Scientific Journal of Informatics
Vol. 9, No. 2, Nov 2022
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm

Alamsyah[1*], P. Eko Prasetyo[2], Sunyoto[3], Siti Harnina Bintari[4], Danang Dwi Saputro[5], Shohihatur Rohman[6], Rizka Nur Pratama[7]
[1,7]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia
[2]Department of Development Economics, Faculty of Economics, Universitas Negeri Semarang, Indonesia
[3,5,6]Department of Mechanical Engineering, Faculty of Engineering, Universitas Negeri Semarang, Indonesia
[4]Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: This research aims to do customer segmentation in retail companies by implementing the Recency Frequency Monetary (RFM) K-Means cluster model and algorithm optimized by the Elbow method.
Methods: This study uses several methods. The RFM model method was chosen to segment customers because it is one of the optimal methods for segmenting customers. The K-Means cluster algorithm method was chosen because it is easy to interpret, implement, fast in convergence, and adapt, but lacks sensitivity to the initial partitioning of the number of clusters. To help classify each category of customers and know the level of loyalty, they use a combination of the RFM model and the K-Means method. The Elbow method is used to improve the performance of the K-Means algorithm by correcting the weakness of the K-Means algorithm, which helps to choose the optimal *k* value to be used when clustering.
Result: This research produces customer segmentation 3 clusters with a Sum of Square Error (SSE) value of 25,829.39 and a Callinski-Harabaz Index (CHI) value of 36,625.89. The SSE and CHI values are the largest ones, so they are the optimal cluster values.
Novelty: The application of the integrated RFM model and the K-Means cluster algorithm optimized by the Elbow method can be used as a method for customer segmentation.

Keywords: Customer Segmentation, RFM, K-Means Algorithm, Elbow Method
Received October 2022 / **Revised** October 2022 / **Accepted** November 2022

INTRODUCTION

The development of Information Technology (IT), which is increasing, is certainly in line with the conditions of human civilization, including in the business sector [1]. In business, competition requires companies to maximize existing skills as well as possible to compete with other companies [2]. Companies must be able to understand and incorporate customer characteristics into an important thing to consider [3].

Changes in people's economic conditions have an effect on the increasingly fierce competition in the business industry [4]. Along with the increasing competition in the business industry in retail companies, retail companies are required to shift their focus not only to product-oriented but also to implement strategies that also focus on customer-oriented [5]. In doing customer oriented, it is required to know the characteristics of each customer [6]. Customer segmentation is needed to classify customers with the same characteristics. In one customer segmentation is a unified customer group that has the same characteristics [7].

*Corresponding author.
Email addresses: alamsyah@mail.unnes.ac.id (Alamsyah), pekoprasetyo@mail.unnes.ac.id (Prasetyo), sunyoto@mail.unnes.ac.id (Sunyoto), harnina@mail.unnes.ac.id (Bintari), danangdwisaputro@mail.unnes.ac.id (Saputro), prof.shoheh@gmail.com (Rohman), rzkprtma@students.unnes.ac.id (Pratama)
DOI: 10.15294/sji.v9i2.39437

One of the methods used to segment customers is the Recency Frequency Monetary (RFM) model. The RFM model is a model used to determine the difference between important customers from big data based on three variables, namely recency, frequency and monetary [8]. After knowing the customer segmentation group, customer characteristics can be analyzed based on the generated segmentation, so that planning can be done to retain old customers with the most appropriate company service strategy for each customer [9].

Data mining is used to assist in decision making [10], among the methods used are Recurrent Neural Network (RNN) [11], [12], naïve bayesian classifier algorithm [13], and K-Nearest Neighbor (KNN) algorithm [14]. Clustering is a data mining technique by dividing data in a set into groups with the similarity of data in one group greater than the similarity of the data with data in other groups [7]. The K-Means cluster is the most popular and widely studied clustering method to minimize clustering errors [15]. The K-Means approach uses a greedy strategy to generate new partitions by assigning each pattern to the nearest cluster center and calculating the new cluster center [16]. K-Means groups a specified data through several clusters (k clusters). The idea that arises is to provide a definition of the value of the center of k (k centroid), one by one in each cluster. In placing the center value, it must be done smartly because the difference in location is also a factor in the difference in results [17].

To help the process of grouping each category of customers and to know their loyalty level is to use a combination of the RFM model and the K-Means algorithm [18]. The Elbow method is used to improve the performance of the K-Means algorithm by correcting the weakness of the K-Means algorithm, which is helping to choose the optimal k value to be used when clustering [19]. The K-Means method is more optimal in its performance by adding the Elbow method to select the value of k clusters [20].

METHODS

The method used in this study uses the proposed method, namely the integration of the RFM model and the K-Means cluster algorithm. It is optimized by the Elbow method for retail companies in conducting customer segmentation, as shown in Figure 1.

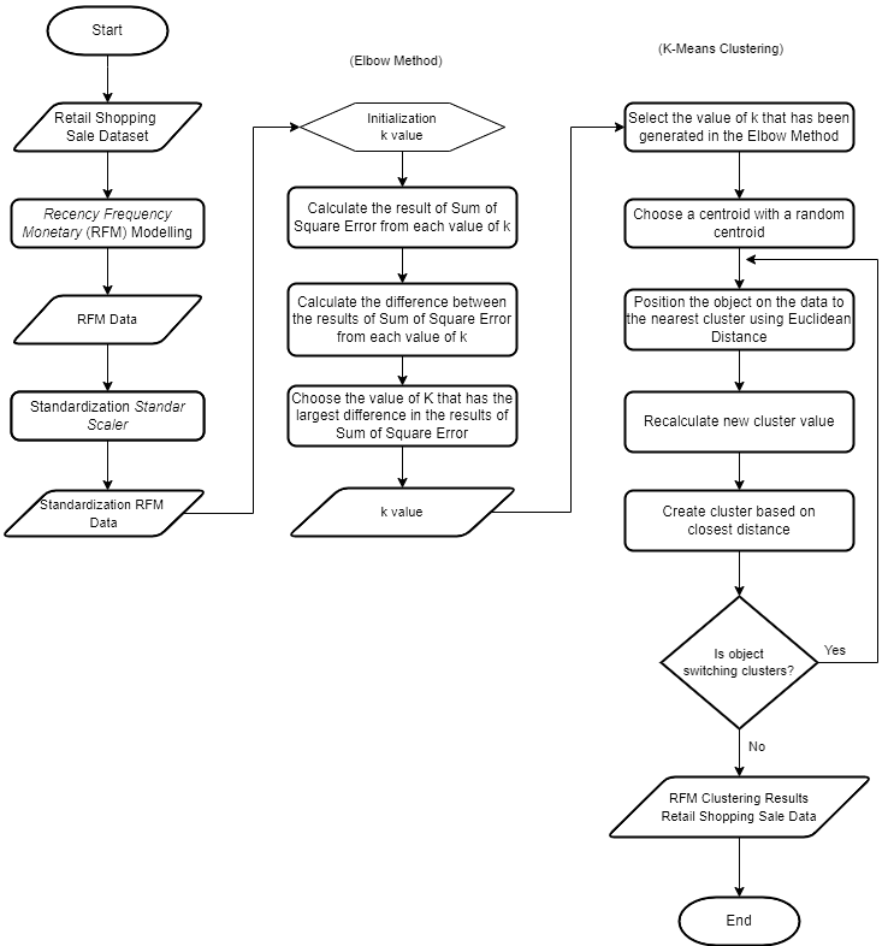

Figure 1. Flowchart of proposed method

Data Collection

In the process of collecting data, it is carried out to collect the required dataset. The dataset was obtained from Retail Shopping Sale Data, Kaggle. The dataset used is a collection of transactional data from E-commerce from Turkey in the period from February 1, 2020 to February 1, 2021. The dataset consists of 4 attributes with 62,295 rows. The attributes of the Retail Shopping Sale Data are described in Table 1.

Table 1. Attribute dataset

No.	Attributes	Type
1.	User ID	Nominal
2.	Invoice No	Numerical
3.	Total Payment	Numerical
4.	Invoice Date	Numerical

Recency Frequency Monetary Modeling

Recency is the amount of difference between the last transaction time and the current transaction time, or the time when the transaction data was published. Frequency is how often the customer makes transactions at a certain time. Monetary is the nominal amount of money owned by the customer at the time of the transaction, if the amount of money is greater, the value of M will also be greater [21].

The process of transforming the data into a Recency Frequency Monetary (RFM) model. The recency value in this study is the difference between the dataset upload time and the last time the customer made a transaction. The dataset upload time used in this study was January 4, 2022. In modeling the Recency value is obtained from the calculation of the distance from the most recent date. last attribute Invoice Date with the date of research by grouping from the User ID attribute. The smaller the Recency value, the closer the distance to the last transaction made by the customer, and the better the Recency value. The frequency value calculates how many times the customer has made a transaction. The Frequency value is obtained from

calculating the number of transactions from the Invoice No attribute based on the grouping of the User ID attribute. Monetary value is the total cost generated by the customer during the transaction. The monetary value uses the sum of the Total Payment attribute's values by grouping from the User ID attribute.

Data Standardization

In this process, standardization of data is carried out on the dataset. Data standardization was carried out in this study using the Standard Scaler Standardization. In Retail Shopping Sale Data, after data transformation using the RFM model method, the data range is quite far, with a fairly high range, the clustering process on the dataset may not be successful, so standardization of data is required using the following Equation (1) [22].

$$z = \frac{(x - u)}{s} \tag{1}$$

with description,
u = average sample value
s = standard deviation of the *training data*

Elbow Method

In the clustering process, it is necessary to input the initial k value of the cluster. The Elbow method is applied to assist in determining the value of k as input to the K-Means clustering algorithm. To obtain the optimal input k value, one of the steps that can be done is by applying the Elbow method. The results of the *k* value selected are obtained from the results of the initiation of the *k* value range through the right-angled point shown by the graph. The output results of the cluster value are used as input at the clustering stage with the K-Means algorithm. The workflow of the Elbow method is shown in Figure 2.

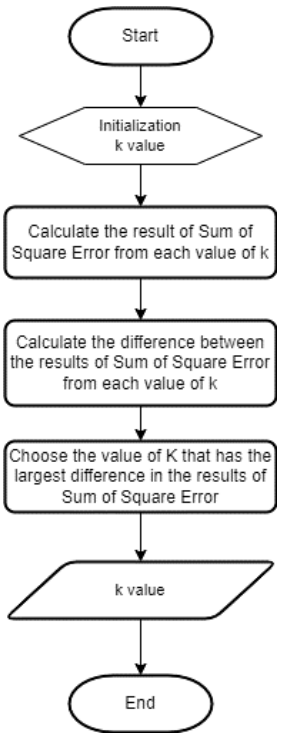

Figure 2. Elbow method flowchart

K-Means Cluster Algorithm

The k-means algorithm is a clustering algorithm to group data into certain groups. To retrieve data on this algorithm by not using class labels (unsupervised learning). The K-Means clustering process divides the data groups that become the input independently. Each cluster has a central point (centroid) that represents the cluster [23]. The K-Means algorithm is used to perform clustering in this study. The input to the K-Means process is the result of the RFM model that has been normalized and the k value obtained by the

Elbow method. The K-Means algorithm is used as an algorithm to perform clustering in this study. The input to the K-Means process is the result of the RFM model that has been normalized and the k value obtained by the Elbow method. The following is a flowchart of the K-Means cluster algorithm, which can be seen in Figure 3 [24].

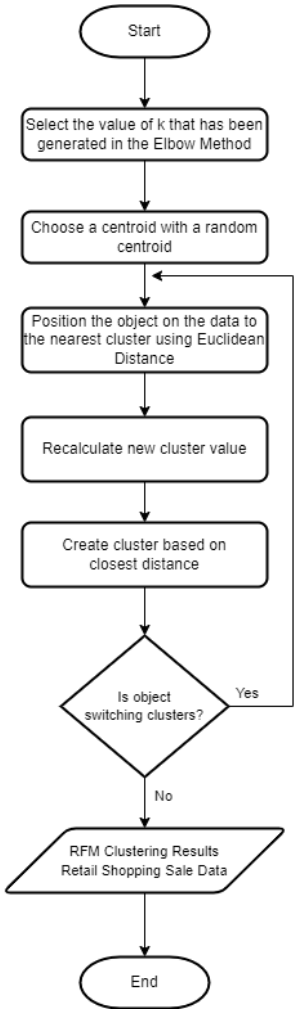

Figure 3. Flowchart of the k-means cluster algorithm

RESULT AND DISCUSSION

From the experiments that have been carried out, the Elbow method is used to determine the optimal number of clusters so that the difference in SSE values can be greater. The number of clusters with an elbow and the largest difference in the Sum of Square Error (SSE) value is the value of $k = 3$. The results of the Elbow method are shown in Figure 4.

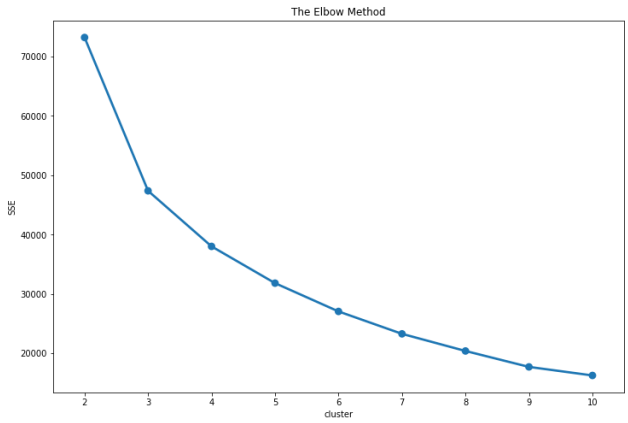

Figure 4. Graph chart elbow method

The number of clusters is not the value that has the largest SSE value difference but has the largest SSE difference value after the cluster value $k = 2$ and the point that shows the elbow on the Elbow chart. In this study, the value of $k = 3$ was used to cluster data, then an evaluation was carried out using the Callinski Harabaz Index (CHI) method, which is shown in Figure 5.

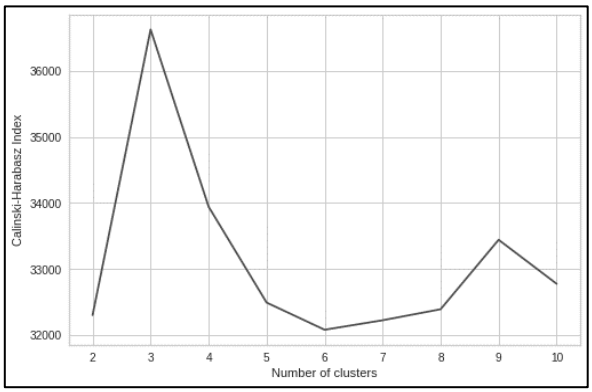

Figure 5. Chart of callinski-harabaz index

This study records the SSE value of each cluster and is evaluated using the Callinski Harabaz Index method. The results of the comparison of the Elbow and Callinski-Harabaz Index methods can be seen in Table 2.

Table 2. Comparison results of SSE and CHI

Number of Clusters	SSE score difference	Callinski Harabaz Index
2	55,232.94	32,300.87
3	**25,829.39**	**36,625.89**
4	9,360.17	33,942.48
5	6,190.64	32,490.18
6	4,765.53	32,077.30
7	3,774.45	32,221.07
8	2,883.74	32,385.81
9	2,684.35	33,441.17
10	1,440.19	32,776.14

Based on Table 2, it is known that the number of clusters that have **the largest difference in SSE values** and have **the largest Callinski Harabaz Index values** are clusters with a **value of *k*=3**. The cluster value generated using the Elbow method after being evaluated using the Callinski Harabaz Index is the cluster value that is proven to be the most optimal. The results of customer segmentation are analyzed to determine the number of members of each cluster that is formed. Figures 6 and 7 show a visualization of the number of members of each cluster and a visualization of the presentation of members of each cluster. Table 3 shows the average RFM of each cluster.

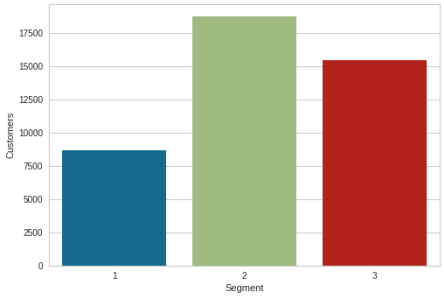

Figure 6. Number of members in each cluster

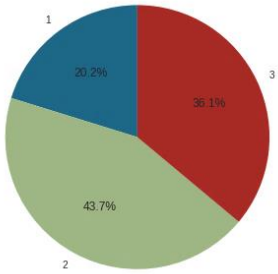

Figure 7. Percentage of members of each cluster

Table 3. Average RFM per cluster

Cluster	Average Recency	Average Frequency	Average Monetary
1	441	3	1,763.49
2	588	1	457.76
3	411	1	575.32

The results of the number of clusters based on the Elbow method and cluster performance tests using the Callinski Harabaz Index method determine the number of clusters to 3, in the process of determining the number of clusters it can make the separation of customer groups still active and tend to have Recency, Frequency and Monetary values that are superior to other groups. The shape of the character above divides the segment into 3, namely cluster 1 in the future which is considered capable of contributing very large profits, so the customers who are included in it by the company really need to be maintained.

The Proposed method in this study resulted in 3 customer segments that were more accurate and optimal than 8 customer segments which were shown through the calculation results of the SSE and CHI values that had been carried out. The drawback in this study is the determination of the centroid that still uses the random centroid method on K-Means, so it needs to be further developed related to the method for initial centroids, and not all clustering algorithms methods have been tested in this study.

CONCLUSION

Applying the RFM model integration and the K-Means cluster algorithm optimized by the Elbow method can be used for customer segmentation in retail companies. Applying the Elbow method to determine the optimal value of k clusters in clustering, so that the most optimal number of clusters is obtained and has the largest difference in SSE values, which is 25.829.39. The Elbow method was evaluated using the Callinski Harabaz Index method value of 36,625.89, and the results obtained were the same optimal cluster values. Thus, it can be concluded that the application of the Elbow method can be used as an optimization to determine the most optimal number of k cluster values in clustering.

REFERENCES

[1] B. E. Adiana, I. Soesanti, and A. E. Permanasari, "Analysis of Customer Segmentation Using a Combination of RFM Models and Clustering Techniques," *J. Terap. Teknol. Inf.*, vol. 2, no. 1, pp. 23–32, 2018.

[2] H. Zhao and C. He, "Objective Cluster Analysis in Value-Based Customer Segmentation Method," in *2009 Second Int. Workshop Knowl. Discov. Data Min.*, 2009, pp. 484–487.

[3] Y. Chen, G. Zhang, D. Hu, and S. Wang, "Customer Segmentation in Customer Relationship Management based on Data Mining," in *Int. Conf. Program. Lang. Manuf.*, 2006, pp. 288–293.
[4] I. Soesanti, "Web-based Monitoring System on the Production Process of Yogyakarta Batik Industry," *J. Theor. Appl. Inf. Technol.*, vol. 87, no. 1, pp. 146–152, 2016.
[5] N. W. Wardani, G. R. Dantes, and G. Indrawan, "Prediction of Customer Churn Using the C4.5 Decision Tree Algorithm based on Customer Segmentation to Retain Customers in Retail Companies," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 1, no. 1, pp. 16–24, 2018.
[6] J. T. Wei, S.-Y. Lin, Y.-Z. Yang, and H.-H. Wu, "Applying Data Mining and RFM Model to Analyze Customers' Values of a Veterinary Hospital," in *2016 Int. Symp. Comput. Consum. Control (IS3C)*, 2016, pp. 481–484.
[7] J. Wu and Z. Lin, "Research on Customer Segmentation Model by Clustering," in *Proc. 7th int. conf. Electron. commer. - ICEC '05*, 2005, p. 316.
[8] C.-H. Cheng and Y.-S. Chen, "Classifying the Segmentation of Customer Value Via RFM Model and RS Theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4176–4184, 2009.
[9] M. A. Berry, "Mastering Data Mining: The Art and Science of Customer Relationship Management," *Ind. Manag. Data Syst.*, vol. 100, no. 5, pp. 245–246, 2000.
[10] L. Ye, C. Qiu-ru, X. Hai-xu, L. Yi-jun, and Y. Zhi-min, "Telecom Customer Segmentation with K-Means Clustering," in *2012 7th Int. Conf. Comput. Sci. Educ. (ICCSE)*, 2012, pp. 648–651.
[11] Walid and Alamsyah, "Recurrent Neural Network for Forecasting Time Series With Long Memory Pattern," *J. Phys. Conf. Ser.*, vol. 824, p. 012038, 2017.
[12] A. Alamsyah, B. Prasetiyo, M. F. Al Hakim, and F. D. Pradana, "Prediction of COVID-19 Using Recurrent Neural Network Model," *Sci. J. Informatics*, vol. 8, no. 1, pp. 98–103, 2021.
[13] B. Prasetiyo, Alamsyah, and M. A. Muslim, "Analysis of Building Energy Efficiency Dataset Using Naive Bayes Classification Classifier," *J. Phys. Conf. Ser.*, vol. 1321, no. 3, p. 032016, 2019.
[14] I. Dinariyah and Alamsyah, "Accuracy Enhancement in Face Recognition Using 1D-PCA &Amp; 2D-PCA based on Multilevel Reverse-Biorthogonal Wavelet Transform with KNN Classifier," *J. Phys. Conf. Ser.*, vol. 1918, no. 4, p. 042144, 2021.
[15] N. Kurinjivendhan and K. Thangadurai, "Modified K-Means Algorithm and Genetic Approach for Cluster Optimization," in *2016 Int. Conf. Data Min. Adv. Comput. (SAPIENCE)*, 2016, pp. 53–56.
[16] A. K. Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[17] P. Bholowalia and A. Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 975–8887, 2014.
[18] O. Doğan, E. Ayçin, and Z. A. Bulut, "Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry," *Int. J. Contemp. Econ. Adm. Sci.*, vol. 8, no. 1, pp. 1–19, 2018.
[19] T. M. Kodinariya and P. R. Makwana, "Review on Determining of Cluster in K-Means," *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.
[20] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The Application of K-Means Clustering for Province Clustering in Indonesia of the Risk of the COVID-19 Pandemic based on COVID-19 Data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, 2022.
[21] A. M. Hughes, *Strategic Database Marketing*. McGraw –Hill, 2005.
[22] A. Ambarwari, Q. Jafar Adrian, and Y. Herdiyeni, "Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, 2020.
[23] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," Multidiscip. Sci. J., vol. 2, no. 2, pp. 226–235, 2019.
[24] P. N. Tan, M. Steinbach, and V. Kumar, "Data mining introduction," 2006.