Journal of
Computational and Theoretical Nanoscience
Vol. 17, 1–8, 2020

Implementation of Clustering Algorithm Method for Customer Segmentation

Nurmalasari[1,*], Anna Mukhayaroh[1], Siti Marlina[1], Sari Hartini[2], Sri Muryani[2], Ahmad Sinnun[3], Siti Nurajizah[4], and Cep Adiwihardja[4]

[1]*Information System, STMIK Nusa Mandiri, Jalan Damai No. 8, Warung Jati Barat, South Jakarta, 12740, Indonesia*
[2]*Technical Information, STMIK Nusa Mandiri, Jalan Damai No. 8, Warung Jati Barat, South Jakarta, 12740, Indonesia*
[3]*Software Engineering, Bina Sarana Informatika University, Jalan Kamal Raya No. 18, West Jakarta, 11730, Indonesia*
[4]*Information System, Bina Sarana Informatika University, Jalan Kamal Raya No. 18, West Jakarta, 11730, Indonesia*

The intense competition in the sale of goods and services in the digital era of e-commerce requires to manage customers optimally. Some online shops try to improve their marketing strategies by classifying their customers. This study aims to determine potential customers, namely loyal customers. Potential customers can be determined by customer segmentation. Sampling from several online shops in Indonesia. The model used for segmentation is RFM (Recency, Frequency, and Monetary) and data mining techniques, namely clustering method with the K-Means algorithm. The results of this segmentation research divide the customer into 2 clusters. The best number of clusters is determined based on the Davies Bouldin index. The first cluster is cluster 0 consisting of 261 customers with RFM Score between 111–543. The first cluster includes the Everyday Shopper group. The second cluster, cluster 1 consists of 102 customers with RFM Score 443–555. The second cluster includes the Golden Customer group. With the existence of research on customer segmentation, it is expected to help in grouping customers so that companies can determine the right strategy for each group of customers.

Keywords: Clustering, K-Means, Customer Segmentation.

RESEARCH ARTICLE

1. INTRODUCTION

Marketplace is a website or online application that facilitates the process of buying and selling from various stores. Actually, the online marketplace has a concept that is more or less at the same as the traditional market. Basically, marketplace owners are not responsible for goods sold because their job is to provide a place for sellers who want to sell and help them to meet customers and make transactions more simply and easily. The transaction itself is regulated by the marketplace. Then after receiving payment, the seller will send the item to the buyer. One reason why a well-known marketplace is due to ease and convenience in use. Many describe online marketplaces such as department stores. In Indonesia alone, we already have several well-known local marketplaces such as Tokopedia, Buka Lapak, Shopee and Lazada. Every marketplace has an operational system in every transaction of its operations, always recorded and documented. The data is stored in a large capacity database. For companies, the data stored in the database can be used to make sales reports, inventory controls, and so on. In recent years, data has become increasingly heterogeneous and complex with volumes increasing rapidly exponentially.

According to Turner in Ref. [1] in 2013 the volume of data had become 44 zettabytes in 2020. Therefore, currently known as big data, which describes the volume of data is very large, structured and unstructured, which flooded the business world.

In a business world that is always dynamic and full of competition, business people must always think of ways to continue to survive and if possible grow their business. To achieve this, there are three business needs that can be done, namely the addition of types and increase in product capacity, reduction in the company's operating costs, and increased marketing effectiveness and profits [2]. In order to be able to meet business needs above there are many ways that can be taken, one of which is by analyzing company data.

Some marketplaces such as Tokopedia, Buka Lapak, Shopee and Lazada are a growing Starup industry for

* Author to whom correspondence should be addressed.

sales from various product categories. With the support of android marketplace applications, it will make it easier for the customers to transact anywhere and anytime. This of course leads to competition between companies. The main focus of the company to compete with its competitors is customers [3]. Customers occupy important positions in developing business strategies, customers are also a source of profit in the company [4]. For that we need a good understanding of the customer. The problem that is often faced is the difficulty in analyzing customer value. Many companies have difficulty identifying the right customer or customer, this can result in the company losing potential customers and of course it will be very detrimental to the company.

The purpose of this research is to determine customer profile segmentation more precisely and to segment customers which will then be used to determine the level of customer loyalty related to the marketing strategy. Data mining techniques used to find customer segmentation using the *K*-Means Clustering method. The results of the Clustering will then be classified to determine customer segmentation using the RFM Model. RFM Model is a model for determining consumer segmentation based on Recency, Frequency, and Monetary. Determination of the most optimal number of clusters using Davies Bouldin Index. In this study we took several samples related to sales data from four marketplaces in Indonesia and applied RFM (Recency, Frequency, Monetary) and *K*-Means clustering methods as a method of analyzing customer segmentation in one sales area of nine existing sales areas in Jakarta at the marketplace. The data to be analyzed is sales data in 2017 as many as 7574 transactions by 363 customers.

2. LITERATURE REVIEW

2.1. Data Mining

Data mining is used to find knowledge hidden in a database. Data mining is a semi-automatic process is that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful potential and useful knowledge information stored in large databases [5]. According to Gartner Group data mining is a process of finding meaningful relationships, patterns, and tendencies by examining in large groups of data stored in storage using pattern recognition techniques such as statistical and mathematical techniques [6].

According to Fayyad et al. in Ref. [1] explained that Data mining or data mining is a technique that is relatively fast and easy to find knowledge, patterns and or relationships between data, automatically. By combining the four computer science disciplines as above, knowledge can be found in five sequential processes: selection, preprocessing, transformation, data mining, and interpretation/evaluation. The extraordinary progress that continues in the field of data mining is driven by several factors, including [6]:

1. Rapid growth in the data set.
2. Storage of data in the data warehouse, so that all companies have access to a good database.
3. Increased data access through web and intranet navigation.
4. Pressure of business competition to increase market control in economic globalization.
5. The development of software technology for data mining (availability of technology).
6. A great development in computing capabilities and the development of storage media capacity.

Relationships sought in data mining can be a relationship between two or more in one dimension. For example, in the dimensions of the product, it can be seen the relationship of purchasing a product with another product. In addition, relationships can also be seen between two or more attributes and two or more objects [7]. Meanwhile, the discovery of patterns is another output of data mining. Suppose a company will increase credit card facilities from customers, then the company will look for patterns from existing customers to find out potential customers and potential customers. Some initial definitions of data mining include a focus on the automation process. Reference [8] in the Data Mining Technique for Marketing, Sales, and Customer Support book defines data mining as a process of exploring and analyzing automatically or semi-automatically for large amounts of data with the aim of finding meaningful patterns or rules [6].

The statement emphasizes that in data mining automation does not replace human intervention. Humans must be active in every phase of the data mining process. The greatness of the ability of the data mining algorithm contained in the analysis software that is currently available allows the use of errors that are fatal. Users may apply inappropriate analysis of data sets using different approaches. Therefore, an understanding of the statistics and structure of mathematical models that underlie the work of software is needed [6].

Data mining is not a completely new field. One of the difficulties in defining data mining is the fact that data mining inherits many aspects and techniques from established fields of science first.

According to Ref. [9] argues that "Data mining is a series of processes to explore added value in the form of information that has not been known manually from a data." Data mining is mainly used to search for knowledge contained in large databases so that it is often called Knowledge Discovery Databases (KDD). The stages of the KDD process according to are as follows:

1. Data cleaning (for removing inconsistent data and noise).
2. Integration of data (merging data from several sources).
3. Data transformation (data is converted into the appropriate form for mining).

4. Application of data mining techniques, pattern extraction processes from existing data.
5. Evaluation of patterns found (the process of interpreting patterns into knowledge that can be used to support decision making)
6. Presentation of knowledge with visualization techniques.

2.2. Data Mining Grouping

Data mining is divided into several groups based on tasks that can be done, namely:
1. Description Sometimes researchers and analysts simply want to try to find ways to describe the patterns and trends contained in the data.
2. Estimates
3. Estimates are almost the same as classifications, except the target variable is estimated more in the numerical direction than in the direction of the category. The model built with a complete record provides the value of the target variable as a predictive value. Furthermore, in the next review the estimated value of the target variable is based on the value of the predictive variable.
4. Prediction
5. Prediction is almost the same as classification and estimation, except that in predicting the value of the results there will be in the future. Some methods and techniques used in classification and estimation can also be used (for the right conditions) for predictions.
6. Classification: In classification, there are target categorical variables. For example, income classification can be separated into three categories, namely high income, medium income, and low income.
7. Clustering
8. Clustering is a grouping of records, observations, or observing and forming a class of objects that have similarities. Cluster is a collection of records that have similarities with one another and have an incompatibility with records in other clusters. Clustering is different from classification, namely there is no target variable in clustering.
9. Association: The task of associations in data mining is to find attributes that appear at one time. In the business world it is more commonly called shopping basket analysis [6].

2.3. Architecture of the Data Mining System

The main architecture of the data mining system, generally consists of several components as follows:
1. Databases, data warehouses, or information storage media, consist of one or several databases, data warehouses, or other forms of data. Data cleaning and data integration is done on the data.
2. Database, data warehousing, is responsible for finding relevant data in accordance with what the user or user wants.
3. The Knowledge Base is the knowledge base used as a guide in pattern searching.
4. Data mining engine is an important part of the system and ideally consists of a collection of function modules used in characterization, classification, and cluster analysis. And is part of the software that runs programs based on existing algorithms.
5. Evaluation of patterns (pattern evaluation), these components generally interact with data mining modules. And part of the software that functions to find patterns or patterns contained in the database that is processed so that later the data mining process can find the appropriate knowledge.
6. Interface (Graphical user interface) is a communication module between users or users with a system that allows users to interact with the system to determine the data mining process itself.

2.4. Market Basket Analysis

Market basketball analysis is one way that is used to analyze sales data from a company. This process analyzes the buying habits of consumers by finding associations between different items that consumers put in shopping for basketball. These obtained results can later be utilized by retail companies such as shops or supermarkets to develop marketing strategies by looking at items that are often purchased simultaneously by consumers [10].

For some cases, the pattern of items purchased simultaneously by consumers is easy to guess, for example milk is bought along with bread. However, there might be a pattern of purchasing items that had never been thought of before. For example, purchasing cooking oil with detergent. It is possible that this pattern has never been thought of before because cooking oil and detergent have no connection at all, both as a complementary item and a substitute for goods. This may not have been thought of before, so that it cannot be anticipated if something happens, such as a detergent stock shortage, for example. This is one of the benefits that can be obtained from doing market basket analysis. By doing this process and using a computer, automatically a manager does not need to experience difficulties in finding patterns about what items might be purchased simultaneously, because the data from the sales transaction will tell itself.

2.5. Cluster Analysis

Reference [9] finding that "clustering is an unsupervised data mining method, because there is no one attribute that is used to guide the learning process. So all input attributes are performed equally." Cluster analyst or clustering is the process of dividing data in a set into several groups whose similarity of data in a group is greater than the similarity of the data with the data in other groups.

Cluster analysis is a multivariate analysis technique that aims to cluster observational data or variables into clusters

in such a way that each cluster is homogeneous according to the factors used for clustering. Because what is desired is to get a cluster that is as homogeneous as possible, then what is used as the basis for clustering is the similarity of the score scores analyzed. Data on the size of similarity can be analyzed by cluster analysis so that it can be determined who entered which cluster [11].

2.5.1. Formulate a Problem

The most important thing in the problem of cluster analysis is the selection of variables that will be used for clustering (cluster formation). Include one or two variables that are not relevant to the clustering problem so that it will lead to deviation from clustering results which are likely to be very useful.

2.5.2. Choose the Size of Distance

The purpose of cluster analysis is to group similar objects into the same cluster. Therefore it requires several measures to find out how similar or different these objects are. The usual approach is to measure the similarity expressed in distance between pairs of objects. In cluster analysis there are three measures to measure the similarity between objects, namely the size of the association, the size of the correlation, and the measure of closeness.

2.5.3. Choosing a Clustering Procedure

The cluster formation process can be done in two ways, namely by hierarchical and non-hierarchical methods. The hierarchy method consists of agglomerative methods and devisive methods. The agglomerative method itself consists of 3 methods, namely the linkage method, the variance method, and the centroid method, where the linkage consists of a single linkage method, complete linkage, and average linkage. Whereas the variance method consists of the Ward method.

2.5.4. Determine the Number of Clusters

The main problem in cluster analysis is determining how many clusters. Actually there are no standard rules for determining how many clusters actually are, but there are some clues that can be used, namely [12]

- Theoretical, conceptual, practical considerations may be suggested/suggested to determine how many clusters actually are. For example, if the purpose of clustering is to identify/identify market segments, management might want a certain number of clusters (say 3, 4, or 5 clusters).
- The relative magnitude of the cluster should be useful.

2.5.5. Interpret the Profile of Clusters (Formed Clusters)

At the stage of interpretation it involves testing each cluster formed to give the name or description correctly as a description of the nature of the cluster, explaining how they can be relevant in each dimension differently. When starting the process of interpretation the average (centroid) is used for each cluster on each variable. The purpose of Cluster Analysis is to group objects based on characteristic similarities between these objects. Thus, the characteristics of a good cluster are having:

- Internal homogeneity (within clusters); namely the similarity between members in one cluster.
- External heterogeneity (between clusters); that is the difference between one cluster with another cluster.

2.6. *K*-Means Cluster

K-Means are included in partitioning clustering, which is that each data must be included in a particular cluster and allow for each data included in a particular cluster at a process stage, in the next step to move to another cluster [13]. Data Clustering is one of the Data Mining methods that is non-directive (unsupervised). There are two types of data clustering that are often used in the process of group- ing data, namely hierarchical (hierarchical) data cluster- ing and non-hierarchical (non hierarchical) data clustering. *K*-Means is one method of non-hierarchical clustering data that attempts to partition existing data into one or more clusters/groups. This method partition the data into clusters/groups so that data that has the same characteristics are grouped into one and the same cluster of data that has different characteristics grouped into other groups. The purpose of this clustering data is to minimize objective functions that are set in the clustering process, which generally attempts to minimize variations in a cluster and maximize variations between clusters. The benefits of Clustering are as Object Identification (Recognition) for example in the field of Image Processing, Computer Vision or robot vision. Besides that, it is a Decision Support System and Data Mining such as market segmentation, regional mapping, marketing management and others.

The *K*-Means method according to Ref. [14] is as follows:

1. Determine the number of clusters.
2. Allocate data in accordance with the number of clusters that have been determined.
3. Calculate the centroid value in each cluster.
4. Allocate each data to the nearest centroid.
5. Return to step 3, if there is still data transfer from one cluster to another cluster, or if the change in the centroid value is still above the specified threshold value, or if changes to the objective function value are still above the specified threshold value.

K-means characteristics:

1. *K*-means are very fast in the clustering process.
2. *K*-means are very sensitive to random generation of early centroids.
3. Allows a cluster to have no members.
4. *K*-means clustering results are unique (always changing, sometimes good, sometimes not good).

RESEARCH ARTICLE

To calculate the centroid cluster i, v_i, use the following formula:

$$v_{ij} = \frac{\sum_{k=1}^{N_1} X_{kj}}{N_i} \quad (1)$$

Where N_i: Amount of data that is a member of the i cluster.

2.7. Recency Frequency Monetary (RFM) Model

Reference [15] stated that the RFM model is a behavior-based model used to analyze customer behavior and then make predictions based on database behavior. This RFM model is an old and popular method for measuring customer relationships. The definition of the RFM model: Recency is when the last transaction was made. Frequency is the number of transactions made by customers. For example, twice a year or three times a month. While monetary is the magnitude of the value of the transactions carried out.

The value of recency, frequency, monetary is divided into five parts with values 5, 4, 3, 2 and 1. The recency value is calculated based on the date of the last transaction or the time interval of the last transaction with the current. Customers with the date of the latest transaction have a value of 5 while the customer with the farthest transaction date in the past has a value of 1. Likewise with the value of frequency, customers who often transact have a high frequency value, i.e., 5. Customers who rarely transact have a value of 1. Customer which has a large deposit balance that has a high monetary value, with a value of 5. Conversely a customer who has a small deposit balance has a low monetary value, namely 1.

2.8. Customer Segmentation

Segmentation is the process of dividing customers into clusters with customer loyalty categories to build marketing strategies. Customer segmentation is divided into 6 characteristics based on RFM values [16].

3. METHODOLOGY/MATERIALS

The methodology used is CRISP-DM (Cross Industry Standard Process for Data Mining). There are 6 stages in CRISP-DM, namely [15].

3.1. Business Understanding

Understand the goals and needs in the business scope or research unit, translating this knowledge into data mining problems.

3.2. Data Understanding

Collecting data, if data comes from more than one database then the data integration process is carried out. Furthermore, understanding data, identifying data quality, checking data and cleaning invalid data or data cleaning processes.

3.3. Data Preparation

At this stage, collect data that will be used for the next stage or process the data selection. Select the variables to be analyzed, prepare the initial data so that it is ready for modeling or data transformation.

3.4. Modeling

This stage includes the selection and application of various modeling techniques to obtain optimal values. There are several different techniques that are applied to the same data mining problems and there are also modeling techniques that require special data formats.

3.5. Evaluation

Determine whether the model is in accordance with the objectives at the initial stage (business understanding).

3.6. Deployment

In this stage the knowledge or information that has been obtained is presented.

4. RESULTS AND FINDINGS

4.1. Business Understanding

The case study of this study was conducted in several marketplaces in Indonesia, namely Tokopedia, Open Lapak, Shopee and Lazada in Jakarta area customers. The business patterns that occur from the sales process at the marketplace get transaction data of 7574 records from a database of several marketplace. The business objectives in this study increase and maintain the number of customers, especially potential customers.

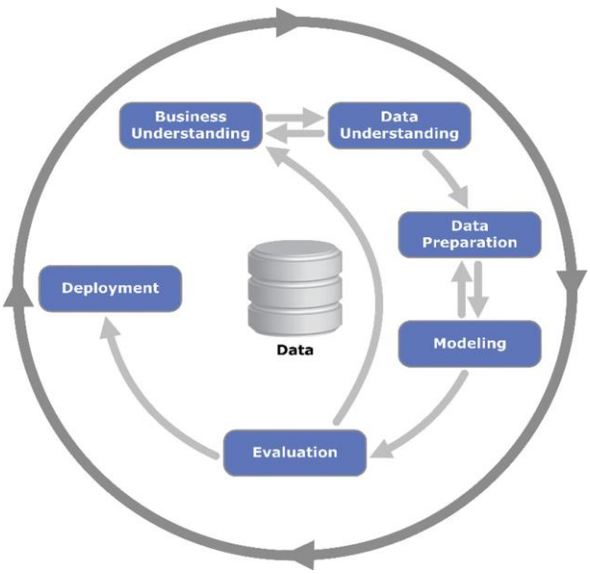

Fig. 1. CRISP-DM process cycle in this researched.

RESEARCH ARTICLE

Translating business objectives into data mining objectives in this study is customer segmentation, which is customer segmentation which will then be used to determine the level of customer loyalty and related marketing strategies.

4.2. Data Understanding

The data used in this research is 1-year transaction data during 2017. From the results of the stages of data collection in the research methodology, the researcher got 7574 records of data. Data is obtained from accessing the marketplace database server. To access the database server, permission from the relevant administrator is required. Data from the database is then exported to an excel file.

4.3. Data Preparation

Data preparation is the most important and often time-consuming data mining projects. This phase consists of Cleaning data, data reduction, feature selection and data transformation. Data cleaning is done on the data obtained. In the dataset, there are attributes of salesmen, years, months, customers, date_l, SumOfNetto, and invoices. Sales, year, month and invoice attributes are deleted because the data is not used in this study. As for data cleaning, data processing is done through Microsoft Excel so data is formed according to research needs. Of the 7574 records, processed by calculating the number of transactions for each customer and producing 363 records with customer attributes, the date of the last transaction, the end date of the period, the total transaction, and the total sales balance. Furthermore, attribute changes are made to make it easier to process RFM models by producing attributes of Recency (R), Frequency (F), and Monetary (M).

The next step is the RFM Model Process. After the RFM data is obtained, the next process is determining the RFM score, where each attribute is calculated as the score.

Table I. RFM model.

Customer class	Characteristics
Superstars	1. Customers with high loyalty. 2. Having high monetary value. 3. Has a high frequency. 4. Having the highest transaction.
Golden customer	1. Having the second highest monetary value. 2. High frequency. 3. Having an average transaction.
Typical customer	Has an average monetary value and average transaction.
Occasional customer	1. The second lowest monetary value after the dormant customer. 2. The lowest recency value 3. The highest transaction.
Everyday shopper	1. Having an increase in transactions. 2. Low transactions. 3. Has moderate to low monetary value
Dormant customer	1. Has the lowest frequency and monetary. 2. The lowest recency value.

Table II. Mapping of RFM initial data.

Initial attribute	Final attribute
Distance between the date of the last transaction and the date of the end of the study period	Recency (R)
Number of transactions for one year	Frequency (F)
Total sales balance for each customer	Monetary (M)

The RFM calculation method in this study uses a simple fixed method. Determination of scoring by this method depends on the business patterns of each company. The company has a business pattern that has not changed for one year such as payment methods, customer order patterns through salesmen so that the pattern of Recency, Frequency, and Monetary is possible. As long as the company does not change the business patterns they use such as payment methods, customer order patterns through salesmen, RFM calculations through the simple fixed range method can be used. The stages of determining the RFM Score are as follows.

4.3.1. Determination of R Score

In determining the Recency Score, researchers use the distance of the Recency value based on the longest Recency, which is 348 days until the latest Recency is 1 day. The criteria used were Recency 348 days as the lower limit for R score 1, average Recency included in R score 3 or middle, Recency 1 day for the upper limit of R score 5.

4.3.2. Determination of F Score

In determining the Frequency Score, the researcher uses the lowest frequency value based on Frequency, which is 1 time to the highest Frequency, which is 51 times. The criteria used are Frequency 1 time for the lower limit of the F score 1, the average Frequency becomes the upper limit in the F score 3, and the Frequency is 51 times for the upper limit of the F score.

4.3.3. Determination of M Score

In determining the Monetary Score, researchers used the lowest Monetary based Monetary value range of 117,600 to the highest Monetary value of 463,664,300. The criteria used are Monetary as many as 117,600 for the lower limit of the M Score value, Monetary average included in M Score 3, and Monetary as much as 463,664,300 for the upper limit of the M score.

Table III. Criteria for determining R score.

R score	Range of R values
5	1 to 4 days
4	15 to 45 days
3	46 to 120 days
2	121 to 179 days
1	180 to 348 days

RESEARCH ARTICLE

Table IV. Criteria for determining *F* score.

F score	Range of *F* values
5	Above 36 to 51 times
4	21 to 35 times
3	9 to 20 times
2	4 to 8 times
1	1 to 3 times

4.3.4. Result of RFM Score

Calculation of RFM Score through Microsoft Excel, namely:

$$\text{RFM Score} = (\text{Recency Values} \times 100) + (\text{Frequency Values} \times 10) + \text{Monetary Values}$$

The normalization process is carried out before the dataset is processed through the Rapidminer application. Normalization is done so that the scale of the data is not too far away. There are several methods/techniques applied for data normalization. In this study the normalization technique used was the normalization of *Z* Score. Also called zero-mean normalization, where the value of an attribute A is normalized based on the average value and the standard deviation of attribute A. A v value of attribute A is normalized to attribute v.

$$v' = \frac{v - \bar{A}}{aA}$$

Dimanav': Normalized value of *z* score. v: Initial Values. $\bar{A}$: Average. aA: Standard deviation.

Normalizing the *z* score on RFM data using Microsoft Excel software. Each Recency, Frequency, and Monetary value is transformed into a normalization *z* score.

4.4. Modelling

Data mining techniques used are clustering, with the *K*-Means algorithm. In *K*-Means the number of clusters must be determined by the decision maker. To identify optimal k, various tests can be used. In this study optimal cluster testing uses davies bouldin index.

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i=j} \left\{ \frac{A(i) + A(j)}{D(i,j)} \right\}$$

$$D(i,j) = \min x \in Si,\ y \in Sj\ \text{dist}(x,y) \quad (2)$$

Table V. Criteria for determining *M* score.

Nilai M	Rentang Nilai M
5	150.000.001 sampai 463.664.300
4	50.000.001 sampai 150.000.0000
3	10.000.001 sampai 50.000.000
2	2.000.001 sampai 10.000.000
1	117.600 sampai 2.000.000

Table VI. Table attribute *Z* score.

	$\bar{A}$ (Mean)	aA (Standar Deviasi)
Recency	35,95	75,46
Frequency	20,87	13,01
Monetary	32.791.863	45.304.807,47

$$A(i) = \max x,\ y \in Si\ \text{dist}(x,y) \quad (3)$$

$$A(j) = \max x,\ y \in Sj\ \text{dist}(x,y) \quad (4)$$

The small Davies Bouldin Index value is a good number of clusters. The smaller the davies bouldin index the more optimal the cluster results. In this study the clustering process and testing of the bouldin index using the Rapid Miner 9.0 software.

From the results of the Clustering conducted in the previous sub-chapter, the number of Cluster 2 is the most optimal cluster. This is because the value of the davies bouldin index in Cluster 2 shows the smallest value. A high negative value indicates a good performance of the index. Comparison of the test results of the bouldin index davies can be seen in Table VII.

The two optimal clusters formed are Cluster 0 of 261 id, and Cluster 1 of 102 id. Of the two Clusters, it is categorized into RFM Model in Table II. 1 based on the RFM Score. A summary of the results of segmentation based on RFM can be seen in Table VIII.

Cluster 0 includes the Everyday Shopper group because it has an increase in transactions and has a moderate to low Monetary value with an RFM Score range between 111 to 543. Cluster 1 belongs to the Golden Customer group because it has a high transaction frequency value and has a high to highest Monetary value. Cluster 1 has an RFM Score of 443 to 555.

Based on the two clusters in processing the data above, cluster 1 as a Golden Customer that has high frequency characteristics, namely 102 customers has an average frequency of transactions 36 times in one year. The average transaction in cluster 1 is higher than cluster 0 as Everyday Shopper which has an average transaction of 14.95 times in one year. Cluster 1 as a Golden Customer can prove that customers with high order frequencies have an influence on mapping potential customers.

4.5. Evaluation

Evaluation of the model used by testing davies bouldin index. From these tests generated groupings with 2 clusters

Table VII. Test results for the davies bouldin index.

Number of clusters	Davies Bouldin index
2	−0,344
3	−0,244
4	−0,221
5	−0,247
6	−0,234

RESEARCH ARTICLE

Table VIII. Summary of segmentation results.

Cluster	Number of customer	Recency average	Frequency average	Monetary average	RFM score
Cluster 0	261	47,90	14,95	14.736.711	111–543
Cluster 1	102	5,39	36	78.942.790	443–555

have the smallest value of davies boulden index. Therefore, customers are grouped into 2 clusters.

4.6. Deployment

The deployment process has not been carried out.

5. CONCLUSION

Customer segmentation generated in the research through the *K*-Means Clustering process in several Marketplace, namely as many as 2 clusters. The optimal number of clusters in customer segmentation is based on the Davies Bouldin Index. A high negative value shows good performance from Index. The Davies Bouldin Index value of cluster 2 is 0,344 while the number of cluster 3 is − 0,244, the number of cluster 4 is − 0,221, the number of cluster 5 is − 0,247, and the number of cluster 6 is − 0,234. The two clusters produced are cluster 0 as many as 261 customers and cluster 1 as many as 102 customers. Cluster 0 has an RFM Score between 111 to 533 including the Everyday Shopper group because it has increased transactions and has moderate to low monetary value. Cluster 1 has RFM Score 434 to 555 including the Golden Customer group because having a high transaction frequency value has a high to the highest monetary value. Customers with high order frequency have an influence on the mapping of potential customers that can be answered with the formation of cluster 1 as a Golden Customer who has a high transaction average of 36 transactions over a year. With this research on customer segmentation, it is expected to help companies in grouping Marketplace customers so that companies can determine the right strategy for each customer group.

Acknowledgments: Thanks to all my friends from Information System and Technical Information at STMIK Nusa Mandiri that is Sari Hartini, Anna Mukhayaroh, Sri Muryani, Siti Marlina, and friends from Software Engineering and Information System at Bina Sarana Informatika University, Ahmad Sinnun, Siti Nurajizah, Cep Adiwihardja who have contributed to the writing of this article and companies that have helped in research this. Specially thanks to my beloved husband and my children for their support. I hope this research is useful for us and helps other researchers who are conducting research in this field.

References

1. Suyanto, **2017**. *Data Mining for Data Classification and Clustering*. Bandung, Informatika Bandung.
2. Sutrisno, S., Afriyudi, A. and Widiyanto, W., **2013**. Application of data mining at sales using clustering method case study Pt. indomarco palembang. *Penerapan Data Mining Pada Penjualan Menggunakan Metode Clustering*, pp.1–11.
3. Wulandari, G.F., **2014**. Customer segmentation using *K*-Means algorithm for customer relationship management (CRM) at Miulan Hijab. pp.5–6.
4. Widiarina and Wahono, R.S., **2015**. Dynamic cluster algorithm for cluster optimization in *K*-means algorithm in mapping potential customers. *Journal of Intelligent Systems, 1*(1), pp.33–36.
5. Turban, E., Aronson, J. and Liang, T., **2005**. *Decission Support System and Intelligent System*. Indiana, Pearsan Hall Universitas.
6. Larose, D.T. and Larose, C.D., **2014**. *Discovering Knowledge in Data: An Introduction to Data Mining*. Second edn., New Jersey, John Willey & Sons, Inc.
7. Thavalingam, A., Bicanic, N., Robinson, J. and Ponniah, D., **2001**. *Computational Framework for Discontinuous Modelling of Masonry Arch Bridges*. Elsevier.
8. Berry, M.J. and Linoff, G.S., **2004**. *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*. United Stated, Wiley Publishing.
9. Vulandari, T.R., **2017**. *Data Mining Teori dan Aplikasi Rapidminer*. Surakarta, Java Media.
10. Han, J., Kamber, M. and Pei, J., **2011**. *Concepts and Techniques*.
11. Gudono, **2011**. *First Edition Multivariate Data Analysis*. Yogyakarta, BPFE.
12. Laeli, S., **2014**. *Cluster Analysis with Average Linkage Method and Ward's Method for Unit Link Life Insurance Customer Respondents Data*. Universitas Negeri Yogyakarta.
13. Azis, W.S. and Atmajaya, D., **2016**. Grouping of student interests using the *K*-means method. *ILKOM Scientific Journal*, *8*(2), pp.89–94.
14. Munigsih, E. and Kiswati, S., **2015**. Application of the *K*-means method for clustering online shop products in stock determination. *Journal Bianglala Informatika, 3*(1), pp.10–17.
15. Hardiani, T., Sulistyo, S. and Hartanto, R., **2015**. Savings Customer Segmentation Using RFM (Recency, Frequency, Monetary) and *K*-Means Models in Microfinance Institutions. *Seminar Nasional Teknologi Informasi dan Komunikasi Terapan (SEMANTIK)*, pp.463–468, DOI: 10.1016/S1875-5364(14)60001-7.
16. Yuliari, N.P.P., Putra, I.K.G.D. and Rusjayanti, N.K.D., **2015**. Customer segmentation through fuzzy C-means and fuzzy RFM method. *Journal of Theoretical and Applied Information Technology, 78*(3), pp.380–385.

Received: 23 May 2019. Accepted: 13 September 2019.

RESEARCH ARTICLE