Customer Segmentation with RFM Model using Fuzzy C-Means and Genetic Programming

Anas Syaifudin , Purwanto , Heribertus Himawan , M. Arief Soeleman
Universitas Dian Nuswantoro, Semarang, Indonesia

ABSTRACT

One of the strategies a company uses to retain its customers is Customer Relationship Management (CRM). CRM manages interactions and supports business strategies to build mutually beneficial relationships between companies and customers. The utilization of information technology, such as data mining used to manage the data, is critical in order to be able to find out patterns made by customers when processing transactions. Clustering techniques are possible in data mining to find out the patterns generated from customer transaction data. Fuzzy C-Means (FCM) is one of the best-known and most widely used fuzzy grouping methods. The iteration process is carried out to determine which data is in the right cluster based on the objective function. The local minimum is the condition where the resulting value is not the lowest value from the solution set. This research aims to solve the minimum local problem in the FCM algorithm using Genetic Programming (GP), which is one of the evolution-based algorithms to produce better data clusters. The result of the research is to compare the application of fuzzy c-means (FCM) and genetic programming fuzzy c-means (GP-FCM) for customer segmentation applied to the Cahaya Estetika clinic dataset. The test results of the GP-FCM yielded an objective function of 20.3091, while for the FCM algorithm, it was 32.44741. Furthermore, evaluating cluster validity using Partition Coefficient (PC), Classification Entropy (CE), and Silhouette Index proves that the results of cluster quality from gp-fcm are more optimal than fcm. The results of this study indicate that the application of genetic programming in the fuzzy c-means algorithm produces more optimal cluster quality than the fuzzy c-means algorithm.

Corresponding Author:

Anas Syaifudin, +6285803894744,
Faculty of Computer Science, Master of Computer Science,
Dian Nuswantoro University, Semarang, Indonesia,
Email: anzt.07@gmail.com

1. INTRODUCTION

A companys business is required to understand market conditions in order to be able to adjust to consumer needs. A company must be able to meet the needs following consumer demand. One of the important aspects of advancing a company is establishing good relationships with consumers [1]. Improper management between consumers and companies will impact poorly on the company. Customer Relationship Management (CRM) is a method that can be used to retain customers in a company [2]. CRM manages interactions and supports business strategies to build customer relationships that can benefit the company and drive sales growth [3].

The implementation of CRM is based on customer data and the utilization of information technology [4]. Utilizing the right information technology can benefit the company and its customers. Especially to find out the various characteristics of customers from transaction data carried out in a company by segmenting customers. Customer segmentation is the process of grouping customer data into several groups according to the transaction data carried out [5]. A modeling technique used to group transaction data is the RFM (Recency, Frequency, and Monetary) model [6]. RFM is used to assign values or perform calculations to get the Recency, Frequency, and Monetary values of the attributes that have been selected for RFM [7].

The RFM model is a modeling technique used to divide transaction data into three variables, namely the distance between the last time and the specified time (R), the total number of transactions carried out within the specified time range (F), and the total product value in the time range (M) [8]. RFM modeling techniques can be applied to data and information processing techniques to obtain customer groups from transaction data. The utilization of information technology, such as data mining used to manage customer data, is necessary to reveal patterns generated by customers when processing transactions [9]. The following stage is to find out the pattern generated from customer transaction data after the preprocessing process using RFM modeling, which can be used as one of the data mining grouping techniques is the clustering technique.

Several clustering techniques that are widely used are Fuzzy C-Means (FCM) [10, 11] In terms of grouping customer data, the FCM algorithm was successfully carried out by Saputra et al. [12] to group customer data. The FCM algorithm is used to help determine the promotion strategy to be carried out. Research conducted by Dharmawan et al. [13] grouped data using FCM to develop business strategies to acquire and retain customers. Prasetyo et al. [14] used the FCM algorithm to classify customer data to determine the right marketing strategy to acquire, retain and partner with e-commerce customers.

The problem often encountered in grouping using the FCM algorithm is that getting stuck in a minimum local condition is very easy. To overcome the minimum local problems, some researchers have succeeded in implementing evolution-based algorithms, as has been done by Adhzima et al. [15]. In his research, the genetic algorithm was successfully applied to optimize the FCM algorithm. In a study conducted by Andersen et al. [16] applied a genetic programming algorithm to produce better cluster quality. Another study conducted by Syelly et al. [17] successfully applied a genetic programming algorithm to group gambier plants.

This research aims to solve the minimum local problem in the FCM algorithm using Genetic Programming (GP), which is one of the evolution-based algorithms to produce better data clusters. Thus, the novelty of this paper is implementing GP to produce clusters that are more optimal in the FCM algorithm. The application of FCM and Genetic Programming Fuzzy C-Means (GP-FCM) for customer segmentation is applied to the Cahaya Estetika clinic dataset.

2. RESEARCH METHOD

The proposed method to solve the local minimum is an algorithm based on evolutionary computation, in which several candidate solutions to the clustering problem are randomly generated. First, the initial value of the cluster center point V is generated randomly, then the V value is used to calculate the u matrix. Furthermore, this V value will be evolved using selection, crossover, and mutation to get the most optimum V value. The proposed model can be seen in Figure 1. This research contributes to utilizing a genetic programming algorithm to optimize the objective function so that the grouping performance can be determined.

The initial step that must be taken in Figure 1 is to prepare a solution representation and input and output specifications by normalizing the sales transaction dataset into the RFM model. First, determine the size of genetic parameters, including population size, number of generations, crossover probability, mutation probability, and max iteration. Next, determine the fitness function used; in this case, the fitness function is the objective function that minimizes the distance between the data and the cluster center point (V) on the Fuzzy C-meansselection of the selection function according to a predetermined probability. Finally, update the u matrix, which is the degree of data membership at each cluster center point (V).
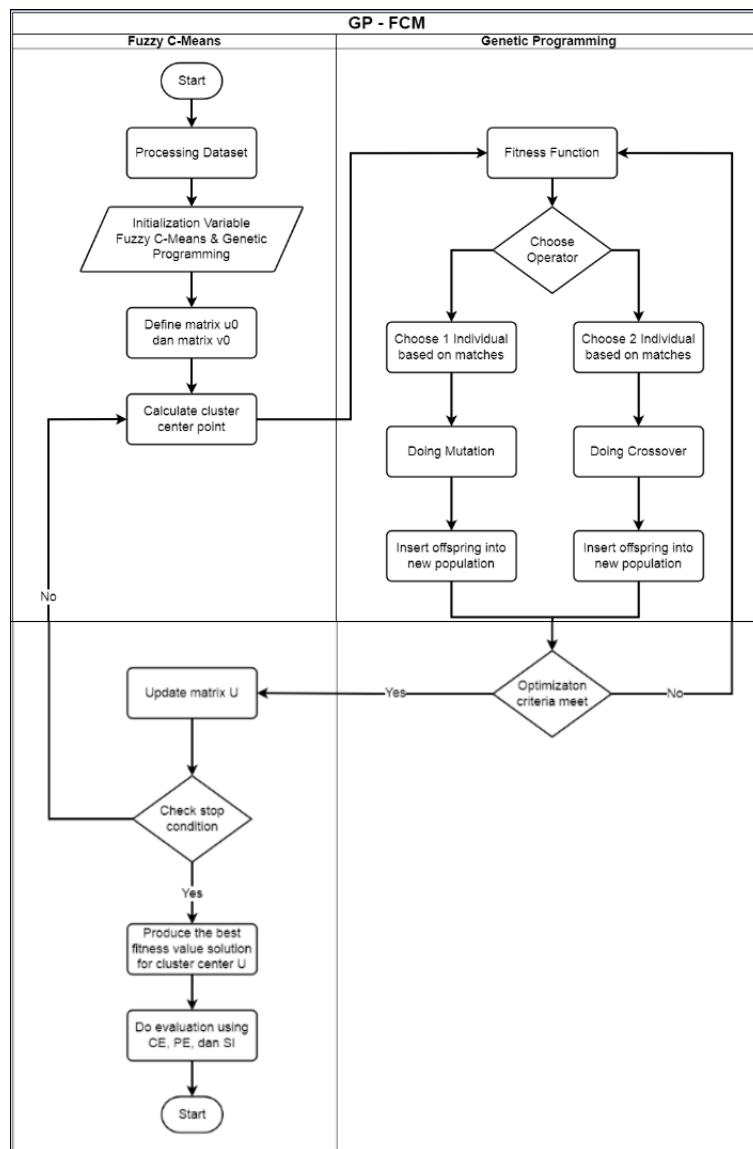
Figure 1. Flowchart of the proposed method

2.1. RFM Model

The concept of RFM was introduced by Bult and Wansbeek (1995) to analyze customer behavior. RFM analysis depends on three variables: Recency, Frequency, and Monetary. These three variables affect the possibility of customer purchases in the future. The determination of attributes in the RFM method is based on the existing transactions. Basically, data retrieval with RFM attributes is based on transaction activities carried out by consumers [18]. Recency is the distance between the last transaction and the current time. The smaller the time interval, the greater the R-value, while Frequency is the total number of transactions carried out during a certain period. The greater the number of transactions, the greater the value of F, and Monetary is the total value of the product in terms of money in a certain period. The greater the products value, the greater the value of M.

Segmentation is the process of dividing customers into several customer loyalty categories to build a marketing strategy. Customer segmentation is divided into six characteristics based on the RFM value, as shown in Table 1.

Table 1. Customer Characters Based On RFM Value [19]

Customer Class	Characteristics
Superstar	a. High loyalty customers
	b. High monetary value.
	c. High frequency.
	d. Has the highest transaction.
Golden customer	a. Has the second-highest monetary value.
	b. High frequency.
	c. Has an average transaction.
Typical customer	Has an average monetary value and an average transaction.
Occasional customer	a. The second lowest monetary value after dormant customers.
	b. The lowest recency value.
	c. Highest transaction.
Everyday shopper	a. Has increased transactions.
	b. Relatively low transaction.
	c. Has a moderate to low monetary value
Dormant customer	a. Has the lowest frequency and monetary.
	b. The lowest recency value.

2.2. Fuzzy C-Means

Fuzzy c-means is a clustering method that allows one part of the data to have two or more clusters. This method is often used in pattern recognition. The FCM clustering algorithm produces a membership value between 0 and 1, indicating the degree of membership for each object for each cluster [20]. This is based on the minimization of the objective function, which can be seen in equation (1).

$$J_t = \Sigma_{i=1}^{N} \Sigma_{j=1}^{c} \mu_{ij}^{m} ||x_i - c_j||^2 \tag{1}$$

Where m $(m > 1)$ is a scalar called the weighting exponent and controls for the fuzziness of the cluster, m is set to a value of 2.00. Where μ_{ij} is membership degree of x_i in cluster j. x_i is data dimension and c_j is cluster center dimension as well as $||x_i - c_j||$ = Euclidean distance between x_i and c_j. Generally, the measurement of data point distance measurement (x_i) to the cluster center (c_j) based on similarity measurements [10]. One of the measurements used is the Euclidean Distance formula, which can be seen in equation (2).

$$||x_i - c_j|| = \sqrt{\Sigma_{j=1}^{c} (x_i - c_j)^2} \tag{2}$$

2.3. Genetic Programming

Genetic Programming (GP) is a model of programming that uses the ideas and terminology of biological evolution to deal with complex problems [21]. First introduced by J. Koza, from several programs that usually function as small programs in larger applications, the most effective programs will survive and compete or cross-breed with other programs to get closer to the required solution approach. Genetic Programming can produce computer programs represented in a tree or graph structure. The chromosomes have different lengths in one population because the resulting daughter chromosomes can be longer or shorter than the parent chromosomes. According to Suyanto, the preparatory steps for Genetic Programming can be seen in Figure 2.

Five major preparatory steps for the basic version of genetic programming [16], among others, are:
1. Set of Terminal.
2. Set of Function.
3. Fitness Measure.
4. Parameters for process control.
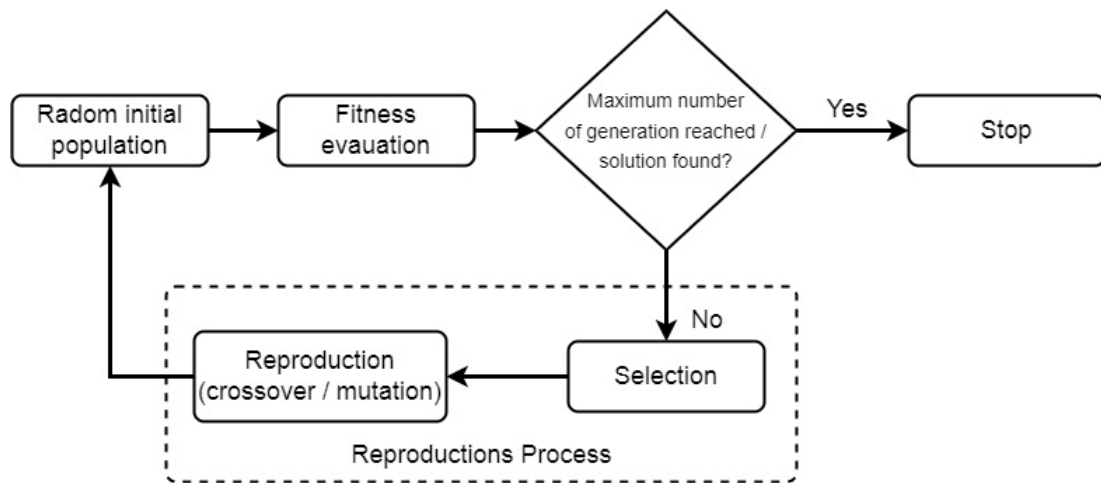5. Termination Criterion and Result Designation.

Figure 2. Flow Chart Of The Genetic Programming Approach [16]

Algorithm 1: Pseudocode for Genetic Programming

1. The initial population is formed randomly based on the terminal scope and primitive functions defined.
2. Perform the process of reproduction in all individuals to form a new population:
   (a) Calculate the match value by running the program.
   (b) Select individuals who will experience reproduction and crossing based on the suitability value.
3. Replace the previous population with a new population; if the conditions cannot be met, return to step 2.

---

3.   RESULT AND ANALYSIS

Experiments were carried out using transaction data at the Cahaya Estetika health clinic, as shown in Table 2. The number of data records used is 55908, using the RFM model to analyze customer behavior based on the transactions made. Furthermore, Genetic Programming is used to determine the optimal number of clusters, and then the data can be grouped into the Fuzzy C-Means algorithm. The experimental process uses RStudio 1.1.463 with version R-3.6.1.

Table 2. Clinic Transaction dataset

Transaction Code	Date of Transaction	Customer Code	Total
R43NA010214001	02/01/2014	S1739	105000
R43NA010214002	02/01/2014	R0651	90000
R43NA010214003	02/01/2014	N2012	125000
R43NA010214004	02/01/2014	S8520	90000
R43NA010214005	02/01/2014	S9805	90000
…	…	…	…
R43NA010214049	02/01/2014	Z0075	90000
R43NA010214001	02/01/2014	S1739	105000

Then carry out the process of determining the RFM attribute for the attribute or recency variable obtained from the sales transaction date field from the period range of December 2014, while the date that is used as a reference for determining the recency variable is December 31, 2014, so that it can be seen the period of the time span of each customer in purchasing during January 2014 to December 2014. For the attribute or frequency variable obtained from the Customer Code field, through the Customer Code, it can be seen how many times the customer made a purchase in that period. The more frequently customers make transactions, the greater the frequency value. In contrast, attributes or monetary variables are obtained from the amount field by adding up all payments from each customer every time they make a purchase within a predetermined period. The results of changes in transaction data can be seen in Table 3.

Table 3. Dataset After RFM Model Applied

Customer Code	Recency	Frequency	Monetary
A0005	325	2	115000
A0011	147	29	2801000
A0015	226	2	275000
A0017	181	8	535000
A0032	230	6	365000
…	…	…	…
A0163	259	2	140000
A0164	166	16	1272500

From the data converted into the RFM model, the dataset will be normalized using min-max normalization as in equation (3). Min-max normalization is a normalization process by performing a linear transformation on the original data to equalize the comparison value between the data before and after the process. Notation A is the data value that will be calculated by taking the minA and maxA values, the smallest values from the data set. For the notation, D and C are the desired conversion range; in this case, the conversion range will be used between 0 to 1.

$$B = \left(\frac{(A - minA)}{maxA - minA} \right) * (D - C) + C \tag{3}$$

After the normalization process is carried out, there are records that produce a value of 0. Records with a value of 0 will be deleted so that the number of records becomes 11746. The results can be seen in Table 4. Normalization is done because the range of values between RFM attributes is too far.

Table 4. Data After Normalization Applied

Customer Code	Recency	Frequency	Monetary
A0005	0.828193833	0.006666667	0.014276847
A0011	0.044052863	0.186666667	0.347734327
A0015	0.392070485	0.006666667	0.034140286
A0017	0.193832599	0.046666667	0.066418374
A0032	0.409691630	0.033333333	0.045313470
…	…	…	…
A0163	0.537444934	0.006666667	0.017380509
A0164	0.127753304	0.100000000	0.157976412

To perform fuzzy c-means calculations in this study using RStudio with the ppclust package, for initialization in the Fuzzy C-Means calculation, the number of clusters is determined to be 6 clusters; this value is determined based on the characteristics of the RFM value [19]. Details of initial values for fuzzy c-means are as follows:

1. Total Cluster : 6
2. Weighting : 2
3. Max Iteration : 1000
4. Least Error : 10-5 or 0,00001

After calculating the fuzzy c-means matrix, the center point of the cluster will also change as in Table 5.

Table 5. Matrix Final Result Of FCM Cluster Center

	Recency	Frequency	Monetary
Cluster 1	0.62733879	0.007586154	0.02924597
Cluster 2	0.38166979	0.011721047	0.03806050
Cluster 3	0.18404324	0.019746756	0.05093926
Cluster 4	0.05086348	0.067068823	0.14634308
Cluster 5	0.89451844	0.002251011	0.01835804

The results of the visualization of clustering using Fuzzy C-Means on customer transaction data with the RFM model can be seen in Figure 3.
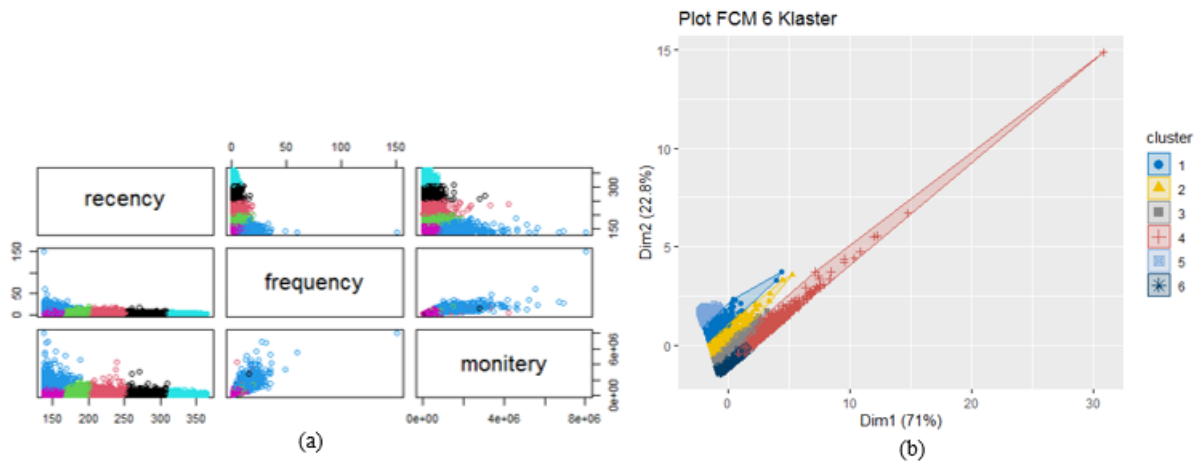
Figure 3. Visualization result using Fuzzy C-Means

After the clustering process uses FCM, the data is returned to the form of an RFM model. Furthermore, an assessment process will be carried out to determine the segment of the customer by taking the minimum and maximum values from each existing cluster. Then the value will be obtained as in Table 6.

Table 6. Segmentation Results Using FCM

Cluster	Min	Max
1	121	144
2	121	244
3	221	344
4	234	444
5	121	143
6	321	443

The conclusion can be drawn from Table 6 that Cluster 1 is an Everyday Shopper customer class because there is an increase in the monetary value. Cluster 2 is an Occasional Customer class because there is an increase in the recency value. Cluster 3 is a Typical Customer class because it has an average monetary and transaction value. Cluster 4 is a Superstar Customer class because it has the highest value. Cluster 5 is a class of Dormant Customers because it has the lowest value. Finally, cluster 6 is the Golden Customer class because it has the second-highest value.

For the GP-FCM experiment, the dataset used is the same as the one used in the FCM algorithm. In addition, the dataset used is the same as one of the normalized with a total of 11746 data lines. In carrying out the genetic programming calculation process, the Euclidean distance formula will be used, which will then be evolved until there is the best model of a population size of 50.

After the process is carried out using GP, the matrix of the initial center point of the cluster from the FCM is obtained, as shown in Table 7.

Table 7. Final Result Of Center Cluster Matric Using GP-FCM

Cluster	Recency	Frequency	Monetary
1	0.0004851637	0.0004852637	0.0773330158
2	0.7005763745	0.7005763745	0.0002219379
3	0.3642414951	0. 3642414951	0.0001864895
4	0.4968076229	0.4968076229	0.0037005056
5	0.0280230550	0.0280230550	0.0062146455
6	0.0675541928	0. 0675541928	0.0003702817

The results of the visualization of clusterization using genetic programming & fuzzy c-means on customer transaction data with the RFM model can be seen in Figure 4.
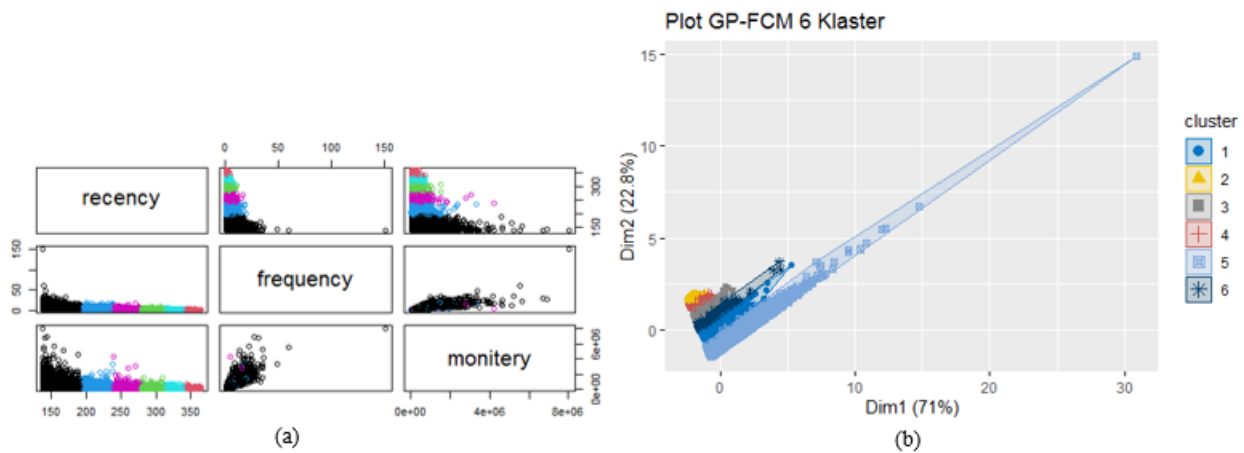
Figure 4. Visualization Results Using Genetic Programming Fuzzy C-Means

Furthermore, an assessment process will be carried out to determine the segment of the customer by taking the minimum and maximum values from each existing cluster. Then the value will be obtained as in Table 8.

Table 8. Segmentation results using GP-FCM

Cluster	Min	Max
1	111	144
2	111	244
3	211	344
4	234	444
5	111	143
6	311	443

In table 8, it can be concluded that Cluster 1 is an Everyday Shopper customer class because there is an increase in the monetary value. Cluster 2 is an Occasional Customer class because there is an increase in the recency value. Cluster 3 is a Typical Customer class because it has an average monetary and transaction value. Cluster 4 is a Superstar Customer class because it has the highest value. Cluster 5 is a Dormant Customers class because it has the lowest value. Therefore, Cluster 6 is the Golden Customer class because it has the second-highest value.

Furthermore, an evaluation of the results of clustering will be carried out using Partition Coefficient (PC), Classification Entropy (CE), and Silhouette Index (SI) methods can be seen in Table 9.

Table 9. FCM and GP-FCM validation comparison

Algorithm	PC	CE	SI
FCM	0.6168213	0.7633299	0.6851368
GP-FCM	0.8487171	0.3069980	0.8798065

Table 9 shows that the CE validity test is close to 0 in the GP-FCM algorithm compared to CE in the FCM algorithm. To test the validity of the PC on the GP-FCM algorithm, it is close to 1 compared to the FCM algorithm. This indicates that the use of GP-FCM in classifying data is better than the FCM algorithm. However, in the Silhouette Index test, the GP-FCM algorithm is closer to the value 1 compared to FCM; this indicates that the data is not in the right cluster.

The results obtained in this study will be compared with previous studies [15]; the silhouette index value generated by GA-FCM is 0.655, while GP-FCM is 0.8798. the silhouette index value generated by the GA-FCM is 0.655, while the GP-FCM is 0.8798. It shows that GP-FCM is better than GA-FCM because the closer the silhouette index value is to 1 indicates that the resulting cluster is better. The results of the comparison of the Silhouette index values can be seen in Table 10.

Table 10. GP-FCM and GA-FCM validation comparison

Algorithm	SI
GA-FCM	0.655
GP-FCM	0.8798

4. CONCLUSION

Based on the results of research that has been carried out from the problem analysis stage and literature review to the results evaluation stage, the proposed methods evaluation results can be seen. This quality improvement can be seen from the value of the objective function, which is more optimal than the fuzzy c-means algorithm. Testing the genetic programming fuzzy c-means (GP-FCM) algorithm for the Cahaya Estetika clinic data set produces an objective function of 20,3091 while the FCM algorithm is 32,44741. Applying Genetic Programming to Fuzzy C-Means affects the evaluation results of the Fuzzy C-Means method. Cluster validity evaluation using Partition Coefficient (PC), Classification Entropy (CE), and Silhouette Index (SI) also proves that fuzzy c-means genetic programming produces optimal cluster quality compared to the fcm algorithm. The results of the silhouette index on the FCM algorithm produce a value of 0,6851368, and GP-FCM produces a value of 0,8798065; this shows that the grouping data from the fcm algorithm produces a better data cluster than the gp-fcm algorithm. While the results of cluster validity on the FCM algorithm using a PC produce a value of 0,6168213, and CE produces a value of 0,7633299. For cluster validity, the GP-FCM algorithm using PC produces a value of 0,8487171, and CE produces a value of 0,3069980. These results indicate that the GP-FCM algorithm produces better data grouping. Although the proposed model has given better results, further research can be carried out as follows. For further experiments, use a large dataset so that the results can look more significant. Use other evolution-based algorithms for more optimal results.

REFERENCES

[1] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-Based Repurchase Behavior for Customer Classification and Segmentation," *Journal of Retailing and Consumer Services*, vol. 61, no. April, pp. 1–9, 2021.

[2] V. Guerola-Navarro, H. Gil-Gomez, R. Oltra-Badenes, and J. Sendra-García, "Customer Relationship Management and Its Impact on Innovation: A Literature Review," *Journal of Business Research*, vol. 129, no. March, pp. 83–87, 2021.

[3] Z. Soltani, B. Zareie, F. S. Milani, and N. J. Navimipour, "The Impact of the Customer Relationship Management on the Organization Performance," *Journal of High Technology Management Research*, vol. 29, no. 2, pp. 237–246, 2018.

[4] S. Monalisa and A. S. Roza, "Penerapan Association Rules untuk Elemen Cross Selling Pada Sistem Informasi Customer Development," *Techno.Com*, vol. 20, no. 3, pp. 420–429, 2021.

[5] S. Tripathi, A. Bhardwaj, and P. E, "Approaches to Clustering in Customer Segmentation," *International Journal of Engineering & Technology*, vol. 7, no. 3.12, p. 802, 2018.

[6] P. P. Pramono, I. Surjandari, and E. Laoh, "Estimating Customer Segmentation Based on Customer Lifetime Value Using Two-Stage Clustering Method," in *2019 16th International Conference on Service Systems and Service Management, ICSSSM 2019*. IEEE, 2019, pp. 1–5.

[7] N. G. A. Dasriani, M. Mayadi, and A. Anggrawan, "Klasterisasi Lokasi Promosi PMB Dengan Fuzzy C-means Masa Pandemi Covid 19," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 327–336, 2022.

[8] V. Kumar and W. Reinartz, *Customer Relationship Management Concept, Strategy and Tools*, third edit ed. Springer Texts in Business and Economics (STBE), 2018.

[9] A. M. Pater, S. Vari-Kakas, O. Poszet, and I. G. Pintea, "Segmenting Users of an Online Store Using Data Mining Techniques," in *2019 15th International Conference on Engineering of Modern Electric Systems, EMES 2019*, 2019, pp. 205–208.

[10] A. M. Anter, A. E. Hassenian, and D. Oliva, "An Improved Fast Fuzzy C-Means Using Crow Search Optimization Algorithm for Crop Identification in Agricultural," *Expert Systems with Applications*, vol. 118, no. March, pp. 340–354, mar 2019.

[11] A. Seal, A. Karlekar, O. Krejcar, and C. Gonzalo-Martin, "Fuzzy C-Means Clustering Using Jeffreys-Divergence Based Similarity Measure," *Applied Soft Computing*, vol. 88, no. March, pp. 1–11, mar 2020.

[12] D. B. Saputra and E. Riksakomara, "Implementasi Fuzzy C-Means dan Model RFM untuk Segmentasi Pelanggan (Studi Kasus : PT. XYZ)," *Jurnal Teknik ITS*, vol. 7, no. 1, pp. 1–6, 2018.

[13] M. T. Dharmawan, N. Y. Setiawan, and F. A. Bachtiar, "Segmentasi Pelanggan Menggunakan Metode Fuzzy C-Means Clustering Berdasarkan LRFM Model Pada Toko Sepatu (Studi Kasus: Ride Inc Kota Malang)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 2, pp. 1978–1985, 2019.

[14] S. S. Prasetyo, M. Mustafid, and A. R. Hakim, "Penerapan Fuzzy C-Means Kluster untuk Segmentasi Pelanggan E-Commerce Dengan Metode Recency Frequency Monetary (RFM)," *Jurnal Gaussian*, vol. 9, no. 4, pp. 421–433, 2020.

[15] Faza Adhzima, Yandra Arkeman, and Irman Hermadi, "The Clustering Rice Plant Diseases Using Fuzzy C-Means and Genetic Algorithm," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 2, pp. 240–245, 2022.

[16] H. Andersen, A. Lensen, and B. Xue, "Genetic Programming for Evolving Similarity Functions Tailored to Clustering Algorithms," in *2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings*, 2021, pp. 688–695.

[17] R. Syelly, I. Laksmana, T. Komputer, S. Tinggi Teknologi Payakumbuh, and J. Khatib Sulaiman, "Rancang Bangun Sistem Pakar Model Identifikasi untuk Klasifikasi Varietas Unggul Tanaman Gambir Menggunakan Genetic Programming," *AGRIEKSTENSIA: Jurnal Penelitian Terapan Bidang Pertanian*, vol. 19, no. 1, pp. 38–48, jul 2020.

[18] H. Hairani, D. Susilowati, I. Puji Lestari, K. Marzuki, and L. Z. A. Mardedi, "Segmentasi Lokasi Promosi Penerimaan Mahasiswa Baru Menggunakan Metode RFM dan K-Means Clustering," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 275–282, 2022.

[19] N. Puspitasari, J. A. Widians, and N. B. Setiawan, "Customer Segmentation Using Bisecting K-means Algorithm Based on Recency, Frequency, and Monetary (RFM) Model," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 78–83, 2020.

[20] S. T. Wierzchoń and M. A. Kłopotek, "Introduction in: Modern Algorithms of Cluster Analysis," in *Studies in Big Data*, 2018, vol. 34, pp. 1–7.

[21] B. Tran, B. Xue, and M. Zhang, "Genetic Programming for Multiple-Feature Construction on High-Dimensional Classification," *Pattern Recognition*, vol. 93, no. September, pp. 404–417, sep 2019.