Jurnal CoreIT, Vol. 8, No.1, June 2022
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Determination of Discounts Using K-Means Clustering with RFM Models in Retail Business

Rina Sisca Zebua[1], Rahmat Izwan Heroza[2*], Ari Wedhasmara[3], Ken Ditha Tania[4], Monterico Adrian[5], Lovinta Happy Atrinawati[6]
[1,2,3,4]Information Systems Department, Universitas Sriwijaya
Jl. Raya Palembang - Prabumulih Km. 32 Indralaya, Sumatera Selatan
[5]Information Technology Department, Telkom University
Jl. Telekomunikasi No. 1, Terusan Buahbatu - Bojongsoang, Bandung
[6]Information Systems Department, Institut Teknologi Kalimantan
Jl. Soekarno Hatta No.KM 15, Karang Joang, Balikpapan

[1]rinasiscazebua@gmail.com, [2]rahmatheroza@unsri.ac.id*(Correspinding Author),
[3]a_wedhasmara@unsri.ac.id, [4]ken_tania@unsri.ac.id,
[5]monterico@telkomuniversity.ac.id, [6]lovinta@lecturer.itk.ac.id

***Abstract*—** Intense competition in the business sector motivates every company to manage services to regular consumers to the fullest. Increase customer loyalty can be done by grouping customers into several groups and determine appropriate and effective marketing strategies for each group. This study aims to propose the right targeting of discounts that can increase customer loyalty in the retail business. Customer grouping uses data mining techniques with the Cross-Industry Standard Process for Data Mining (CRISP-DM) method, which is divided into six phases, namely business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The formation of this cluster uses k-means clustering method and is based on RFM (recency, frequency, monetary) analysis. From the results of the silhouette test on 2734 transaction data from 210 customers of PT. XYZ from October 2019 to March 2020, three customer clusters were formed. From these three clusters, one cluster that has the best frequency and monetary values is chosen so that it is considered the worthiest group to be given a discount in order to maintain its loyalty.

Keywords — Customer Grouping, Data Mining, K-Means Algorithm, CRISP-DM Method, RFM Model, Discounts

Received November 2021 / **Revised** February 2022 / **Accepted** June 2022

INTRODUCTION

Customer loyalty is something that the company needs to pay attention to in the current condition. Customer loyalty is also one of the most researched concepts in marketing (Zephaniah et al., 2020). Retaining existing customers provides greater benefits than finding new customers (Noyan & Gölba, 2014). The author sees an opportunity for the company to overcome and prepare for the competition that will exist later, namely by determining the target discount to customers to increase customer loyalty. Discount is an independent variable that has a major influence on customer loyalty (Suryani, 2013). However, giving a discount in this company is only based on a decision from the superior, without any clear procedure. In addition, there is no customer mapping which makes it difficult for the company to determine which customers deserve to be loyal to it through the provision of discounts. PT XYZ is a private company that works in the distribution sector of medical and laboratory equipment. PT XYZ already has many consumers in the Palembang area, starting from hospitals and also pharmacies, it will be very important for the company to maintain loyal customers. In previous research, research have conducted related to the Decision Support System for the distribution of discounts to resellers using *Simple Multi-Attribute Rating Technique Exploiting Ranks* (SMARTER) method (Haris et al., 2017).

DOI: 10.24014/coreit.v8i1.14695

Fredich Reincheld in his book Loyalty Rules (2007) says rewards (in this case in the form of discounts) cannot be given arbitrarily. Rewards are given to customers who carry out certain transactions that bring great results to the company (Sari, n.d.). Discount is a price reduction given by the seller to the buyer as a form of appreciation for the special activities of a profitable transaction for the seller (Fandy, 1997). Therefore, the authors propose appropriate discount strategies that can increase customer loyalty. In this case, it is necessary to do customer segmentation based on the profile of each customer, after that analyse the results of the segmentation so that loyal customers are known or not. The model used is the RFM (Recency, Frequency, Monetary) model, which is to group customers according to the customer's last transaction time interval, purchase frequency and the amount of value issued. (Savitri et al., 2018). The technique used in this research is data mining using the clustering method for grouping customers. Clustering is a data mining technique that divides data into several groups/clusters where data in a group has the same characteristics or forms compared to data in other groups. (Zheng, 2013). The clustering method used is k-means clustering. This method is used because the k-means clustering method is an interactive method that is easy to understand or interpret, implement, and has an active nature on scattered data.(Hughes, 1994).

The combination of the k-means method and the RFM model can support the process of grouping each type of customer and knowing the level of loyalty of existing customers (Adiana et al., 2018). The smaller the time interval of the last transaction with the current time or R value, the greater the probability that the customer will repurchase with the company, as well as the F value, the greater the F value, the greater the probability that the customer will repurchase with the company. the company and the greater the M value, the customer will respond to the company's products and services (Cheng & Chen, 2009). The results of this study will be obtained from several clusters of data grouping. The largest cluster with the highest RFM value is a loyal customer group and deserves a discount to maintain their loyalty.

LITERATURE REVIEW

A. Discount

Discount is a price reduction given by the seller to the buyer as a form of appreciation for the special activities of a profitable transaction for the seller (Fandy, 2008). Discounts are also the provision of prices that are lower than the price that should be paid based on a number of things including faster payment times, the level and amount of purchases and purchases in certain seasons. (Gitosudarmo, 2000). The company will provide discounted prices or allowances to customers who make early payments, bulk purchases, and off-season purchases. (Kotler et al., 2007).

B. Customer loyalty

Customer loyalty is defined as a customer who always shows a positive attitude towards a product, has loyalty to a particular product, and intends to continue to buy it in future transactions. (Sari, n.d.). Previous research that discusses loyalty is research conducted by Noyan et al, regarding the antecedents of customer loyalty and also research by Hesti Kartika Sari regarding the Effectiveness of Loyalty Programs in customer relationship management on customer satisfaction and loyalty.

C. Data Mining

Data mining is a technique used to extract understanding discoveries in databases. Data mining is a process of looking for patterns and relationships that are not visible in some large data that aims to perform grouping, estimation, estimation, association rules, clustering, description and visualization. (Han & Kamber, 2006).

The k-means algorithm is an algorithm for solving clustering problems (Zhao et al., 2010). The k-means method is a non-hierarchical method. This method is a limitation technique (partition) that sorts objects in the form of one or more groups (clusters). (Zheng, 2013).

The clustering process using the k-means algorithm is shown in the Figure 1.

DOI: 10.24014/coreit.v8i1.14695

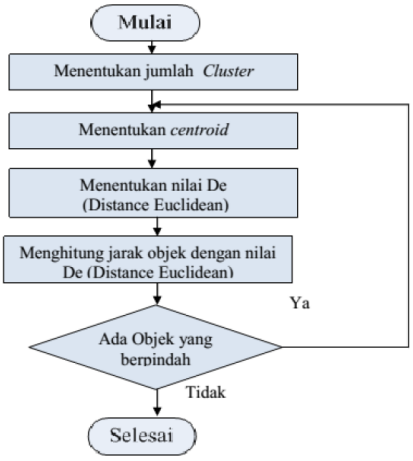

Figure 1 K-Means Flowchart (Santoso & Budi, 2007)

D. RFM

RFM (Recency, Frequency, Monetary) is a model of customer grouping based on the distance between the customer's last purchase, frequency of purchase, and the amount of value issued as company royalties. (Aggelis, n.d.). RFM is a method that is often used to determine whether a customer is valuable (valuable customer) in the following ways: (Birant, 2011):

1. *Recency*

Recency is the last date the customer made a purchase transaction.

2. *Frequency*

Frequency shows how active customers are in making purchases, so the more often customers make purchases, the greater the profits obtained by the company, but this is also influenced by the monetary value.

3. *Monetary*

Monetary is the value or value of the nominal amount in each customer transaction. Customers who make transactions with high monetary values tend to provide big profits for a company.

E. CRISP DM

In CRISP-DM, a data mining process has a cycle consisting of six phases. The following is a picture of the data mining cycle,

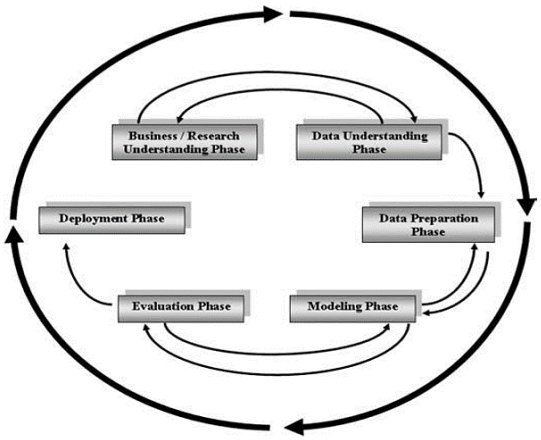

Figure 2 CRISP-DM Cycle (Larose, 2006)

Note that the phase sequence is adaptive. That is, the next phase, often depends on the results associated with the previous phase. The most significant dependencies between phases are indicated by arrows. For example, suppose we are in a modelling stage that depends on the actions and characteristics of the model, we may have to return to the data preparation stage for further refinement before moving on to the model evaluation stage. The six phases are as follows (Larose, 2006):

DOI: 10.24014/coreit.v8i1.14695

a. *Business Understanding Phase*
b. *Data Understanding Phase*
c. *Data Preparation Phase*
d. *Modelling Phase*
e. *Evaluation Phase*
f. *Deployment Phase*

METHODS

A. Research Material

1) Research Object

The research that will be conducted by the author discusses customer grouping using the k-means algorithm based on the RFM model. The location of the research to be conducted is at PT XYZ.

2) Data collection

The data used in this study comes from the sales transaction data section at PT XYZ. In collecting the data, the author conducted several interviews with the accounting department regarding sales transactions and also the procedures for giving discounts. Through this interview, the author obtained 2734 transaction data from 210 PT XYZ customers in October 2019 to March 2020 consisting of the transaction date, customer name, name of the item purchased, purchase price and purchase amount. In addition to interviews, the author collects several studies or literature related to the field being researched and analysed, in the form of journals, books, websites or other sources to support the author's research.

F. Research Methods

1) Business Understanding Phase

In this study, the author will solve the problems that exist in the sales transaction process at PT XYZ. Giving discounts to customers when making sales transactions is only based on a decision from the superior, without any clear procedure. This causes an error in the targeting of the discount. In addition, there is no customer mapping which makes it difficult for the company to determine which customers deserve to be loyal to it through the provision of discounts.

The implementation of data mining in this research is related to sales transaction data. The stored transaction data is used to group PT XYZ customers based on their level of loyalty. From the results of the grouping, it will be obtained potential customers who deserve to be given a discount. The purpose of implementing data mining in this study is to obtain additional knowledge about potential customers on existing data which will later be assisted by using the RFM model.

2) Data Understanding Phase

This study uses sales transaction data from October 2019 to March 2020 in spreadsheet format. Data analysis will be carried out regarding sales transaction data along with influencing factors such as recency (when buying and selling transactions occur between sellers and customers), frequency (how active the customer is in making transactions) and monetary (the nominal amount in each buying and selling transaction carried out). This factor will later become a reference in obtaining information/knowledge

3) Data Preparation Phase

At this stage the stored data will be prepared so as to facilitate the mining process. There are several processes carried out in preparing the data including selecting the variables to be analysed, cleaning the data, preparing the initial data so that it is ready for data transformation. The grouping of customers in this study is based on the RFM model (recency, frequency, monetary), then the data selection from the RFM model is the last time the customer made a transaction using the analytical method, the number of frequencies (how often) transactions made by the customer during the study period, and the number of nominal transactions for each customer during the study period.

DOI: 10.24014/coreit.v8i1.14695

Table 1 Customer RFM Value

No	R	F	M
1	4	9	8	638.144.317
2	3	2	0	1.385.454
3	2	9	9	69.922.330
4	3	1	9	22.221.087
5	3	9	1	5.168.182

4) Modelling Phase

The method proposed in this research modelling is the K-Means algorithm. The input data for the clustering process is the normalized customer transaction RFM data. The clustering process uses the k-means algorithm using the Elbow method to find the optimal k value. The modelling flow is shown in Figure 3.

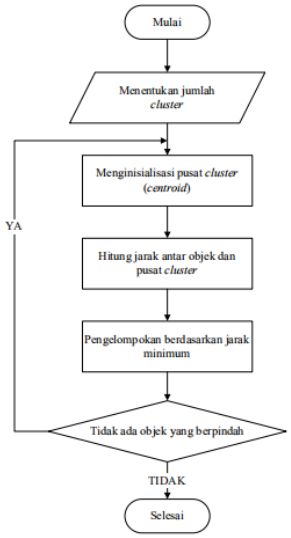

Figure 3 *Flowchart* of K-Means Algorithm

5) Evaluation Phase

The implementation phase or also known as the Deployment Phase is the stage where reports on the results of data mining activities will be made. Reports are generated in the form of visualizations to assist the author in making decisions on the results of the analysis of sales transaction data at PT XYZ. The visualization that will be used is a scatter plot visualization that shows the distribution of each attribute of recency, frequency, and monetary.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (1)$$

The value of the silhouette ranges from -1 (S) 1, where the clustering results are good if the silhouette value is positive (0-1). This indicates that the data is in the right cluster.

6) Deployment Phase

The implementation phase or also known as the Deployment Phase is the stage where reports on the results of data mining activities will be made. Reports are generated in the form of visualizations to assist the author in making decisions on the results of the analysis of sales transaction data at PT XYZ. The visualization that will be used is a scatter plot visualization that shows the distribution of each attribute of recency, frequency, and monetary.

RESULT AND DISCUSSION

A. Result

The data used in this study is sales transaction data at PT XYZ in the period October 2019 to March 2020 in excel format. The following is a table of data snippets of Sales Transactions at PT XYZ Palembang.

DOI: 10.24014/coreit.v8i1.14695

Table 2 Sales transaction data

Last Transaction	F	Amount
21/02/2020	1	7252000
10/03/2020	9	638144322
20/11/2019	1	409091
07/01/2019	1	136364
16/10/2019	1	209091
27/03/2020	10	1385458
20/03/2020	1	5168184
14/02/2020	1	227273
27/03/2020	1	818182
27/03/2020	1	327273
27/02/2020	1	370000
26/03/2020	1	163636

After the data is prepared, the next step is data selection based on the RFM features, namely Recency, Frequency and Monetary.

Table 3 RFM

No	R	F	M
1	49	8	638.144.317
2	32	10	1.385.454
3	29	19	69.922.330
4	31	9	22.221.087
5	39	1	5.168.182

The next step is to select the data, divide it into three parts, namely recency, frequency, and monetary, then perform data transformation. At this stage, the normalization process is carried out. The normalization stage is carried out so that the data ratio does not have too much difference between one another, which is between 0 to 1. The following is a snippet of RFM data normalized using the min max method.

Table 4 Normalization

No	R	F	M
1	0,20	0,25	0,90
2	0,12	0,26	0,10
3	0,10	0,43	0,19
4	0,11	0,25	0,13
5	0,15	0,10	0,11

After all the RFM data are normalized, the next step is the modelling stage. The method used in the modelling of this research is the K-Means algorithm. The input data for the clustering process is the normalized customer transaction RFM data. In this K-Means method, the author needs to determine the number of clusters that will be used later in the clustering process. This study uses the Elbow method to find the optimal k value. The following is a graphic description of the calculation results to determine the optimal number of clusters from the existing RFM data using the elbow method.

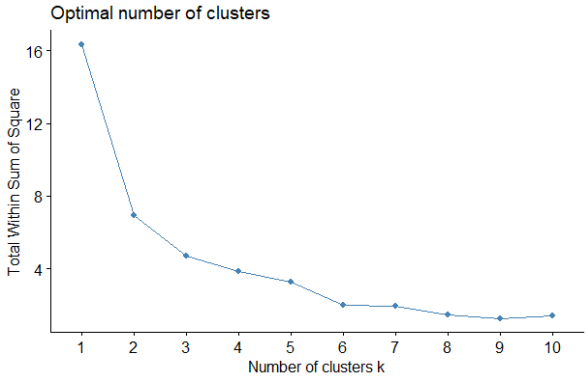

Figure 4 Elbow Method

DOI: 10.24014/coreit.v8i1.14695

Through the calculation of the elbow method using the help of RStudio, the results of the graph of the elbow method show that the elbow point between the lines is at point 3. The line after point 3 there is no significant drop in point. So it can be concluded that the optimal number of clusters based on the elbow method is 3 clusters. In addition to using the elbow method, the author also determines the right number of clusters by visually seeing the spread of RFM data through a scatter plot.

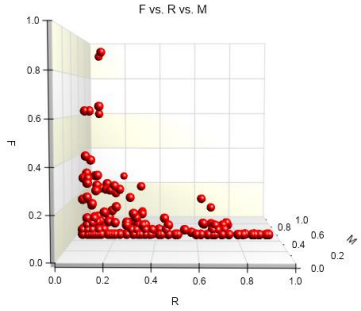

Figure 5 3D Scatter Plot of RFM

Figure 5 shows that the plot generated from the RFM data can be divided into 3 parts, which means that it supports the results of the elbow method, namely 3 clusters.

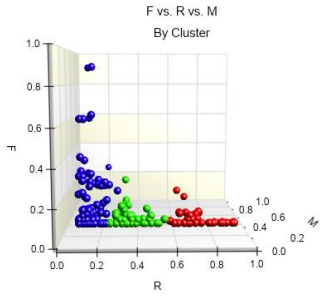

Figure 6 *Scatter Plot* 3D with 3 *clusters*

After obtaining the number of clusters through the calculation of the elbow method and also by looking at the distribution of the data visually, the next stage is the evaluation or testing stage. This phase has the aim of evaluating whether the cluster results are optimal or not. In this study using the Global Silhouette evaluation process. The silhouette method is one of the methods used to evaluate the quality of the clusters resulting from the clustering process.

The silhouette test was carried out on transaction data that had been collected, from October 2019 to March 2020. The test was carried out by entering 3 clusters and calculating them using the silhouette equation. The result of the silhouette calculation is 0.589. The silhouette value with three clusters produces a positive value and is the largest value from the number of other clusters. This shows that the average distance between objects in the same cluster is smaller than the average distance between objects and objects in other clusters, which means that each customer is in the right cluster. Visualization of the accuracy test with a silhouette can be seen in Figure 7. The accuracy test or evaluation is carried out using the RStudio tools. From Figure 7 it can be seen that the division of the cluster into 3 has a good silhouette value.

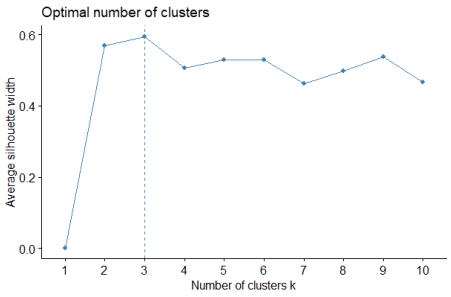

Figure 7 *Global Silhouette*

DOI: 10.24014/coreit.v8i1.14695

The last phase in the CRISP DM method is the implementation phase or deployment phase. After getting the optimal number of clusters, the next step is to do the grouping process using K-Means Clustering. In this study, the author uses the help of the RStudio application in the clustering process with the K-Means method.

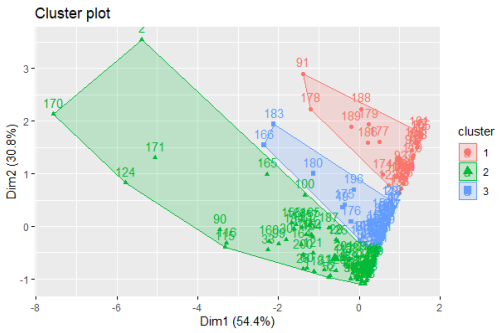

Figure 8 Cluster Plot 2D

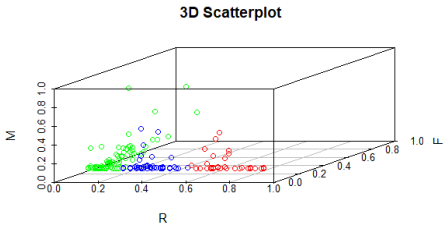

Figure 9 Cluster Plot 3D

Figure 8 is the result of customer transaction RFM data clustering at PT XYZ Palembang. The red area is a customer group that is in the first cluster, the blue area is a customer group that is in cluster 2 and the green area is a customer group that is in cluster 3. After grouping, the author also gets the centroid value of each cluster. Table 5 shows the centroid value or the average RFM of each cluster.

Table 5 Average RFM

Cluster	Count	R	F	M
1	47	0.72	0.11	0.13
2	104	0.15	0.19	0.15
3	50	0.38	0.12	0.13

From 201 customer data at PT XYZ, there are 47 customers in the first cluster, 104 customers in the second cluster and the remaining 50 customers in the third cluster.

B. Discussion

Based on the average values in Table 5 the authors make a graph that can make it easier to see the magnitude of the value of each RFM.

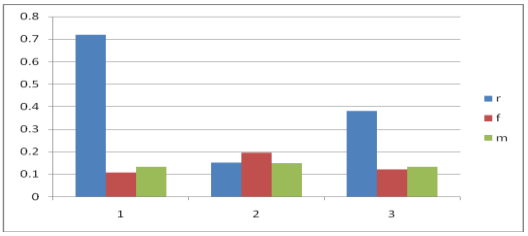

Figure 10 Clustering Graph

Through the graph in Figure 10, the author performs an analysis based on the distribution of RFM data in each cluster.

DOI: 10.24014/coreit.v8i1.14695

1. *Cluster* #1

In cluster 1 there are 47 customer data with the following characteristics:

- The value of R (recency) for cluster 1 is the largest of the other two clusters, meaning that customers in this cluster have not made transactions for a long time.
- The lowest F (frequency) value, it means that customers in this cluster very rarely make purchase transactions within a certain period.
- The value of M (monetary) has a moderate value, which means that the customers in this cluster provide sufficient financial value for the company.

2. *Cluster* #2

In cluster 2 there are 104 customers with the following characteristics:

- The R (recency) value in this group is low, which means it has a good value and the customers in this cluster made the last transaction recently, this gives the potential for these customers to make transactions again (Soeni, 2012).
- The F (frequency) value is the highest from the other 2 groups, meaning that customers in this cluster often make purchase transactions at a certain time.
- The value of M (monetary) in this cluster also has the highest value from other clusters, meaning that customers in this cluster provide high financial for the company.

3. *Cluster* #3

In cluster 3 there are 50 customers with the following characteristics:

- The R (recency) value in this cluster is on average, meaning that the last transaction was carried out recently, this can also provide potential customers to repurchase (Soeni, 2012).
- The value of F (frequency) is at the average value, which means that customers in this cluster make transactions quite often during a certain period of time.
- The value of M (monetary) is at the smallest value, meaning that the level of spending in this cluster is quite profitable for the company.

Recency is the time interval between the last purchase and the current time, a lower value corresponds to a higher probability that the customer will make a repeat purchase. Frequency is the number of transactions that have been carried out by customers within a certain period of time and monetary is the amount of money spent in a specified time period (Wang, 2009).

Based on the above analysis, cluster 1 contains a group of customers who have low loyalty, cluster 2 is a group of customers with good loyalty, while cluster 3 is a group of customers who have moderate loyalty. Fredich Reincheld in his book Loyalty Rules (2007) says awards or rewards (in this case in the form of discounts) cannot be given arbitrarily. Rewards are given to customers who carry out certain transactions that bring great results to the company (Sari, n.d.). Therefore, the customer group that deserves a discount to maintain their loyalty is the customer group that is in the second cluster.

After getting the best cluster, the author performs a cluster test scenario that shows whether the number of customers in a cluster will remain the same if the total number of customers in PT XYZ is different. In this study, the authors tested using 201 customers, 100 customers and 50 customers of PT XYZ Palembang.

Table 6 Cluster Testing Scenario

#Total	#Cluster 1	#Cluster 2	#Cluster 3
201	47	104	50
100	21	63	16
50	34	5	11

Based on Table 6, the number of customers in each cluster will be different according to the total number of customers. This is because each customer will enter a cluster containing other customers with the closest distance. Therefore, the number of customers in each cluster which has 201 total customers will be different from the number of customers in each cluster which has 100 total customers, as well as 50 total customers at PT XYZ Palembang.

CONCLUSION

DOI: 10.24014/coreit.v8i1.14695

The conclusion of this research is as follows:

1) This study can determine the grouping of customers at PT XYZ which is divided into 3 groups or clusters using the k-means clustering method based on the RFM model (recency, frequency and monetary) which produces 3 categories, namely good loyalty, moderate loyalty and low loyalty.
2) Discounts in this study are given to customers in the second cluster who have the best frequency and monetary values so that they are considered the worthiest group to be given discounts in order to maintain their loyalty.

REFERENCES

Adiana, B. E., Soesanti, I., Permanasari, A. E., No, J. G., No, J. G., & No, J. G. (2018). Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering. 2, 23–32. https://doi.org/10.21460/jutei.2017.21.76

Aggelis, V. (n.d.). Customer Clustering using RFM analysis. 1–5.

Birant, D. (2011). Data Mining Using RFM Analysis. Knowledge-Oriented Applications in Data Mining, 92–108.

Cheng, C., & Chen, Y. (2009). Expert Systems with Applications Classifying the segmentation of customer value via RFM model and RS theory. Expert Systems With Applications, 36(3), 4176–4184. https://doi.org/10.1016/j.eswa.2008.04.003

Fandy, T. (1997). Strategi Pemasaran (Edisi 2). Penerbit Andi.

Fandy, T. (2008). Strategi Pemasaran (Edisi Keti). ANDI Yogyakarta.

Gitosudarmo. (2000). Manajemen Pemasaran (Edisi Pert). BPFE.

Han, J., & Kamber, M. (2006). Data Mining Concepts and Techniques Second Edition. Morgan Kauffman.

Haris, A., Satria, B., & Ukkas, M. I. (2017). Penerapan Sistem Pendukung Keputusan Pemberian Diskon Pada Reseller Dengan Metode Simple Multi-Attribute Rating Technique Exploiting Ranks (SMARTER). 7(2), 31–37.

Hughes, A. M. (1994). Strategic Database Marketing. Probus Publishing.

Kotler, Philip, Lane, K., & Keller. (2007). Manajemen Pemasaran (Edisi Kedu). PT INDEKS.

Larose, D. T. (2006). Data Mining Methods and Models. Wiley-Interscience.

Noyan, F., & Gölba, G. (2014). Fatma Noyan. 109(2002), 1220–1224. https://doi.org/10.1016/j.sbspro.2013.12.615

Sari, H. K. (n.d.). dalam Customer Relationship Management terhadap Kepuasan dan Loyalitas Pelanggan. 177–206.

Savitri, A. D., Bachtiar, F. A., & Setiawan, N. Y. (2018). Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus : Belle Crown Malang). 2(9), 2957–2966.

Soeni, R. A. (2012). Customer Segmentation based on Modified RFM Model in the insurance Industry. International Conference on Machine Learning and Computing.

Suryani, A. (2013). Generalized Structured Componen Analysis (GSCA) Tentang Atmosfir Mall, Kualitas Pelayanan, Diskon dan Loyalitas Pelanggan. Doctor Thesis, Universitas Brawijaya.

Wang, C. (2009). Robust Segmentation for the Service Industry Using Karnel Induced Fuzzy Clustering Techniques. IEEE IEEM, 2197–2201.

Zephaniah, C. O., Ogba, I. E., & Izogo, E. E. (2020). Examining the effect of customers' perception of bank marketing communication on customer loyalty. Scientific African, 8, e00383. https://doi.org/10.1016/j.sciaf.2020.e00383

Zhao, J., Zhang, W., & Liu, Y. (2010). Improved K-Means cluster algorithm in telecommunications enterprises customer segmentation. Proc. 2010 IEEE Int. Conf. Inf. Theory Inf. Secur. ICITIS 2010, 167–169.

Zheng, D. (2013). Application of Silence Customer Segmentation in Securities Industry Based on Fuzzy Cluster Algorithm. 13, 4337–4347. https://doi.org/10.12733/jics20102432

DOI: 10.24014/coreit.v8i1.14695