Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

Master Stockist Customer Segmentation Using RFM Model and Self-Organizing Maps Algorithm

Ni Kadek Ayu Nirwana[1)] , Ni Putu Wahyuni Dewi[2)] , I Made Dwi Putra Asana[3)*] , Ni Wayan Jeri Kusuma Dewi[4)] , Gusti Ayu Shinta Dwi Astari[5)]
[1,2,3,4,5)]Department of Informatics, Institute of Business and Technology, Indonesia
[1)]ayu.nirwana@instiki.ac.id,[2)] wahyunidewi3108@gmail.com,[3)] dwiputraasana@instiki.ac.id,
[4)]wayan.kusumadewi@instiki.ac.id, [5)]shinta.astari@instiki.ac.id

Submitted: Sep 15, 2024 | **Accepted**: Oct 6, 2024 | **Published**: Oct 17, 2024

Abstract: Master Stockist PT SNS 21 Bali struggles to identify member performance based on purchasing behavior because the applicable system only records transactions and stock of goods without providing insight into customers. Customer segmentation can be carried out to identify and understand differences in customer purchasing behavior. Therefore, this study aims to determine customer segmentation using the RFM (Recency, Frequency, Monetary) model and the Self-Organizing Maps (SOM) algorithm. Segmentation development uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach. The RFM model numerically represents customer behavior through three variables, while the Self-Organizing Maps algorithm groups customers into segments with similar characteristics. In this research, the best SOM parameters are 750 iterations, learning rate 0.5, radius 0.5, and grid size 1x3, resulting in 3 clusters with a Silhouette Score of 0.647608 and a Davies-Bouldin Index of 0.536503. Cluster 1 consists of 226 new customers with low RFM values who need encouragement to be more active. Cluster 2, comprising seven members, has low recency, high frequency, and high monetary values, representing loyal customers who need to be retained. Cluster 3 consists of 239 inactive customers with high recency, low frequency, and low monetary values, requiring a reactivation strategy.

Keywords: Customer Segmentation; RFM; Self-Organizing Maps;Master Stockist; Data Mining

INTRODUCTION

In today's competitive business era, a good understanding of customers is one of the key elements that determine the sustainability of a business. As the main stakeholder, customers are not only a source of revenue but also a measure of the success of a business. Sustainability and success are closely related to customers who play an important role in determining business strategy. The right business strategy tends to be able to maintain the business in the long term because the main purpose of the strategy is to understand, serve, and meet the needs of customers of a business (Wicaksana et al., 2022). This has led to changes, namely the strategy of a business or business is now not only focused on products (*product/service oriented*) but also considers *customers* (*customer oriented*).

PT Sukses Nusantara Sakti 21 abbreviated as PT SNS 21 headquartered in Purwokerto has a role as an authorized distributor and sole agent of PT Laba Asia Food as a *Multi-level Marketing* product manufacturer. *Master Stockist* PT SNS 21 Bali is a business that plays an important role in product distribution to customers or partners and all consumers in the Bali and Nusa Tenggara regions. As an authorized distributor, *Master Stockist* PT. SNS 21 Bali is certainly obliged to serve the needs of its customers with a total of 60,414 id partners registered as members within the region. Not only includes product distribution but also related to building sustainable relationships with partners and consumers to meet the evolving needs of partners and consumers or customers.

Building good relationships with customers can be done in many ways such as good service. Services related to product availability at *Master Stockist* have been well implemented as seen from the system that records integrated sales with the availability of stock items, but not with customer management. Based on interviews with the owner of *Master Stockist* PT SNS 21 Bali, the applicable system is only used in transaction activities that are connected to stock items while information about customers cannot be accessed on the system. So it can be said that the data available on the system has not provided insight into customers. With the large number of members in Bali without a system that provides customer insight, it becomes an obstacle faced by *Master Stockist*, making

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

it difficult to identify member performance based on their purchasing behavior. Based on the current data, none of the 60,414 registered partners at Master Stockist PT SNS 21 Bali have undergone customer segmentation. This is due to the limitations of the existing system, which is solely designed to manage transactions and stock availability, without offering insights into customer behavior or facilitating segmentation. Consequently, the percentage of potential customers that have been clustered is effectively zero percent, as no segmentation has been conducted using the available data.

Customers have various behaviors and characteristics. For example, customers who make purchases in large quantities at once but only once a few months and there are also those who regularly make purchases but in small quantities. Customers are called potential if they have loyalty which can be seen from the tendency to continue purchasing from the same company even though other alternatives are given. (Wicaksana et al., 2022). Attention cannot only be paid to potential customers. Companies need to identify customers who show potential to become pillars of business sustainability. Therefore, a different approach is needed to find out the potential and characteristics of customers. To identify and understand the differences in customer purchasing behavior can be done by customer segmentation.

Customer segmentation is carried out to divide customers into groups that have similar characteristics (Bellotti et al., 2021). Customer segmentation can be done by doing data mining. Data mining is a term used to describe knowledge discovery in databases. (Tholib, 2020). Data mining is done to find *valuable customers* from sales history data that shows past customer consumption patterns which are later used to develop appropriate strategies. (Wei et al., 2020) (Juhari & Juarna, 2022). Customer identification can be done by implementing the RFM (*Recency, Frequency, Monetary*) model as a method to translate customer behavior into numbers which are divided into three attributes, namely *recency* obtained from the last purchase time, *frequency* obtained from the number of transactions in a certain period, and *monetary* obtained from the value of customer transactions in a certain period.

To get a good segmentation, a further approach is needed by grouping customers into segments through the implementation of algorithms in *clustering.* In this research, clustering is done using the *Self-Organizing Maps* algorithm. The use of this algorithm is done because based on a comparison between two algorithms, namely Fuzzy C-Means and SOM, the SOM algorithm provides a more consistent centroid movement in changing the centroid every iteration compared to FCM (Fawaz et al., 2022). Then in the K-Means algorithm which is one of the popular algorithms used in segmentation cases, this algorithm requires a value of k as an input parameter, namely the number of clusters to be formed and the *K-means* algorithm can be affected by *outlier* data. For small datasets, there will be difficulties in accurately clustering the data. If there is overlapping data, then *K-means* will not be able to separate the *clusters* clearly (Wisnuwardhana et al., 2020). In other data clustering cases, SOM has the best clustering quality with the best test scores compared to clustering quality with *K-means* and *Agglomerative Clustering.* (Muqsit & Swanjaya, 2021). Therefore, the author chooses to use the SOM algorithm because it is proven to produce stable *clusters* seen from the value of the point with the group (centroid) which does not change every time the test and has quite good accuracy (Wijaya et al., 2023).

LITERATURE REVIEW

Related Research

Research on clustering analysis using clustering algorithms has been widely conducted in various contexts and in this study, five previous studies with similar methods were used as references. The first study entitled "Using A Combination of RFM Model and Cluster Analysis to Analyze Customers' Values of A Veterinary Hospital" (Wei et al., 2020) aims to identify customers with different behaviors and then develop adequate marketing strategies to maintain good relations with existing customers and attract new customers for animal hospitals. This study used two-stage clustering, a combination of Self Organizing Maps and K-Means methods, and the RFM model was used to analyze customer value from transaction data focusing on dogs in veterinary hospitals in Taichung City, Taiwan in 2014. The results showed that 4,472 customers were classified into twelve clusters, and seven of the twelve clusters were found to be the best or loyal customers. However, the other five clusters were uncertain customers. Among the five clusters, three clusters are lost customers and two clusters with relatively higher than average recency values can be considered as new customers.

Furthermore, research entitled "Customer Segmentation Based on RFM Analysis Using the K-Means Algorithm as a Basis for Marketing Strategy (Case Study of PT Coversuper Indonesia Global)" (Widiyanto & Witanti, 2021) aims to produce a customer segmentation application using a combination of the K-Means method and the RFM model that can assist companies in finding the characteristics of each customer so that companies are able to prioritize resources and energy for certain customers. The results of this study are 4 customer segments from a total of 736 customers, Consumers 226, Ordinary 186, Big Consumers 299 and Top Class 25. Based on the accuracy obtained in calculations using the customer segmentation system from 29 tests, there are 29 customers (100%) with characteristics produced by the system in accordance with user knowledge, 0 customers (0%) are not in accordance with user knowledge. With this accuracy value, it can be categorized as good. These customer characteristics will help PT Coversuper Indonesia Global to make decisions in prioritizing its energy and resources to certain customers (potential).

* name of corresponding author

In the same year, research entitled "Analysis of Customer Satisfaction Survey Data Grouping Models Using the Self Organizing Maps Method" (Muqsit & Swanjaya, 2021) which aims to model the clustering of customer satisfaction survey data using the SOM method to gain a better understanding of customer preferences and patterns. The results obtained from this study are the relationship between many groups and the quality of clustering, the relationship between the max_epoch value and the quality of clustering and the time to train the network, as well as a comparison of the quality of clustering of the SOM method, with K-Means Clustering and Agglomerative Clustering. The SOM method has the best clustering quality with its Silhouette Coefficient value of 0.1849.

In the following year, research entitled "Implementation of RFM Analysis Model for Customer Segmentation Using The K-Means Algorithm Case Study XYZ Online Bookstore" (Juhari & Juarna, 2022) The research was conducted which aims to improve marketing strategies by implementing Customer Relationship Management and customer segmentation using the RFM model (Recency, Frequency, Monetary) and the K-means clustering algorithm. The results of the study stated that cluster analysis based on customer value using RFM and Customer Value Matrix methods shows that based on the RFM method, it produces 3 types of customer characteristics, namely loyal customers, new customers, and lost customers. While based on the customer value matrix method produces 2 types of customer characteristics, namely the best customers and uncertain customers.

The last research with the title "Implementation of the Self Organizing Map Algorithm to Identify Grouping Patterns of Family Welfare Levels in Siak District" (Aziz & Mustakim, 2022) was conducted with the aim of clustering family welfare data based on three aspects namely health, education, and economy using the SOM algorithm. The research conducted 36 experiments using 2016 welfare data totaling 22,047 data. The results of this study with validation using DBI show cluster 3, learning rate 0.20 and 500 iterations on the 10th trial with a DBI value of 0.9398 is the optimal cluster with 2 clusters and 1 outlier. The results in cluster 0 from the health aspect as much as 25.22% have no disability, from the education aspect as much as 20.92% do not go to school anymore, from the economic aspect as much as 16.59% do not have a main job. Then in cluster 2 from the health aspect as many as 72.89% do not have disabilities, from the education aspect as many as 28.49% have elementary / equivalent diplomas and economic aspects as many as 33.51% of the main jobs are gardening. The results of these three aspects are used as recommendations for the Social Service of Siak Regency to be more optimal in classifying family welfare in providing assistance.

Customer Segmentation

Customer segmentation is a process of grouping customers into homogeneous groups based on characteristics that can be used as a way to determine marketing strategies. (ASTUTI, 2019). Customer segmentation is an important element in the context of *relationship marketing* to improve relationships with customers for the better related to customer needs (Bellotti et al., 2021). Segmentation aims to customize products, services, and marketing messages for each segment. In addition, segmentation also allows companies to understand the behavior of customers, customer preferences, and gain knowledge about various customer groups. (Juhari & Juarna, 2022).

RFM

RFM (*Recency, Frequency, Monetary)* is a popular model used to understand customer behavior. The RFM model is a method that is often used to segment customers (Juhari & Juarna, 2022). *Recency*, is information about the time of the customer's most recent purchase which shows how recently the customer made a purchase. *Frequency*, shows how many times the customer has made a purchase or the number of transactions the customer has made within a certain period of time. *Monetary*, is the amount of money spent by customers cumulatively. Related to *recency*, marketers believe that customers who have made a recent purchase are more likely to make another purchase compared to customers who made a purchase in the past. While *frequency*, this metric is used with the assumption that customers who have made more purchases are more likely to make future purchases compared to customers with fewer transactions (Gustriansyah et al., 2020).

Clustering

Clustering is the process of grouping an unlabeled pattern that is *unsupervised learning* into a group that has certain characteristics. *Clustering* is a very important process when it comes to problems such as pattern analysis, decision making, machine learning, data mining and so on. This technique is a popular technique that is widely used in data mining (Paembonan & Abduh, 2021). Data mining is the process of searching and analyzing large amounts of raw data to identify patterns and extract useful information (Twin, 2023). Data mining can also be defined as the process of extracting useful information hidden from data. *Cluster* analysis has two basic concepts, namely similarity and distance measurements. The concept of similarity of characteristics is the closeness between one object and another. While the concept of distance is a measure of separation between two objects. Cluster analysis has several characteristics such as having high similarity between members with each other in one group, having a high difference between groups with one another. The concept most often used to measure the similarity of data with each other is the euclidean distance measure. Euclidean distance is a way to measure the distance between objects and cluster centers (Mardiyah, 2022).

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

Self-Organizing Maps

The SOM (*Self-Organizing Maps)* algorithm or often called the Kohonen artificial neural network is an artificial neural network method introduced by Professor Teuvo Kohonen in 1981. SOM networks are used for *clustering* data based on the characteristics or features of the data included in *unsupervised* learning. (Aziz & Mustakim, 2022). The SOM network consists of two layers, namely the *input layer* and the *output layer*. Each neuron in the *input layer* is connected to each *output layer*. Each neuron in the *output layer* represents the class of the given input. During the process, the *cluster* that has the weight vector with the closest distance to the *input layer* will be selected as the winner or called the best matching unit (MUSDARI et al., 2019). The stages in the SOM algorithm are basically initialization, random sampling, calculation of the closest neuron, weight update, then repeating the second to fourth until a certain iteration threshold is reached. (Purbasari et al., 2020). Before applying SOM, it is necessary to normalize the data first using *min-max scaling* which aims to make the data have a smaller range with the same weight between 0-1 with equation (1) (Mardiyah, 2022).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

Where x is the data, x_{min} is the minimum data, and x_{max} is the maximum data. After normalizing, the input vector is initialized, i.e. the set of m *field* values for the nth record becomes an input vector $x_n = x_{n1}, x_{n2}, x_{n3}, ..., x_{nm}$ and the set of m weights for a particular output node j becomes a w*eight* vector $w_j = w_{1j}, w_{2j}, w_{3j}, ..., w_{mj}$ (Henderi, 2021). The SOM algorithm step begins with initializing the initial weights randomly as W_{ij}, determining the *learning rate* (α), and the radius value R. Then calculate the distance between the weights w_{ij} and the input vector x_i using equation (2) *euclidean distance* calculation.

$$D = \sqrt{\sum_{i=1}^{m} i = \left(w_{ij} - x_i\right)^2} \qquad (2)$$

Where W_{ij} is the connecting weight between the input neuron and the output neuron and x_i is the neuron in the i-th input layer and m is the number of variables. The smallest value of D is called the winner. Then the weights W_{ij} are updated using equation (3).

$$W_{ij}(baru) = W_{ij}(lama) + \alpha[x_i - W_{ij}(lama)] \qquad (3)$$

Where W_{ij} (new) is the new W_{ij} weight, W_{ij} (old) is the initial weight, α is the *learning rate* value, and x_i is the i-th input vector. The *learning rate* value is $0 \leq \alpha \leq 1$ which can be updated using the following formula where t is the number of iterations. The learning rate and radius are updated if it has not reached the maximum iteration. The test state stops when the iteration value reaches the maximum and can be known by calculating the weight W_{ij} (new) with W_{ij} (old), when the weight W_{ij} is known not to change or the change is not much, it can be said that the test stops and has reached convergence.

Silhouette Score

Silhouette score or *silhouette coefficient* (SC) is one of the *cluster* evaluation methods that combines *cohesian* and *separation*. *Cohessian* aims to determine how close the relationship between data in one group is while *separation* aims to measure how far a *cluster* is from other *clusters*. SC has a range of values between -1 and 1. A group can be said to meet good criteria if its value is close to 1 and not good if it has a value close to -1. (Mardiyah, 2022) (Paembonan & Abduh, 2021). Equation (4) is used to calculate the *silhouette score* value.

$$S\,(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (4)$$

Where S(i) is the *silhouette coefficient of* the i-th *cluster*, b(i) is the average distance of the i-th cluster with other clusters, commonly called neighboring *clusters* or inter-group values, namely from different *clusters*, a(i) is the average distance value of all data in the i-th *cluster* or intra-group distance value, namely in the same *cluster*. (Wijaya et al., 2023).

The average value of all *silhouette scores* for all *clusters* can illustrate how *cohesive* and *separable* the *clusters* are. The interpretation of object *silhouette* scores can be seen in table 1. After getting the *Silhouette* value of each object in the cluster, we can determine the *Silhouette Width* for the *cluster*, namely by calculating the average *Silhouette* value of all different objects in the *cluster*. (Muqsit & Swanjaya, 2021). The value of the *silhouette width* can be seen in table 2.

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

Table 1. Silhouette Object Value

S(x)	Interpretation
Negative	It shows a high *overlapping* structure that x is close to other objects in cluster B instead of A, the previous cluster. Or it can be said that x should not be in cluster A, in other words, it is placed in the wrong *cluster*.
0	Indicates x is at or very close to the decision boundary between neighboring clusters. Or x is said to be similar for clusters A and B.
Positive	Indicates x belongs to cluster A or is far away from cluster B.

Table 2. Silhouette Width Value

Silhouette width	Interpretation
0.71– 1	*Strong cluster*
0.51– 0.7	*Reasonable cluster*
0.26– 0.5	*Weak* or *artificial cluster*
≤ 0.25	*No cluster found*

Davies-Bouldin Index

Davies Bouldin Index (DBI) is one of the cluster validations introduced by D.L. Davies and Donald W. Bouldin, therefore the naming of this method is a combination of the names between the two, namely Davies-Bouldin (Septiani et al., 2022). The *Davies-Bouldin* index is one of the cluster validation methods for quantitative evaluation of *clustering* results. This measurement aims to maximize the intercluster distance between one *cluster* and another (Aziz & Mustakim, 2022).. The smaller the DBI value, the better the *clustering* results.(Manalu & Gunadi, 2022). There are four stages in calculating DBI, the first is calculating the *Sum of Square within Cluster* (SSW) is the attachment of members of one cluster or how similar members one and two are and the smaller the better because the more similar, with equation (5).

$$SSW_i = \frac{1}{m_i} + \sum_{j=i}^{m_i} d(X_j, C_i) \qquad (5)$$

With x is the data in the cluster, d(x,c) is the distance of the data to the centroid, m_i is the number of data in the *i-th* cluster, x_j is the data in cluster *j*, and c_i is the centroid of the *i-th* cluster. Next is to calculate *Sum Of Square Between Clusters* (SBB), which is the distance between clusters large enough to separate into other groups, with formula (6).

$$SSB_{i,j} = d(C_i, C_j) \qquad (6)$$

With $d(C_i, C_j)$ which is the distance between one centroid and another. Next is the calculation of the ratio between *clusters* which serves to be able to find out how good the comparison value of one *cluster* is with other clusters, with equation (7). After obtaining the results of the previous stages, calculating DBI is done with equation (8).

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{ij}} \qquad (7)$$

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \qquad (8)$$

Where k is the existing clusters and max is the largest *inter-cluster* ratio. A smaller DBI value indicates that the clusters are better separated and have a smaller size, which means the clustering is more optimal. The smaller the DBI value, the better the clustering results (Septiani et al., 2022).

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

METHOD

This article applies the *Cross Industry Standard Process Model for Data Mining* (CRISP-DM) method in the customer segmentation process. CRISP-DM consists of several stages, namely *business understanding, data understanding, data preparation, evaluation, modeling,* and *deployment* (Studer et al., 2021). Figure 1 shows an illustration of the phases in developing customer segmentation with RFM and SOM methods.

Fig. 1 CRISP-DM Phases

Business Understanding

This stage is done before segmenting customers by understanding the goals and needs from a business point of view. This research refers to the data contained in the sales report as its main object. The sales report contains information about daily sales transactions, customers involved in the transactions, products sold, and the sales amount associated with each transaction. The data obtained from the report will be used to gain insight into customers. Then, the data will be utilized for customer segmentation purposes that can be used as a consideration in determining business strategies by the company.

Data Understanding

The data in this study were obtained from the sales report of the *Master Stockist* PT SNS 21 Bali which occurred within a span of 1 year from January 1, 2023 to December 31, 2023. The data taken is in pdf format which can be converted into xls which consists of 2102 transactions. In one transaction contains information in the form of transaction date, transaction number, member id and customer name, item code, item name, quantity, price, discount, and subtotal of transactions made by customers which can be seen in figure 2.

PT SUKSES NUSANTARA SAKTI 21
Jl. M. Yamin No. 05 Karangklesem, Purwokerto Selatan, Purwokerto - Jawa Tengah Telp. 0822 2020 4021
LAPORAN PENJUALAN
Area : MASTER STOCKIST BALI (sns0043555)
Periode : 01/01/2023 - 31/12/2023

No	Kode Barang	Nama Barang	Jumlah	Harga	Potongan	Sub Total
1	**Tanggal : 01/01/2023**	**No Transaksi : #10123207152**	**Member ID : MASTER STOCKIST (SNS0043555)**			
	0522019	Brightening Package	1	702.640	0	702.640
				Jumlah :	**0**	**702.640**
2	**Tanggal : 07/01/2023**	**No Transaksi : #20123207574**	**Member ID : DESAK NYOMAN SU (SNS0171193)**			
	0617001	Susu Kambing Etawa Proviterje	12	412.544	120.000	4.950.528
	0617001	Susu Kambing Etawa Proviterje	68	362.544	680.000	24.652.992
	0522017	Brightening Bar Soap Purple	2	140.960	8.000	281.920
				Jumlah :	**808.000**	**29.885.440**
3	**Tanggal : 07/01/2023**	**No Transaksi : #20123207567**	**Member ID : STOCKIS BATUBULA (SNS0203918)**			
	0522017	Brightening Bar Soap Purple	2	140.960	8.000	281.920
	0617001	Susu Kambing Etawa Proviterje	166	362.544	1.660.000	60.182.304
	0617001	Susu Kambing Etawa Proviterje	26	412.544	260.000	10.726.144
				Jumlah :	**1.928.000**	**71.190.368**
4	**Tanggal : 07/01/2023**	**No Transaksi : #10123207576**	**Member ID : MAYLANI BIMANTA (SNS0344243)**			
	0617001	Susu Kambing Etawa Proviterje	2	362.544	0	725.088
				Jumlah :	**0**	**725.088**

Fig. 2 Bali Master Stockist Sales Report

Data Preparation

This stage aims to ensure the quality of the data to be used is of good quality so that the next stages can run more efficiently. This process includes identifying, cleaning, and transforming data to eliminate anomalies, missing values, and reorganizing data to suit the needs of the analysis to be carried out. First, the attributes are selected and adjusted to leave the attributes that will be used in the implementation of the RFM model. The attributes used are attributes that contain the time span of transactions, the number of transaction frequencies made by customers during the specified period, and the nominal amount of transactions that occurred during the specified period, namely from January 1, 2023 to December 31, 2023. Data with the selected attributes will be transformed into RFM form. After obtaining the RFM value, data normalization is then carried out using the *min-max scaling* method to make the data have a smaller range with the same weight of 0 to 1.

Modeling

After the data is normalized, the next step is modeling with the *Self-Organizing Maps* method. The RFM value is used as the input layer in the SOM model to obtain the output layer of the clustering results. Each neuron in the

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

input layer will represent one dimension of the RFM model. For example, if there are three RFM attributes, then there will be three neurons in the input layer, each representing *recency, frequency, and monetary*. This stage starts from initializing the initial weights randomly with a range of 0 to 1. Then the 1st iteration process is carried out, namely calculating the *Euclidean* distance for each input and weight. From the calculation of the *Euclidean* distance, the result with the smallest value is the winning neuron or BMU (Best Matching Unit) which will go through the weight update stage. The process is carried out for each RFM input data then the next iteration is carried out. The new weight value is used to calculate the distance and update the weights in the next iteration, and so on. After reaching the maximum iteration, the data will be grouped into *clusters* based on the location of the winning neuron in each iteration performed.

Evaluation

This stage aims to test the results of SOM clustering. This evaluation stage uses the *silhouette score* method which ranges from -1 to 1. If the value obtained is close to or equal to 1, it indicates that the data is in the right cluster. The *silhouette* value is obtained by calculating the intra-cluster distance with the inter-cluster distance. In addition to evaluation using the *silhouette* method, *cluster* evaluation is also done by calculating DBI. The smaller the resulting value, it shows that the resulting *cluster* is well separated.

RESULT

The data that has been obtained is the sales data of *Master Stockist* PT SNS 21 Bali for 1 year, namely 2023. The data has many columns or attributes that are not needed in model building and has an irregular shape. Therefore, it is necessary to transform the data to ensure that it is ready to be used in the modeling process. In this *data preparation* stage, it will be divided into two parts, namely data structure transformation and RFM transformation. Data structure transformation aims to take the value of the attributes needed in the implementation of RFM only. In addition, the data structure transformation stage makes the data into a unified dataset with uniform format standards tailored to the modeling needs. The result of this stage converts the data in figure 2 into a more structured form by retrieving and rearranging the relevant columns which can be seen in figure 3.

	No	Tanggal	Transaksi	Member ID	Member	Jumlah
0	1	01/01/2023	10123207152	SNS0043555	MASTER STOCKIST	702640
1	2	07/01/2023	20123207574	SNS0171193	DESAK NYOMAN SU	29885440
2	3	07/01/2023	20123207567	SNS0203918	STOCKIS BATUBULA	71190368
3	4	07/01/2023	10123207576	SNS0344243	MAYLANI BIMANTA	725088
4	5	07/01/2023	20123207585	SNS0322769	WAYAN ARDANA 1	5950704
..	..	...	...	...	...	...
64	65	19/01/2023	20123208568	SNS0203918	STOKIS BATUBULA	63493040
65	66	20/01/2023	20123208706	SNS0064391	SUB STOCKIST AB	5950704
66	67	20/01/2023	10123208707	SNS0344285	SIGIT ARIS SANT	362544
67	68	20/01/2023	10123208708	SNS0344286	SIGIT ARIS SANT	4350528
68	69	20/01/2023	10123208710	SNS0344287	SIGIT ARIS SANT	4350528

Fig. 3 Data Structure Transformation Result

The next section, RFM transformation, aims to measure customer value based on three variables: *recency*, *frequency*, and *monetary*. With RFM transformation, the resulting data will be more structured and ready to be incorporated into the *Self-Organizing Maps* model. The results of this stage can be seen in figure 4.

	Member ID	Recency	Frequency	Monetary
0	SNS0043555	40	16	8507244
1	SNS0171193	5	60	987382151
2	SNS0203918	3	66	2584572081
3	SNS0344243	179	2	1450176
4	SNS0322769	2	122	1300389944
..	...	...	...	...
477	SNS0344275	9	1	4350528
478	SNS0355283	9	1	725088
479	SNS0355656	9	1	4350528
480	SNS0060312	7	1	362544
481	SNS0355698	3	1	725088

Fig. 4 RFM Transformation Result

After obtaining RFM values from each member who has made transactions for one year, the RFM data is normalized to ensure that each RFM metric has a balanced weight for the clustering process using SOM. The SOM algorithm works to group data that has similar patterns. The SOM traning process will produce a mapping of clusterization results. To ensure the quality of the clusters formed, cluster evaluation is carried out using *Silhouette Score* and *Davies-Bloudin Index*.

This research uses 297 trial combinations to find the best combination that produces *Silhouette Score* and *Davies Bouldin Index* evaluation values to measure *cluster* quality. This number of combinations is generated by doing manual variations, namely changing parameter values alternately manually, the results of which are poured into excel data. The varied parameters include 1x3 to 5x5 grid sizes that produce 3 to 25 *clusters*; iterations of 500, 750, and 1000; *learning rates of* 0.1, 0.2, and 0.5 with radii of 0.5 and 1. 5x5 grid size is determined as the

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

maximum grid size because based on the Vesanto formula the grid size that produces the maximum cluster is recommended to be the square root of the number of samples in the dataset. Table 3 shows the results obtained from the SOM exploration process with 297 parameter variations.

Table 3. SOM Exploration Result

No.	Iterat -ions	Learning Rate	Radius	SOM Width	SOM Height	Num-Clusters	Silhouette Score	Davies Bouldin Index
1	500	0,1	0,5	1	3	3	0,620305	0,605891
2	750	0,1	0,5	1	3	3	0,590733	0,514946
3	1000	0,1	0,5	1	3	3	0,533632	0,585189
4	500	0,2	0,5	1	3	3	0,501798	0,693794
5	750	0,2	0,5	1	3	3	0,589549	0,511876
6	1000	0,2	0,5	1	3	3	0,645785	0,568777
7	500	0,5	0,5	1	3	3	0,645785	0,568777
8	750	0,5	0,5	1	3	3	0,647608	0,536503
9	1000	0,5	0,5	1	3	3	0,647608	0,536503
10	500	0,1	1	2	2	4	0,439051	0,623599
11	750	0,1	1	2	2	4	0,472746	0,621663
12	1000	0,1	1	2	2	4	0,439855	0,662476
13	500	0,1	0,5	2	2	4	0,556222	0,581127
14	750	0,1	0,5	2	2	4	0,614938	0,57481
15	1000	0,1	0,5	2	2	4	0,486738	0,655144
16	500	0,2	1	2	2	4	0,418108	0,699307
17	750	0,2	1	2	2	4	0,614656	0,721507
...	...	...	...	...	...	...	...	...
292	500	0,5	1	5	5	24	0,380058	0,755328
293	750	0,5	1	5	5	25	0,350426	0,808303
294	1000	0,5	1	5	5	24	0,379028	0,774085
295	500	0,5	0,5	5	5	23	0,410511	0,755638
296	750	0,5	0,5	5	5	23	0,384685	0,779538
297	1000	0,5	0,5	5	5	24	0,411125	0,7431

From the experiments conducted, it can be seen that the larger the grid size used is in line with the number of clusters formed. Based on Table 2 *Silhouette Width Value*, the resulting *cluster* can be said to be good if the resulting evaluation value is more than 0.51 to 1 in the *reasonable cluster* to *strong cluster* category. As for the DBI value, the smaller the value, the better the cluster formed. In the exploration results, data that showed *silhouette* evaluation results of more than 0.51 were only 33 trial combinations. The best combination was found in the 8th trial with an iteration value of 750, learning rate 0.5, radius 0.5, grid size 1x3 resulting in 3 clusters with evaluation test results 0.647608 on *silhouette* and 0.536503 on *Davies-Bouldin Index*. The *clustering* results using the best trial combination can be seen in figure 5.

	Member ID	Recency	Frequency	Monetary	Cluster
0	SNS0043555	40	16	8507244	1
1	SNS0171193	5	60	987382151	2
2	SNS0203918	3	66	2584572081	2
3	SNS0344243	179	2	1450176	3
4	SNS0322769	2	122	1300389944	2
..	...	...	...	...	...
477	SNS0344275	9	1	4350528	1
478	SNS0355283	9	1	725088	1
479	SNS0355656	9	1	4350528	1
480	SNS0060312	7	1	362544	1
481	SNS0355698	3	1	725088	1

Fig. 5 Clustering Results with the Best Trial Combination.

After getting the results of the *clustering* process, all 482 member ids are divided into *cluster* 1 with 236 members, *cluster* 2 with 7 members, and *cluster* 3 with 239 members. Each *cluster* formed has a diverse range of RFM values. The RFM value of each *cluster* reflects the purchasing behavior of customers which can be seen with the RFM boxplot in figure 6.

* name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Nomor 4, October 2024
DOI : https://doi.org/10.33395/sinkron.v8i4.14112
e- ISSN : 2541-2019
p- ISSN : 2541-044X

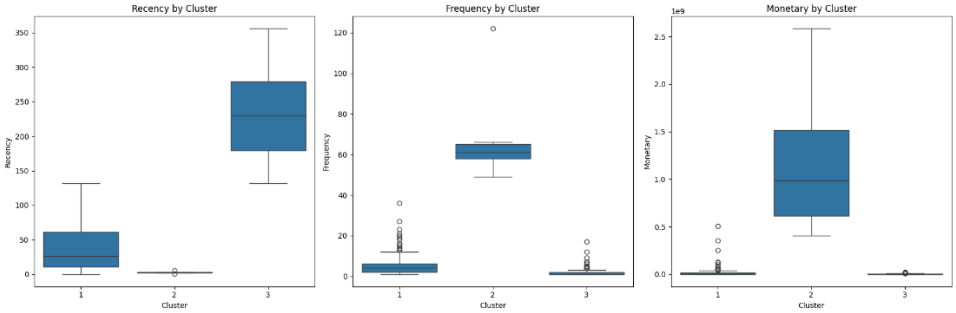

Fig. 6 Recency, Frequency, and Monetary Boxplot by Cluster.

In the recency boxplot, the first cluster has IQR (*Interquartile Range*) values ranging from 10 to 60 indicating that 75% of the members in this *cluster* have *recency* values below 60. While the other 25% are members with recency values above 60 which means the majority of members in this *cluster* have made fairly recent transactions in the last 60 days. In the second *cluster*, a very narrow range of *recency* values with a boundary value that is very close to 0 indicates that all members in *cluster* 2 make very recent transactions or show very recent interactions. The third *cluster* has a high *recency* value as seen from the minimum range of more than 100 to the maximum value of more than 350. This indicates that all members in *cluster* 3 are customers who have not made transactions for a long time.

In the frequency boxplot, the *frequency* value of the first *cluster* is around 1 to 10 transactions with several outliers caused by customers who make transactions much more frequently than the majority of other customers. In the second *cluster*, it can be seen that the *lower whisker* shows the lowest *frequency* value of around 50 transactions and the *upper whisker* is more than 60 with 1 *outlier* with a *frequency* value of more than 120. So it can be said that *cluster* 2 members have a very high *frequency* value compared to *clusters* 1 and 3. *Cluster* 3 shows a very low *frequency* value with several *outliers* that have a *frequency* value below 20.

The monetary boxplot shows the range of *monetary* values in each *cluster* with 1e9 which means $1x10^9$ or in billion. In the first *cluster* there is an outlier in the *monetary* value of around 500 million which is caused by customers who make higher value transactions than the majority of other members in *cluster* 1. *Cluster* 2 has data that has a maximum *monetary* value with a minimum value of around 400 million which indicates that in this *cluster* its members make transactions with a very high monetary value compared to *clusters* 1 and 3 which have very low *monetary* values.

So based on the boxplot analysis in figure 6, the characteristics possessed by *cluster* 1 are low *recency*, low *frequency*, and low *monetary* values. Customers included in *cluster* 1 can be categorized as new customers who need to be encouraged to make transactions more often. To increase customer activity in this *cluster*, there are several strategies that can be carried out by the company. The first is a customer loyalty program by providing special offers to encourage them to make repeat transactions so that there is an increase in their *frequency* and *monetary* value or transaction value. Second is a more personalized promotion based on the *monetary* value that has been spent by the customer. The third is to communicate periodically through monthly emails or text messages containing special offers at certain periods. Periodic communication will remind customers about new offers that can help maintain their *engagement* with the company.

Customers in *cluster* 2 have characteristics of very low *recency* value, high *frequency* seen from the highest number of visits by customers in this group, and very high *monetary* value. So that members of *cluster* 2 are categorized as active and high-value customers or loyal customers. To maintain customer loyalty in this *cluster*, the strategy that can be implemented by the company, for example, is to give loyalty awards because customers in this cluster are very active and have high transaction values. The company can provide special offers that are only available to customers in this cluster. Exclusive loyalty programs are provided to encourage them to continue transacting. In addition, personalized communication is also important to consider to help maintain and increase customer satisfaction and customer retention.

The characteristics possessed by *cluster* 3 are a very high *recency* value that touches the maximum number of *recency* indicating that the customer has not transacted for a long time, a low *frequency* value, and a low monetary value. This indicates that the amount of money spent in their transactions is fairly small. Customers who enter *cluster* 3 can be categorized as inactive customers or old customers who rarely make transactions that need to be reactivated. There are several strategies that companies can do for customers who are in this *cluster*. Given that members in this *cluster* rarely make transactions and have been inactive for a long time, customer reactivation strategies by sending attractive offers or special discounts can help revive their interest. In addition, it is necessary to conduct further root cause analysis to understand why customers in this *cluster* rarely transact related to problems in services or products that could have caused the inactivity. Other strategies include providing special promotions or incentives with the first purchase after a period of inactivity that encourages customers to transact again. Personalized communication is equally important to remind customers that the company is still paying attention to their presence and provide offers related to previous shopping history.

* name of corresponding author

e- ISSN : 2541-2019
p- ISSN : 2541-044X

DISCUSSIONS

When viewed as a whole, there is a pattern that can be recognized from the SOM exploration data in table 3 that has been done. Higher iteration values tend to produce better cluster evaluation values because the model has more opportunities to update neuron weights so that it can achieve good convergence. The learning rate influences how much updating is done to the neuron weights during the training process. Although smaller *learning rates* generally give better results, the SOM exploration showed that a *learning rate of* 0.5 with high iterations gave the best results. This combination allows for large weight changes, which are accelerated by high iterations, helping the model achieve convergence faster. The radius value determines how far the influence of weight updates from the central neuron to its neighboring neurons goes during training. A smaller radius value of 0.5 gives better evaluation results than a larger radius. A larger radius causes more neighboring neurons to be updated, but may cause the changes to spread too widely and reduce the quality of the updates. In addition, a smaller number of clusters tends to provide better evaluation compared to larger clusters. Using a larger grid increases the number of clusters, but decreases the quality of the clusters, resulting in poorer evaluation scores.

CONCLUSION

Customer segmentation using the RFM model and the *Self-Organizing Maps* algorithm can provide insight into customer characteristics that the company does not yet have. By using the best combination of SOM parameters, namely 750 iterations, *learning rate* 0.5, radius 0.5, and 1x3 grid size, this study successfully grouped customers into three *clusters* with an evaluation value of *Silhouette Score* of 0.647608 and *Davies-Bouldin Index* (DBI) of 0.536503. Each *cluster* reflects the characteristics of customers, where from 482 member IDs are divided into 3 groups. *Cluster* 1 with 236 members are new customers with low RFM values who need to be encouraged to make more frequent transactions. Strategies that can be applied are to provide special offers, personalized promotions, and regular communication. *Cluster* 2 with 7 members are customers with low *recency*, high *frequency*, and high *monetary* categorized as loyal customers. Strategies that can be done to maintain customer loyalty are giving loyalty awards, special offers, exclusive loyalty programs, and personalized communication. *Cluster* 3 with 239 members are customers with high *recency*, low *frequency*, and low *monetary* values which are categorized as inactive customers. Strategies that companies can do are reactivation strategies, special promotions or first purchase incentives after a period of inactivity, and personalized communication.

ACKNOWLEDGMENT

This research is supported and funded by the Ministry of Education, Culture, Research, Technology and Higher Education of the Republic of Indonesia with contract numbers 2927/LL8/AL.04/2024, and 004/INSTIKI.R4.D1/PM.03/06/2024.

REFERENCES

ASTUTI, R. D. (2019). *ANALISIS PERBANDINGAN ALGORITMA K-MEANS DAN K-MEDOIDS UNTUK MENERAPKAN SEGMENTASI Diajukan sebagai salah satu syarat untuk memperoleh gelar*. *14002140*.

Aziz, S., & Mustakim. (2022). Implementasi Algoritma Self Organizing Map untuk Identifikasi Pola Pengelompokan Tingkat Kesejahteraan Keluarga Kabupaten Siak. *IJIRSE: Indonesian Journal of Informatic Research and Software Engineering*, *2*(2), 82–92.

Bellotti, W., Davies, D., & Wang, Y. H. (2021). Improved Multi-index Customer Segmentation Model Research. *International Journal of Smart Business and Technology*, *9*(2), 49–64. https://doi.org/10.21742/IJSBT.2021.9.2.04

Fawaz, F., Fitriasari, N. S., & Rosalia, A. A. (2022). Perbandingan Algoritma Self Organizing Map dan Fuzzy C-Means dalam clustering hasil produksi ikan PPN Karangantu. *Explore: Jurnal Sistem Informasi Dan Telematika*, *13*(2), 102. https://doi.org/10.36448/jsit.v13i2.2783

Juhari, T., & Juarna, A. (2022b). Implementation Rfm Analysis Model for Customer Segmentation Using the K-Means Algorithm Case Study Xyz Online Bookstore. *Explore*, *12*(1), 107. https://doi.org/10.35200/explore.v12i1.548

Manalu, D. A., & Gunadi, G. (2022). IMPLEMENTASI METODE DATA MINING K-MEANS CLUSTERING TERHADAP DATA PEMBAYARAN TRANSAKSI MENGGUNAKAN BAHASA PEMROGRAMAN PYTHON PADA CV DIGITAL DIMENSI. *Infotech: Journal of Technology Information*, *8*(1), 45–54.

Mardiyah, I. (2022). *Implementasi Metode Self Organizing Maps Dalam Pengelompokan Wilayah Penyebaran Covid-19 di Provinsi Jawa Timur*. 1–101.

Muqsit, M. A., & Swanjaya, D. (2021). *Analisa Model Pengelompokan Data Survey Kepuasan Pelanggan Menggunakan Metode Self Organizing Maps*. 126–131. https://proceeding.unpkediri.ac.id/index.php/inotek/article/view/1091/702

MUSDARI, M., Amalia, E. L., & Yunhasnawa, Y. (2019). Clustering Untuk Sistem Manajemen Penjualan Di Klinik Laptop Menggunakan Metode Self Organizing Maps (SOM). *Seminar Informatika Aplikatif Polinema*, 18–23.

* name of corresponding author

Paembonan, S., & Abduh, H. (2021). Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, *6*(2), 48. https://doi.org/10.51557/pt_jiit.v6i2.659

Purbasari, I. Y., Puspaningrum, E. Y., & Putra, A. B. S. (2020). Using Self-Organizing Map (SOM) for Clustering and Visualization of New Students based on Grades. *Journal of Physics: Conference Series*, *1569*(2). https://doi.org/10.1088/1742-6596/1569/2/022037

Septiani, I. W., Fauzan, Abd. C., & Huda, M. M. (2022). Implementasi Algoritma K-Medoids Dengan Evaluasi Davies-Bouldin-Index Untuk Klasterisasi Harapan Hidup Pasca Operasi Pada Pasien Penderita Kanker Paru-Paru. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *3*(4), 556. https://doi.org/10.30865/json.v3i4.4055

Tholib, A. (2020). MANAJEMEN KLUSTERISASI PASAR: Penerapan Segmentasi Pelanggan Berbasis Metode Self-Organizing Map (SOM) di CV Karunia Probolinggo. *TRILOGI: Jurnal Ilmu Teknologi, Kesehatan, Dan Humaniora*, *1*(2), 35–45. https://doi.org/10.33650/trilogi.v1i2.1897

Twin, A. (2023). *What Is Data Mining? How It Works, Benefits, Techniques, and Examples*. Investopedia Is Part of the Dotdash Meredith Publishing Family. https://www.investopedia.com/terms/d/datamining.asp

Wei, J. T., Lin, S. Y., Yang, Y. Z., & Wu, H. H. (2020a). Using a combination of RFM model and cluster analysis to analyze customers' values of a veterinary hospital. *IAENG International Journal of Computer Science*, *47*(3), 442–448.

Wei, J. T., Lin, S. Y., Yang, Y. Z., & Wu, H. H. (2020b). Using a combination of RFM model and cluster analysis to analyze customers' values of a veterinary hospital. *IAENG International Journal of Computer Science*, *47*(3), 442–448.

Wicaksana, P. A., Swamardika, I. B. A., & Hartati, R. S. (2022). Literature Review Analisis Perilaku Pelanggan Menggunakan RFM Model. *Majalah Ilmiah Teknologi Elektro*, *21*(1), 21. https://doi.org/10.24843/mite.2022.v21i01.p04

Widiyanto, A. T., & Witanti, A. (2021). Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global). *KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi*, *1*(1), 204–215. https://doi.org/10.24002/konstelasi.v1i1.4293

Wijaya, I. D., Hendrawan, M. A., & Anabela, N. N. (2023). Pengelompokan Obyek Wisata Potensial dengan Self Organizing Maps (SOM) dan Sum Additive Weighting (SAW). *JISKA (Jurnal Informatika Sunan Kalijaga)*, *8*(1), 1–9. https://doi.org/10.14421/jiska.2023.8.1.1-9

* name of corresponding author