Customer Loyalty Clustering Model Using K-Means Algorithm with LRIFMQ Parameters

Aloysius Matz

Institut Sains dan Teknologi Terpadu Surabaya

aloysiusmatz@gmail.com

Abstract— Loyal customers are one of the factors that determine the development of a business. Therefore, businesses need a strategy to keep customers loyal, even making customers who were previously less loyal to become more loyal. The strategy used must be right on target according to customer segmentation. The purpose of this paper is to model a cluster of customer loyalty to help businesses in making the right decisions of marketing strategy. Segmentation is done using the k-means algorithm with LRIFMQ (length, recency, interval, frequency, monetary, quantity) as parameters, and the CLV (customer lifetime value) of each cluster is calculated. Data obtained from PT. XYZ (a company engaged in food processing) for one year (1 January 2019 - 31 December 2019), with 337.739 transactions, and 26.683 customers. AHP (analytical hierarchy process) method is used for LRIFMQ weighting because this method has a consistency index calculation. The silhouette coefficient is used to calculate the cluster quality and determine the optimal number of clusters. The best results are obtained with the silhouette coefficient value of 0,632904 with the number of clusters 6.

Keywords—customer analysis; k-means; LRIFMQ; analytical hierarchy process; AHP; silhouette coefficient.

I. INTRODUCTION

One of the intangible assets owned by a company is a loyal customer [1]. The competition in business is so tight, as an owner must know the customer's needs as well so that customers do not turn away. Companies need to carry out special strategies to establish good relations between the company and its customers. The strategy must be under the customer's need because with the right marketing strategy can increase profits for the company [2].

The step that a company can take is to provide value to customers based on certain criteria that can benefit the company. This concept is known as CLV (customer lifetime value) [3]. CLV is an estimated value. Even so, this value can be used to evaluate the future of customers against the company with data mining techniques in detecting patterns and relationships using data history [4]. One of the models that can measure CLV is RFM (recency, frequency, monetary) [5].

RFM was first introduced by Hughes, which is the most common segmentation method used to identify customer value in the company based on three variables: recency, frequency, dan monetary. Recency is a calculation of how many days from the customer's last transaction up to today. Frequency is a calculation of how many transactions are made during the period, while monetary is how much money has been transacted by the customer [6].

This RFM model was developed by Chang and Tsay with the addition of length variable and is known as the LRFM model (length, recency, frequency, monetary) [7]. Length is the calculation of the distance of days between the first and the last time customer's transaction. In this paper, the author tries to add interval and quantity variables. Interval is the average daily distance from each transaction of the customer, while the quantity is the sum of all items that have been transacted. The addition of these two variables was inserted in the previous theory to LRIFMQ (length, recency, interval, frequency, monetary, quantity). Each of the parameters has a weight that will be determined using the AHP method (analytical hierarchy process).

After assessing customers with LRIFMQ, the next step is to segment customers. This segmentation can be done with the help of data mining techniques known as clustering. Clustering in data mining is the process of forming segments or clusters by looking at the similarities between data based on the parameters given.

Many clustering methods have their advantages and disadvantages. It can be seen in Table I that the majority of previous research uses the k-means algorithm, which is widely used because of its performance quite effective and efficient [8]. Another quite popular algorithm is fuzzy c-means, which has an element of fuzziness that is difficult to solve using the k-means algorithm. Therefore, this paper uses both algorithms.

In research by A.J. Christy show that segmentation performed using k-means requires a shorter time and much less iteration than fuzzy c-means [6]. Also, the average number of silhouette coefficient values using the k-means algorithm is better than the fuzzy c-means. The silhouette coefficient and the arm method or commonly known as the elbow, are used to determine the best number of clusters. Besides, this silhouette coefficient method can also be used to determine the quality of a clustering process.

TABLE I
PREVIOUS RESEARCH

Research	Model	Clustering Method
Hughes (1994)	RFM	-
Miglautsch (2000)	RFM	-
Shih and Liu (2003)	RFM	K-means clustering
Chang and Tsay (2004)	LRFM	Self-organizing maps (SOM)
Hu and Jing (2008)	RFM	K-means clustering
Bin,Peiji,and Dan (2008)	RFM	K-means clustering
Wu ct al. (2009)	RFM	K-means clustering

Research	Model	Clustering Method
Chang et al. (2010)	RFM	K-means clustering
Li et al. (2011)	LRFM	Two-Step clustering
Wei ct al. (2012)	LRFM	SOM
Chen (2012)	RFM	C-means clustering
Kafashpoor and Alizadch (2012)	RFM	Hierarchical Clustering
A.J. Christy et al. (2018)	RFM	K-means, Fuzzy C-means, RM K-means

This paper aims to obtain a cluster model of customer loyalty that is assessed based on CLV with the k-means algorithm and using the silhouette coefficient method to determine the best number of clusters. The customer loyalty cluster model is expected to assist businesses in determining the best marketing strategy.

II. RESEARCH METHODOLOGY

The processes stage in the discussion of this paper uses several steps, following:

A. Data Collection

Sources of the data obtained from PT. XYZ (a company engaged in food processing). The data taken are sales transactions consisting of 2,629,261 transactions and 64,239 customers. While the attributes used are the date of transaction, transaction value, and quantity.

B. Weighting Using Analytical Hierarchy Process (AHP)

AHP method is one of the methods used in determining weight. In this case, the weight that will be determined is LRIFMQ. The steps taken are as follows: [9]
- Prioritize using the pairwise comparison matrix with the pairwise comparison index shown in Table II [5].
- Normalize pairwise comparison matrix.
- Calculates the weight obtained from the average of each row in the pairwise comparison matrix. This calculation will produce W_L, W_R, W_I, W_F, W_M, W_Q. This weight can be described using a 6x1 matrix with W variable. After getting the weight value, the value is multiplied by the initial pairwise comparison matrix before normalization. This matrix multiplication will produce a 6x1 matrix denoted X variable.
- Calculates consistency values based on Equation (1).

$$\lambda_{max} = \frac{1}{n}\sum_{i=1}^{n}(\frac{row\ i\ on\ (X)}{row\ i\ on\ (W)})$$ (1)

- Calculates the value of the consistency index (CI) using Equation (2).

$$CI = \frac{\lambda_{max} - n}{n - 1}$$ (2)

- The value of consistency ratio (CR) calculated using Equation (3).

$$CR = \frac{CI}{RI}$$ (3)

Where *RI* for the number *n = 6* is *1.24*. A consistency ratio below 0.1 is required for weights to be considered valid and can be used [9].

TABLE II
PAIRWISE COMPARISON INDEX

Definition	Index	Definition	Index
Equally important	1	Equally important	1/1
Equally or slightly more important	2	Equally or slightly less important	1/2
Slightly more important	3	Slightly less important	1/3
Slightly too much more important	4	Slightly to way less important	1/4
Much more important	5	Way less important	1/5
Much too far more important	6	Way too far less important	1/6
Far more important	7	Far less important	1/7
Far more important to extremely more important	8	Far less important to extremely less important	1/8
Extremely more important	9	Extremely less important	1/9

C. Data Preparation

The steps taken in this data processing stage are as follows:
- Removes duplicate data.
- Removes inconsistent data.
- Removes customer data that only make one transaction because the interval parameter is required to have at least two transactions.
- Transform data into LRIFMQ:
  a) L (Day) = Last transaction date – first transaction date
  b) R (Day) = 31/12/2019 – last transaction date
  c) I (Day) = Average distance of days between transactions
  d) F (Transaction) = number of transactions that have been made
  e) M (Rupiah)= Total monetary value of all transactions
  f) Q (Pc)= Total quantity of all transactions
- Normalize using min-max:
  Normalization of data using the min-max method, with a range of 0-1. Equation (4) used to calculate min-max normalization.

$$V_1 = \frac{V\text{-}min}{max\text{-}min}\ (D\text{-}C) + C$$ (4)

Where:
V_1 = min-max value
min = smallest value in the attribute
max = largest value in the attribute
D = new max value (1)
C = new min value (0)

D. Clustering Process

The main clustering method used is k-means, while fuzzy c-means is used as a comparison. The steps of clustering using k-means are as follows [10]:

- Determine the number of clusters k
- Determine the centroid point randomly as many as clusters k.
- Calculate the flat distance to centroid using the Euclidean distance formula, as in Equation (5).

$$D_{xy} = \sqrt{\sum_{i=1}^{n} (xi - yi)^2} \tag{5}$$

- Renew centroid points based on the average value in each cluster.
- Repeat points c - d until all centroid points are convergent/immovable.
- The k-means algorithm is run several times to search for global optima.

The steps of clustering with fuzzy c-means are as follows:

- Choose the center point k randomly (the number of k is predetermined).
- Calculate fuzzy membership μ_{ij}, based on Equation (6).

$$\mu_{ij} = \frac{1}{\sum_{c=1}^{k} \left(\frac{d_{ij}}{d_{cj}}\right)^{\frac{2}{m-1}}} \tag{6}$$

Where:
$$d_{ij} = \| x_i - v_j \|$$
$$d_{cj} = \| x_i - v_c \|$$

- Update the centroid on each cluster v_j , using Equation (7).

$$v_j = \frac{\sum_{j=1}^{n} u_{ij}{}^m x_i}{\sum_{j=1}^{n} u_{ij}{}^m} \tag{7}$$

- Repeat until the value ε is between 0 and 1, where ε is the limit specified between 0 and 1 based Equation (8).

$$\varepsilon > max_{ij}\left\{|u_{ij}^{(k+1)} - u_{ij}^{k}|\right\} \tag{8}$$

E. Determine the Optimal Cluster Number

Determine the optimal number of clusters using the silhouette coefficient method and the elbow method based on the following steps:

- Calculate the average distance of data objects with other data objects in 1 cluster and will get the value $a(o)$, based on Equation (9).

$$a(o) = \frac{\sum_{o`\in Ci, O \neq O`} dist(o, o')}{|Ci| - 1} \tag{9}$$

- Calculate the average distance between data objects and data objects in other clusters. From the average distance of each cluster, take the least, and will get the value b(o) based on Equation (10).

$$b(o) = min_{C_j; 1 \leq j \leq k, j \neq i}\left\{\frac{\sum_{o`\in c_j} dist(o, o`)}{|C_j|}\right\} \tag{10}$$

- Calculate the silhouette coefficient with Equation (11).

$$Si = \frac{b(o) - a(o)}{max\{a(o), b(o)\}} \tag{11}$$

Elbow method is a method that determines the optimal number of clusters using visuals, as shown in Figure 1, there is a significant change in cluster point 3. If you use the elbow method as a reference for determining the optimal number of clusters, the optimal number of clusters obtained is three based on Equation (12).

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_k} |x_i - c_k||^2 \tag{12}$$

Where:
K = number of clusters
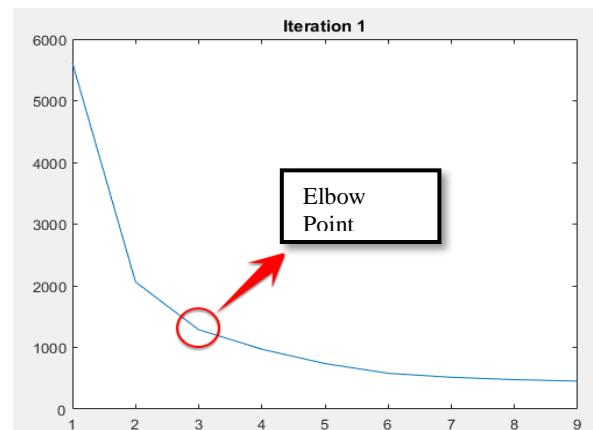x_i = i[th] data
c_k = centroid cluster k

Figure 1. Example for Elbow Method Graph

F. Calculate Customer Lifetime Value (CLV)

Calculate CLV value based on Equation (13).

$$CLV = L * W_L - R * W_R - I * W_i + F * W_F + M * W_M + Q * W_Q \qquad (13)$$

Where:

- L, R, I, F, M, Q is the average customer value
- $W_L, W_R, W_i, W_F, W_M, W_Q$ is the weights of each LRIFMQ.
- The values of R and I are as a deduction due to their inverse values.

G. Evaluation and Analysis of Results

Evaluate and analyze the results of all trials. If the results obtained are not appropriate due to a trial error, then the step can be repeated to the point where the error occurred. The best clustering is the result of trials with the silhouette coefficient that is closest to one. I also performed an analysis of the type of customer for each cluster.

III. RESULT AND DISCUSSION

In the initial preparation stage, the AHP process is carried out to determine the weight of each trial parameter. Giving weight using the pairwise comparison index is done subjectively because it depends on the company how to see the interests of each LRIFMQ parameter. Weights obtained for each $W_L, W_R, W_i, W_F, W_M, W_Q$ are 0,0672, 0,0395, 0,2519, 0,2850, 0,2082, 0,1481 with a value of consistency ratio 0,039. With a consistency ratio below 0.1, the weight is considered valid and can be used.

TABLE III
DATA LRIFMQ

Cust No.	L	R	I	F	M	Q
1	357	3	2,73	132	11129041	35805,4
2	357	0	2,2	163	4993819	83651,6
3	62	298	1,35	47	454513	1410
4	62	299	1,35	47	178907,5	559
5	62	299	1,35	47	580652,5	1769
6	62	298	1,38	46	540590	1656
7	62	298	1,35	47	158350,5	495
8	359	3	0,67	533	36590709	16717,8
…						
25683	62	299	1,35	47	179106,5	560

TABLE IV
DATA LRIFMQ NORMALIZED RESULT

Cust No.	L	R	I	F	M	Q
1	0,9623	0,0083	0,0075	0,1037	0,0252	0,0069
2	0,9623	0	0,0061	0,1284	0,0113	0,0161
3	0,1671	0,8232	0,0037	0,0359	0,0010	0,0003
4	0,1671	0,8260	0,0037	0,0359	0,0004	0,0001
5	0,1671	0,8260	0,0037	0,0359	0,0013	0,0003
6	0,1671	0,8232	0,0038	0,0351	0,0012	0,0003
7	0,1671	0,8232	0,0037	0,0359	0,0004	0,0001
8	0,9677	0,0083	0,0019	0,4234	0,0828	0,0032
…						
25683	0,1671	0,8260	0,0037	0,0359	0,0004	0,0001

A. Trial Using K-Means Algorithm

The trial using the k-means algorithm was repeated 40 times. Each repetition is carried out clustering, starting from k = 2 to k = 9. The reason is dismissed at number 9 is because the value of the silhouette coefficient decreases as the number of clusters increases k.

The silhouette coefficient is used to measure the quality of the clustering process, as well as to determine what the optimal number of clusters is. The silhouette coefficient value has a range between -1 to 1. If the value is close to 1, then the quality of the resulting cluster is getting better. While close to -1, the cluster quality gets worse [12].

TABLE V
TRIAL RESULTS USING K-MEANS

Number of Clusters (k)	Iteration	Silhouette Coefficient
2	1	0,716160
3	1	0,650548
4	2	0,622218
5	1	0,618958
6	5	0,632904
7	35	0,567067
8	13	0,583168
9	10	0,550509

Table V shows the results of trials with the highest silhouette coefficient for each number of clusters from 40 iterations. The silhouette coefficient results that are closest to 1 are in the number of clusters 2, iteration 1, with a value of 0.71616.
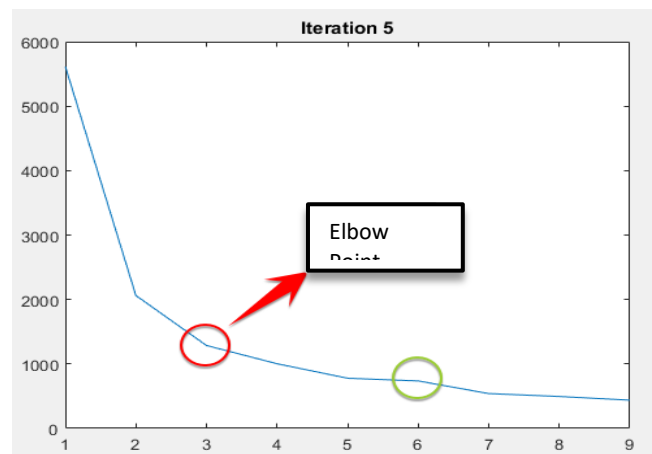

Figure 2. Elbow Method Graph

Whereas the elbow method graph shown in Figure 2 shows the fracture point that previously dropped significantly and then changes to the point at the number of clusters 3. What is unique in this graph is the increase in the number of clusters point 6.

Table VI and Table VII show the number of members in each cluster, along with the CLV value, whether using AHP or not. It can be seen in the number of clusters 2 that cluster 1 gets the first rank when seen from its CLV value. This shows that there are more potentially loyal customers than potentially non-loyal customers. It can also be seen that the AHP weight value does not change the cluster rating. For the number of clusters 3, the number of members in the first rank is more than the number of members in the next rank. This means the number of potential customers loyal to PT. XYZ tends to be more than those that don't.

TABLE VI
TEST OF K-MEANS WITH CLUSTER NUMBER 2

k	Member	CLV	CLV (AHP)	Rank
1	13852	0,545412	0,018295	1
2	11831	-0,45978	0,009742	2
Number of Cluster = 2, Average Silhouette = 0.716160				

TABLE VII
TEST OF K-MEANS WITH CLUSTER NUMBER 3

k	Member	CLV	CLV (AHP)	Rank
1	6711	-0,68532	-0,03534	3
2	11516	0,639201	0,02665	1
3	7456	-0,08671	-0,02232	2
Number of Cluster = 3, Average Silhouette = 0.650548				

Figure 3. Comparison of LRIFMQ k=2 (Without AHP Weighting)

A comparison of LRIFMQ values for each cluster without using AHP weights is shown in Figure 3. The first cluster has a higher L, I, F, M, and Q values and lowers R compared to the second cluster. A high value of I is a minus point for loyalty, but overall the loyalty of the first cluster is higher. A comparison of LRIFMQ values using AHP weights is shown in Figure 4. When compared to the LRIFMQ value that does not use weights, the value of R goes down far enough inversely with the value of I and F, which rises quite significantly. Even so, for the number of cluster 2, the AHP weighting does not change the cluster loyalty rating.
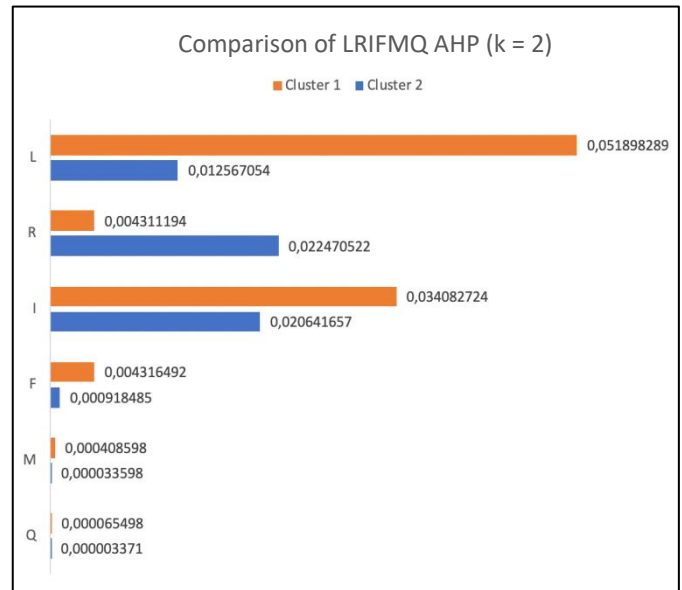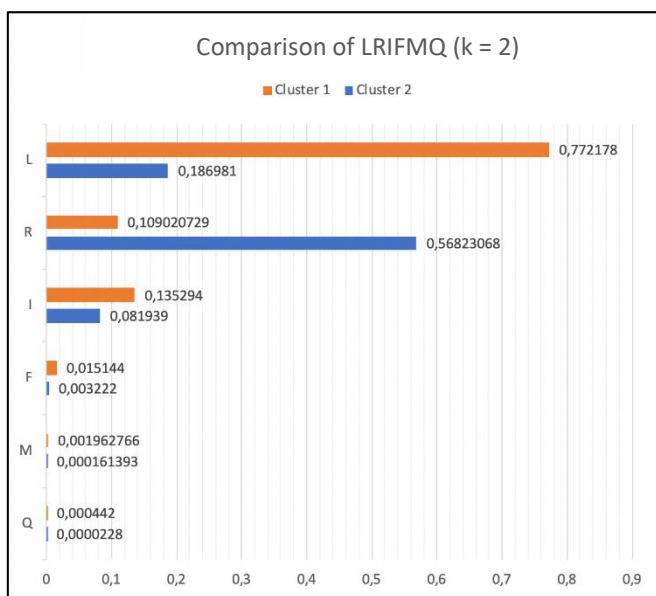

Figure 4. Comparison of LRIFMQ k=2 (With AHP Weighting)

Based on Table V, the silhouette coefficient value for the number of clusters 3, 4, and 5 decreases along with the clusters increase. However, there is a difference in the number of clusters 6, where the value of the silhouette coefficient increased to 0.632904, higher than the number of clusters 4. Details of clustering results from cluster 6 are shown in Table VIII. In the number of clusters 6, the first rank is the sixth cluster, both using AHP weights or not. This clustering process can also separate the most two loyal customers, viewed from the CLV value far adrift with the value of other cluster CLV. Furthermore, for ranks 2 and 3, whether using AHP weights or not, are in the same cluster. However, there is a change in rank after weighting, where the fifth cluster was previously ranked 4, becoming rank 6. The second cluster before weighting was rank 5 to rank 4. The third cluster before weighting was rank 6 to rank 5.

TABLE VIII
TEST OF K-MEANS WITH CLUSTER NUMBER 6

k	Member	CLV	CLV (AHP)	Rank	Rank (AHP)
1	5242	0,045456	-0,01176	3	3
2	3084	-0,11122	-0,01469	5	4
3	6720	-0,68481	-0,03564	6	5
4	9227	0,744866	0,042299	2	2
5	1408	-0,03931	-0,10023	4	6
6	2	2,240244	0,308585	1	1
Number of Cluster = 6, Average Silhouette = 0.632904					

58

A comparison of LRIFMQ values without using AHP weights is shown in

Figure 5. It is clear that cluster 6 has a larger L, F, M, Q, and R, and I value compared to other clusters. While for the last rank, the third cluster has an R-value that is quite large compared to other clusters. The LRIFMQ comparison after weighting using AHP is shown in Figure 6. The third cluster, which previously had a high R-value, turned lower and made this cluster rise to rank 5. Meanwhile, the fifth cluster that already had a top *I* value became even more elevated and made this cluster go down to the last grade.
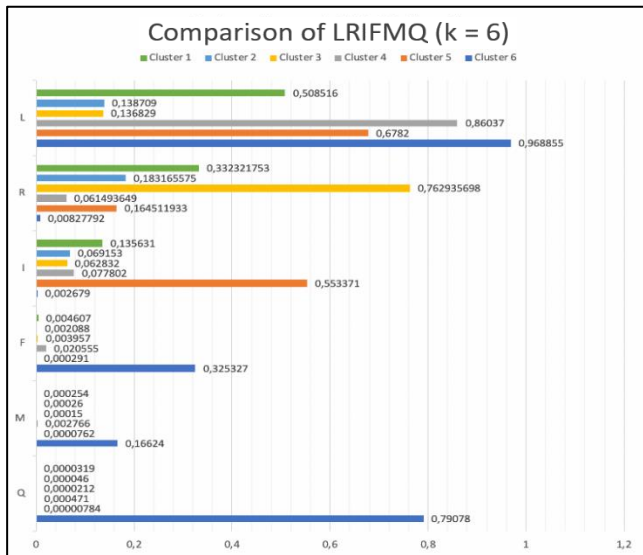

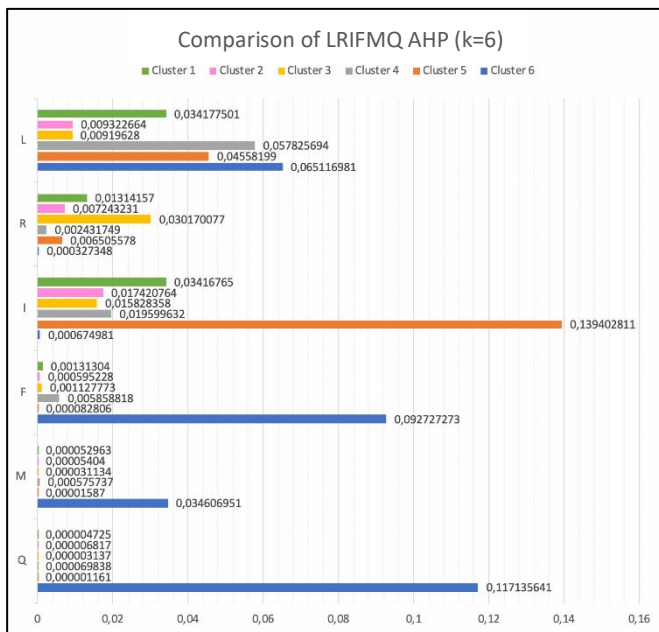Figure 5. Comparison of LRIFMQ k=6 (Without AHP Weighting)

Figure 6. Comparison of LRIFMQ k=6 (Without AHP Weighting)

H. Trial Using Fuzzy C-Means Algorithm

The trial using the fuzzy c-means algorithm is carried out with the same treatment as the experiments using the k-means algorithm, which is done as many as 40 iterations and the number of k = 2 to k = 9.

Figure 7 shows the LRIFMQ comparison without using AHP weights. The second cluster has a higher L, I, F, M, Q values, and lower R values than the first cluster.
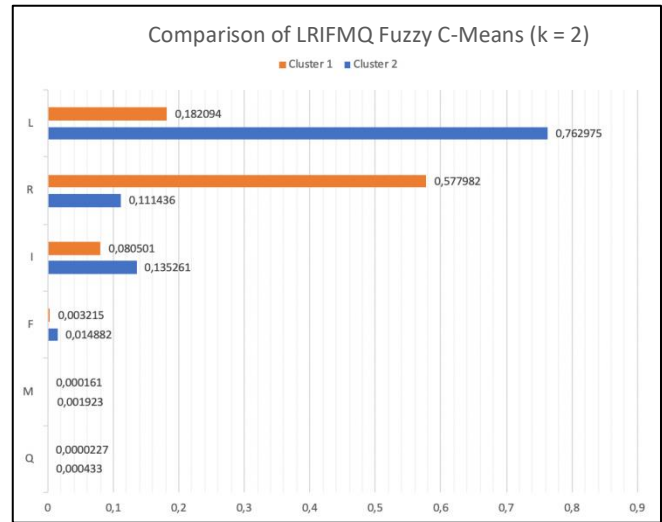

Figure 7. Comparison of LRIFMQ Fuzzy C-Means k=2 (Without AHP Weighting)

Table IX shows the results of trials with the highest silhouette coefficient for each number of clusters from 40 iterations. The silhouette coefficient obtained the closest result to 1 is the number of clusters 2, iteration 1, with a value of 0.716398.

TABLE IX
TRIAL TEST RESULTS USING FUZZY C-MEANS

Number of Clusters (k)	Iteration	Silhouette Coefficient
2	1	0,716398
3	1	0,642447
4	31	0,607515
5	1	0,564791
6	24	0,581685
7	17	0,530723
8	4	0,486646
9	35	0,446805

Table X shows the number of members for each cluster, along with the CLV value, whether using AHP or not. It can be seen that the second cluster gets the first rank when seen from its CLV value. This shows that there are more potential loyal customers than potentially non-loyal customers, seen from the number of members in the second cluster of 14171. This number increases, when compared to the k-means algorithm, where the number of members ranked first is 13852.

TABLE X
FUZZY C-MEANS TRIAL TEST RESULT WITH THE NUMBER OF CLUSTER 2

k	Member	CLV	CLV (AHP)	Rank	Rank (AHP)
1	11512	-0,47299	-0,0299438	2	2
2	14171	0,533515	0,01750469	1	1
Number of Cluster = 2, Average Silhouette = 0.716398					

k	Member	CLV	CLV (AHP)	Rank	Rank (AHP)
4	3719	-0,24517	-0,0271002	5	4
5	1421	-0,05894	-0,1018146	3	1
6	2830	-0,09431	-0,0138249	4	6
Number of Cluster = 6, Average Silhouette = 0.581685					

Table IX It is shown that the greater the number of clusters, the smaller the value of the silhouette coefficient. But in the number of clusters 6, the value of the silhouette coefficient is 0.581685, higher than the number of clusters 5. The value is obtained after the 24th iteration. The trial results in details with cluster number 6, 24th iteration, shown in Table XI. The cluster that gets the first rank in a process that does not use AHP weights is the third cluster. But when using AHP weights, the first rank changes to the fifth cluster. In trials using fuzzy c-means, the highest silhouette coefficient value for the number of clusters is still lower than using k-means. Also, the number of members in the trial results using fuzzy c-means is not as significant a part as trials using k-means.
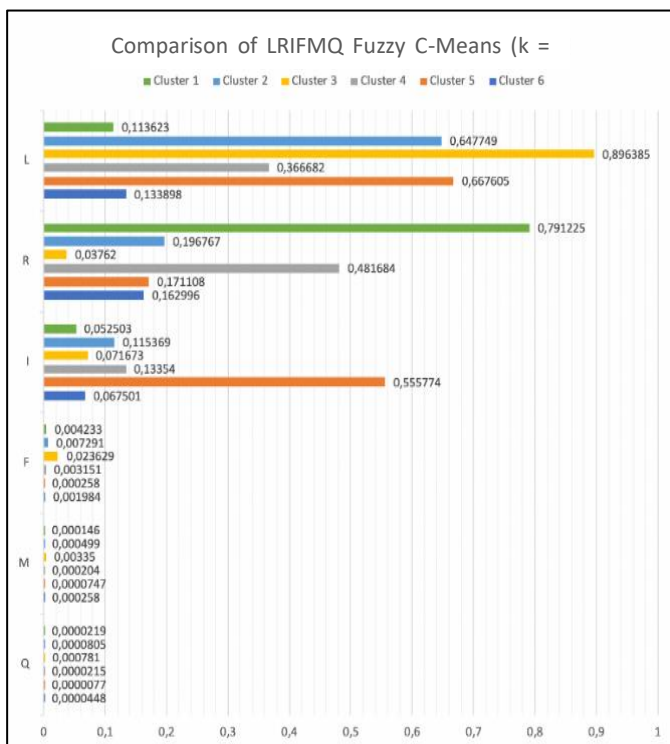

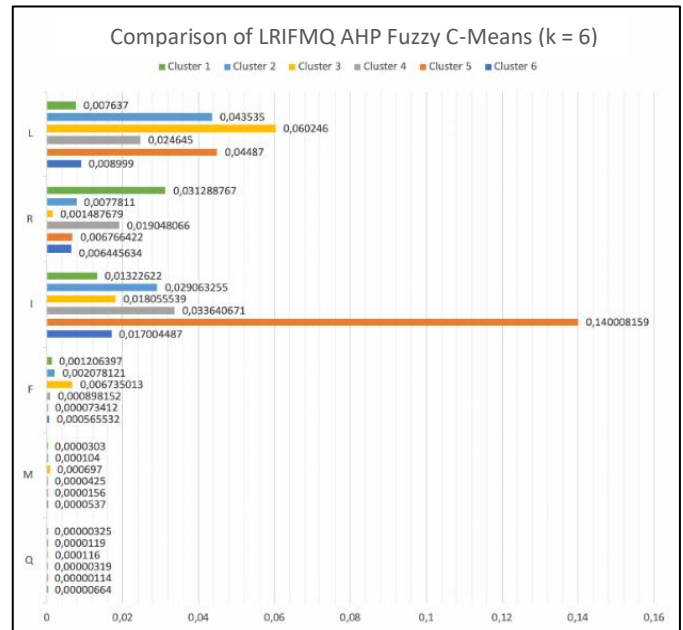Figure 8. Comparison of LRIFMQ Fuzzy C-Means k=6 (Without AHP Weighting)

Figure 9. Comparison of LRIFMQ Fuzzy C-Means k=6 (Using AHP Weighting)

A comparison of LRIFMQ values without using AHP is shown in Figure 8. It appears that the third cluster has significantly higher L and F values than the other clusters. In addition, the third cluster also has R, and I value compared to other clusters. The last rank is in the first cluster because the L value is on average lower, and the R-value is significantly higher than the other cluster. After weighting using AHP, the fifth cluster that was previously ranked 3rd changed to 1st place. This is because the value of I increased significantly. A comparison of LRIFMQ values after weighting can be seen in Figure 9.

TABLE XI
FUZZY C-MEANS TRIAL TEST RESULTS WITH THE AMOUNT OF CLUSTER 6

k	Member	CLV	CLV (AHP)	Rank	Rank (AHP)
1	5731	-0,7257	-0,0356384	6	5
2	4724	0,343482	0,00888484	2	3
3	7258	0,814852	0,0482511	1	2

IV. CONCLUSION

There are two trials conducted in this study, using the k-means algorithm and fuzzy c-means. The trial using k-means can separate a group of customers who are very loyal to the number of clusters 6 with two members. While using fuzzy c-means for the same number of clusters, the number of members in the most loyal cluster is 1421. If the business wants more specific grouping, then the results of this k-means will be quite helpful. In this case, even though the highest silhouette coefficient value is in the number of clusters 2, this value cannot be used as a benchmark that the number of clusters 2 is a cluster that can be used in business decision making. The value of the silhouette coefficient and the elbow

method graph that increases between decreases can also be used as consideration. Meanwhile, the use of AHP weights can also change the CLV rating. Therefore, weighting using AHP is important to do and adjusted to the level of importance of each business. Suggestions for further research are to increase the number of clustering iterations. This needs to be done because if you look at the research that has been done, the highest silhouette coefficient value can only be obtained after dozens of iterations. Besides, the addition of variables other than LRIFMQ also needs to be considered. Examples of variables that can be used are age, region, economic level, and others. The addition of these variables can further assist businesses in getting more specific customer clusters.

V. References

[1] E. Winarti and T. Primadiana, "ANTARA KUALITAS PELAYANAN JASA EKSPEDISI DENGAN LOYALITAS KONSUMEN," JURNAL LENTERA BISNIS, vol. 5, pp. 62-76, 2015.

[2] T. Kristanto and R. Arief, "ANALISA DATA MINING METODE FUZZY UNTUK CUSTOMER RELATIONSHIP MANAGEMENT PADA PERUSAHAAN TOUR & TRAVEL," in Seminar Nasional Sistem Informasi Indonesia, 2013.

[3] S. Monalisa, "Klasterisasi Customer Lifetime Value Dengan Model LRFM Menggunakan Algoritma K-Means," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), vol. 5, no. 2, pp. 247-252, 2018.

[4] F. A. Buttle and S. Maklan, Customer Relationship Management: Concepts and Technologies, 2016.

[5] A. Parvaneh, H. Abbasimehr, and M. J. Tarokh, "Integrating AHP and Data Mining for Effective Retailer Segmentation Based on Retailer Lifetime Value," Journal of Optimization in Industrial Engineering, vol. 11, pp. 25-31, 2012.

[6] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," Journal of King Saud University – Computer and Information Sciences, 2018.

[7] A. A. Zoeram, "A New Approach for Customer Clustering by Integrating the LRFM Model and Fuzzy Inference System," Iranian Journal of Management Studies, vol. 11, no. 2, pp. 351-378, 2018.

[8] D. Suyanto, Data Mining Untuk Klasifikasi dan Klasterisasi Data, Informatika, 2019.

[9] A. T. Rahman, Wiranto and R. Anggrainingsih, "Coal Trade Data Clusterung Using K-Means," ITSMART: Jurnal Ilmiah Teknologi dan Informasi, vol. 6, no. 1, 2017.

[10] N. P. E. Merliana, Ernawati and A. J. Santoso, "ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING (SENDI_U)," PROSIDING SEMINAR NASIONAL MULTI DISIPLIN ILMU & CALL FOR PAPERS UNISBANK, p. 978–979–3649–81–8, 2008.

[11] B. Santoso, I. Cholissodin and B. D. Setiawan, "Optimasi K-Means untuk Clustering Kinerja Akademik Dosen Menggunakan Algoritme Genetika," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, pp. 1652-1659, 2017.

[12] D.-C. Li, W.-L. Dai and W.-T. Tseng, "A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business," Expert Systems with Applications, pp. 7186-7191, 2011.