencing online purchasing intention: The mediation role of consumer attitude," J. Mark. Consum. Res., vol. 42, no. 2003, pp. 66–74, 2018.

[4] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," Neural Comput. Appl., vol. 31, pp. 6893–6908, 2019.

[5] M. R. Kabir, F. Bin Ashraf, and R. Ajwad, "Analysis of different predicting model for online shoppers' purchase intention from empirical data," in 2019 22nd International Conference on Computer and Information Technology (ICCIT), 2019, pp. 1–6.

[6] T. Brian and A. Sanwidi, "Implementasi Algoritma Apriori Untuk Market Basket Analysis Berbasis R," J. ELTIKOM J. Tek. Elektro, Teknol. Inf. dan Komput., vol. 2, no. 1, pp. 1–8, 2018.

[7] C. S. G. Dhas, N. Yuvaraj, N. V Kousik, and T. D. Geleto, "D-PPSOK clustering algorithm with data sampling for clustering big data analysis," in System Assurances, Elsevier, 2022, pp. 503–512.

[8] S. Saeed, H. Bin Haroon, M. Naqvi, N. Z. Jhanjhi, M. Ahmad, and L. Gaur, "A systematic mapping study of low-grade tumor of brain cancer and csf fluid detecting approaches and parameters," Approaches Appl. Deep Learn. Virtual Med. Care, pp. 236–259, 2022.

[9] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, F. Hunaini, and T. M. A. Aziz, "Segmentation model of customer lifetime value in small and medium enterprise (SMEs) using K-means clustering and LRFM model," Int. J. Integr. Eng., vol. 11, no. 3, 2019.

[10] A. D. Savitri, F. A. Bachtiar, and N. Y. Setiawan, "Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus: Belle Crown Malang)," J. Pengemb. Teknol. Inf. Dan Ilmu Komput. E-ISSN, vol. 2548, 2009.

[11] G. Chen, B. Sheng, R. Luo, and P. Jia, "A parallel strategy for predicting the quality of welded joints in automotive bodies based on machine learning," J. Manuf. Syst., vol. 62, pp. 636–649, 2022.

[12] T. Saheb, M. Dehghani, and T. Saheb, "Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis," Sustain. Comput. Informatics Syst., vol. 35, p. 100699, 2022.

[13] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Optimasi Cluster Pada Data Stunting: Teknik Evaluasi Cluster Sum of Square Error dan Davies Bouldin Index," in Prosiding Seminar Nasional Riset Information Science (SENARIS), 2019, vol. 1, pp. 918–926.

[14] S. F. Hussain and M. Haris, "A k-means based co-clustering (kCC) algorithm for sparse, high dimensional data," Expert Syst. Appl., vol. 118, pp. 20–34, 2019.

[15] S. Khodabandehlou, "Designing an e-commerce recommender system based on collaborative filtering using a data mining approach," Int. J. Bus. Inf. Syst., vol. 31, no. 4, pp. 455–478, 2019.

[16] E. Bakhshizadeh, H. Aliasghari, R. Noorossana, and R. Ghousi, "Customer Clustering Based on Factors of Customer Lifetime Value with Data Mining Technique (Case Study: Software Industry)," Int. J. Ind. Eng. Prod. Res., vol. 33, no. 1, pp. 1–16, 2022.

[17] R. Rahmadianti, A. Dhini, and E. Laoh, "Estimating customer lifetime value using LRFM model in pharmaceutical and medical device distribution company," in 2020 International Conference on ICT for Smart Society (ICISS), 2020, pp. 1–5.

[18] T. L. Nikmah, M. Z. Ammar, Y. R. Allatif, R. M. P. Husna, P. A. Kurniasari, and A. S. Bahri, "Comparison of LSTM, SVM, and naive bayes for classifying sexual harassment tweets," J. Soft Comput. Explor., vol. 3, no. 2, pp. 131–137, 2022.

[19] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context," Int. J. Electr. Comput. Eng., vol. 8, no. 4, p. 2367, 2018.

[20] S. Monalisa, "Analysis outlier data on rfm and lrfm models to determining customer loyalty with dbscan algorithm," in 2018 International Symposium on Advanced Intelligent Informatics (SAIN), 2018, pp. 1–5.

[21] S. Fatahi and M. Rabiei, "Users clustering based on search behavior analysis using the LRFM model (case study: Iran scientific information database (Ganj))," Iran. J. Inf. Process. Manag., vol. 36, no. 2, pp. 419–442, 2022.

[22] N. H. Syani, A. Amirullah, M. B. Saputro, and I. A. Tamaroh, "Classification of potential customers using C4. 5 and k-means algorithms to determine customer service priorities to maintain loyalty," J. Soft Comput. Explor., vol. 3, no. 2, pp. 123–130, 2022.

[23] M. I. Dzulhaq, K. W. Sari, S. Ramdhan, and R. Tullah, "Customer segmentation based on RFM value using K-means algorithm," in 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1–7.

[24] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis based on k-means," Indones. J. Electr. Eng. Comput. Sci., vol. 18, no. 1, pp. 470–477, 2020.

[25] D. Marutho, S. H. Handaka, and E. Wijaya, "The determination of cluster number at k-mean using elbow method and purity evaluation on headline news," in 2018 international seminar on application for technology of information and communication, 2018, pp. 533–538.

[26] M. A. Masud, J. Z. Huang, C. Wei, J. Wang, I. Khan, and M. Zhong, "I-nice: A new approach for identifying the number of clusters and initial cluster centres," Inf. Sci. (Ny)., vol. 466, pp. 129–151, 2018.

[27] F. Bin Ashraf, A. Matin, M. S. R. Shafi, and M. U. Islam, "An Improved K-means Clustering Algorithm for Multi-dimensional Multi-cluster data Using Meta-heuristics," in 2021 24th International Conference on Computer and Information Technology (ICCIT), 2021, pp. 1–6.

[28] H. Nguyen, X.-N. Bui, Q.-H. Tran, and N.-L. Mai, "A new soft computing model for estimating and controlling blast-produced ground vibration based on hierarchical K-means clustering and cubist algorithms," Appl. Soft Comput., vol. 77, pp. 376–386, 2019.

[29] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," EURASIP J. Wirel. Commun. Netw., vol. 2021, no. 1, pp. 1–16, 2021.

[30] O. Dogan, E. Ayçin, and Z. Bulut, "Customer segmentation by using RFM model and clustering methods: a case study in retail industry," Int. J. Contemp. Econ. Adm. Sci., vol. 8, 2018.

[31] R. D. Firdaus, T. G. Laksana, and R. D. Ramadhani, "Pengelompokan Data Persediaan Obat Menggunakan Perbandingan Metode K-Means Dengan Hierarchical Clustering Single Linkage," INISTA J. Informatics, Inf. Syst. Softw. Eng. Appl., vol. 2, no. 1, pp. 33–48, 2019.