Customer Segmentation of Personal Credit using Recency, Frequency, Monetary (RFM) and K-means on Financial Industry

Hafidh Rizkyanto[1], Ford Lumban Gaol[2]
Computer Science Department, Bina Nusantara University, Jakarta, Indonesia[1]
Doctor of Computer Science Department, Bina Nusantara University, Jakarta, Indonesia[2]

***Abstract*—This research focuses on how to build a segmentation model for credit customers to identify the potential for defaulting credit customers based on their transaction history. Currently, there is no segmentation available for this possibility of payment failure. Credit scoring helps in minimizing credit risk when applying for credit. However, using RFM (Recency, Frequency, Monetary) models helps to score each transaction variable of the customer's financial activity. K-means then assists in the process of segmenting the results of the RFM model scoring, which occurs in the middle of the customer's repayment schedule. Challenge is how to decide the variable that can be used in RFM models and how to interpret the clusters that have been formed and the actual implementation of the customer. The Bank can divide the clusters that have possibility of payment failure by their customers so that banks can take preventive actions and as information for the collection system to be able to make payment withdrawals or billing.**

***Keywords*—Credit; credit risk; recency; frequency; monetary; K-means**

I. Introduction

A commercial bank is a business entity that obtains funds from the public in the form of savings, term deposits, and other types of funds. These funds are then distributed to the public as credit with the purpose of improving their standard of living. The financial industry in banking offers various alternative money loans to the public, with the provision of loans in the form of credit to bank customers [1].

Regarding types of businesses, financial service industries that provide funds directly include commercial banks, guarantee companies, and factoring companies [2]. In this research, Bank XYZ, a commercial bank, is used as a case study. There are several types of commercial banks in Indonesia, including state banks, private national banks, foreign banks, joint banks, regional government banks, and commercial banks [3].

As of January 11, 2021, there were 198,986 active credit customers, with a total of 621,968 active and closed credit data for Bank XYZ. For the financial industry in banking, data is an important asset that can be used for corporate strategy. Banks have a large amount of raw data, including transaction data, customer data, and statement data, which require processing to provide decision support information [4]. However, Bank XYZ, established on July 10, 1970, currently does not have a decision support system that can provide information to credit officers regarding potential customers or customers who may have payment difficulties, which could significantly impact the bank. Therefore, with the implementation of a data processing system, credit handling can be improved by each account officer who monitors the credit of their customers. This will help to reduce errors in the amount and credit payment process, which often results in changes in collectability and poor customer performance, leading to poor credit quality.

In general, data mining functions are divided into two parts: descriptive and predictive. Other functions include classification, association, clustering, sequencing, and forecasting [5]. Data mining techniques can be used to segment credit customers, with clustering as the algorithm used, specifically K-means. The RFM Model is used because it can be adapted to evaluate the value of customers [6] and classify them in different service areas such as finance, telecommunications, and e-commerce [7]. K-means is a non-hierarchical data grouping method that can partition data into two or more groups [8].

The combination of the RFM and K-Means Models can produce optimal segmentation because RFM establishes variables that are closely related to business needs, and K-Means can group them based on the similarities of each customer. The evaluation of the number of segments is determined using the Calinski-Harabasz Index (CH), which gives better results than clustering evaluation methods such as the Elbow method and others [9].

In providing loans, banks face various problems and risks, including the behavior of customers who do not pay their installments on time or delay payment of installments for several months, resulting in bad credit. Therefore, it is essential for banks to evaluate credit risk by focusing on several aspects, including determining the features impacting credit risk and predicting the possibility of default or payment failure [10]. An intelligent processing system is needed to assist banks in selecting prospective customers who will be given loans.

II. Related Works

This research is related to previous studies. The following are summaries of previous research that are relevant to this study. Table I presents a summary of the related works.

TABLE I. SUMMARY OF RELATED WORKS

The Authors	Methodology	Case
Farida Gultom dan Tober Simanjuntak	Algoritma naïve bayes dan K-Nearest Network	Results with a very low level of accuracy on the success of bank credit payments. Tests were conduct by combining the Naïve Bayes and Algorithms of kNN [11].
Xiaxia Niu, Jun Wu, Li Shi, Xiaodong Cui, Liping Yang, Yuanyuan Li, Sang-Bing Tsai, and Yunbo Zhang	Model of Recency Frecuency Monetary and Algorithm of K -Means++, Metode PCA	The method of PCA was used to determine the indicator RFM weight. Customers were classified based on buying behavior into several groups [12].
A. Neyaa, A. Umamakeswari, A. Joy Christy, and L. Priyatharsini.	Model of RFM Analisys, K-Fuzzy C-Means Clustering, and RM K-Means Clustering	The research Propose a new method to select the initial centroid for algorithm K-means and to apply the method to segment the customers with reduced time and iterations [13].
Laxmiputra Salokhe, Saraswati Jadhav, and Rahul Shirole	Model of RFM Analysis, Algorithm of K-Means Clsutering	Investigating the scope of customer value based on crossselling probability current value and customer loyalty, this paper uses a neural network approach that uses a Self Organization Map (SOM) to form clusters for banking [14].
Guangshu Xu, Yuanyuan Li , Jun Wu, Li Shi, Sang-Bing Tsai, Wen-Pin Lin, and Liping Yang,	RFM Model was improved and combine with Algorithm of K-Means.	Quantitative analysis method to make segmentation and platform clusters. Segmenting customers with clear values and purchasing preferences greatly helps platform for effectively allocating marketing resources to specific customer groups and for building healthy long term relationships with customers [15].

III. RESEARCH METHODOLOGY

Clustering can be used to divide all customers into several clusters based on various criteria that are similar to customers. Clustering is employed to group data naturally based on the similarity of data objects and to reduce their similarity with other clusters. Unlike classification, clustering is unsupervised learning and does not require a data training stage. In this study, K-Means and decision tree are used as data mining approaches. The K-Means algorithm can group bank customers based on their similarity in credit payment statements [16].

The method consists of several stages that need to be considered, as shown in Fig. 1.

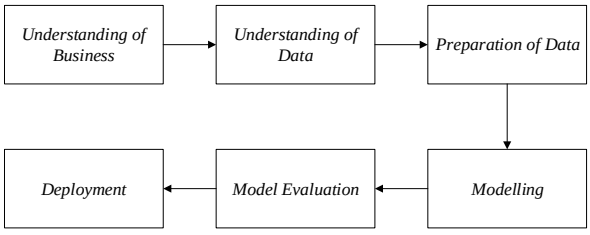

Fig. 1. The cross-industry standard process (CRISP).

This research suggests customer segmentation using the RFM Model and K-Means algorithm. Many studies have used the K-Means algorithm for the segmentation of customers [17] and traffic management systems [18]. The segmentation of customers and profiling of Recency, Frequency, and Monetary (RFM) method can help understand customer loyalty [19].

Furthermore, customers are assigned to three different variables: Recency, Frequency, and Monetary (RFM). By calculating scores for each instance, an assessment is done where the score ranges from 1 to 5, indicating the lowest and highest variable scores [13].

The evaluation of segmentation will be calculated using the Calinski-Harabasz index. The author in [9] pointed out that the Calinski-Harabasz index is the most reasonable metric to measure how well the number of groups formed by K-Means. Meanwhile, the Elbow method only calculates the error between data points X and centroid C, which will produce sum squared errors (SSE).

1) To calculate CH, the first step is to compute the inter-cluster dispersion or the between-group sum of squares (BGSS). In CH, the inter-cluster dispersion measures the weighted sum of squared distances between the centroid of the cluster and the centroid of the entire dataset (barycenter).

$$\text{BGSS} = \sum_{k=1}^{K} N_k \; x \; || \; C_k - \; C \; ||^2 \tag{1}$$

Where, Nk is the number of observations in cluster k, Ck is the center of mass of cluster k, C is the centroid of the barycenter, and K is the number of clusters.

2) The second step involves calculating the intra-cluster dispersion or within-group sum of squares (WGSS). In CH, intra-cluster dispersion measures the sum of the squares of the distances between each observation and the centroid of the same cluster.

$$WGSS_k = \sum_{i=1}^{Nk} \; || \; X_{ik} - \; C_k \; ||^2 \tag{2}$$

Where, Xik is the observation of the-i in cluster k. Then, add up all the individuals in the group and calculate the sum of squares.

$$\text{WGSS} = \sum_{k=1}^{K} \; WGSS_k \tag{3}$$

3) The Calinski-Harabasz index is calculated by summing the inter-cluster dispersions and the intra-cluster dispersions for all clusters.

$$\text{CH} = \frac{BGSS}{WGSS} \; x \; \frac{N-K}{K-1} \tag{4}$$

N is the number of observations and K is the number of clusters formed. From the equation above, it can be concluded that the greater the value of the Calinski-Harabasz index, the better the grouping is made.

To explain the results of customer segmentation, we can find the score of value and customer type, which can be used as a strategy by the company. For example, this information can be used for marketing or collecting strategies to increase the company's profits, as discussed in [17].

The achievement of the research is the application of customer segmentation in industries related to the banking sector, using the K-Means algorithm based on the score of RFM Credit Payment. The performance of the clustering methods in customer segmentation is shown in Fig. 2.

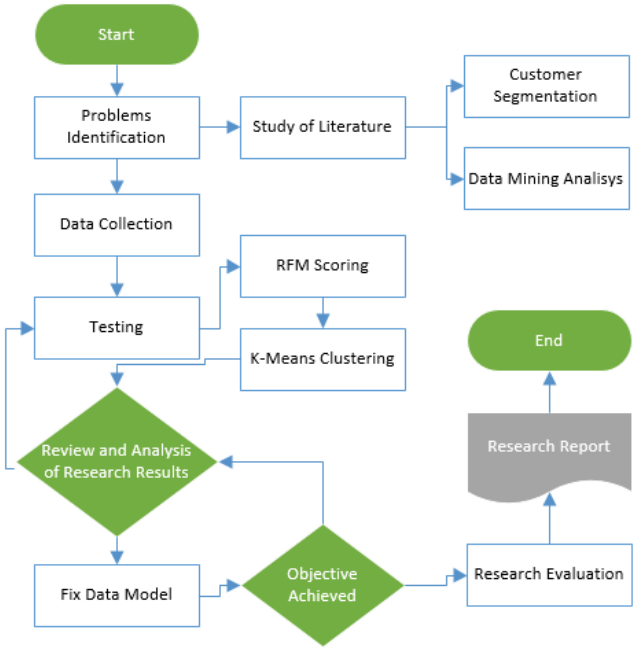

Fig. 2. Research methods.

There are steps to identify variables with good scores that can create better customer segmentation. Each stage of this process uses the Fig. 1 Cross-Industry Standard Process (CRISP), which includes the following details:

4) Preprocessing: This stage is the initial stage in the data processing, and it combines data understanding and data preparation. It contains two stages:

a) Data Cleansing. This stage is purposed to delete or eliminate variables that are not needed or cannot be used in the grouping process, such as variable names, address details, and dates.

b) Merging and *Generating Variable.* After getting good variables, we can connect each data point and change the variable according to the RFM concept. We take variables that are closely related to the last date, how many transactions, and the accumulated nominal transactions.

- This stage has several steps carried out to produce variables that are feasible for the clustering process. They are:

a) Identification of the quality of each variable using summary statistics. Variables that are not qualified, such as lots of empty data, almost all data values being the same, or the data not being varied, will be removed because they don't significantly affect clustering.

b) Handling empty data on variables. Research shows that using the median or middle value gives better results than using the mean and k-NN imputation. This research will change each variable's blank value with the median of each variable.

- Modelling: This stage changes the original value of the RFM variable that is formed into a score according to the concept of the RFM model. The original value changes to 5 bins or 5 score groups. The approach to dividing the original value is based on the frequency, so that an almost equal bin distribution can be formed.
- Modelling Evaluation. The RFM score variable's results stage will be grouped using K-Means. Calinski-Harabasz (CH) will be used to obtain the most optimal number of clusters. Besides that, the results of plotting all the variables on the clusters formed will also be seen. The results will be tested on test data.
- Deployment. This final stage interprets the clusters that have been formed and implements the actual customer collectibility process.

IV. Result and Discussion

A. Data Collection

The provided data pertains to credit, savings summaries, and transactions. The customer's personal data is deleted to maintain confidentiality, ensuring that the data processing focuses on transaction variables that will classify the customer's transaction behavior. The data processing process, from start to finish, employs R programming. Table II displays the total observations and variables for each data point.

TABLE II. Data Source Research

Data Test	Description	Data Load
Credit	Credit Data	10312 obs. of 24 variables
Saving	Saving Account Data	12373 obs. of 9 variables
Transaction	Transaction / Bank Statement Data	1048575 obs. of 6 variables

B. Model Development

This stage divides into four sub-chapters, they are pre-processing, feature engineering, modeling, and cluster interpretation. Generally, the pre-processing stage involves producing feasible variables for processing in the feature engineering stage. Feature engineering, on the other hand, involves transforming original variables into new ones based on the RFM concept. Meanwhile, the modeling stage involves grouping the previous results using K-Means. At the end of this stage, a detailed description of the formed clusters along with their implementation for collectability will be provided.

1) Pre-processing: This stage focuses on eliminating unqualified variables from the modeling process. Each data point will be cleaned, and this stage will remove any variables that cannot be used in the clustering process. Table III shows a list of the savings variables that were deleted because they could not be used.

TABLE III. LIST OF REMOVED VARIABLE SAVING

LIST OF REMOVED VARIABLE SAVINGS				
BRCOD	LOKASI	PDID	AOID	SGBU
SGID	AMBLK	CCCOD	AUTID	CSNO
WARS2	WARS3	TGLOPN	KERJA	SPCL
CTAX	CDATE	AINTR	ATAX	ADATE
INPDAT	INPUID	MNTDAT	MNTUID	AUTDAT
BCCOD	NAMA	ALM1	KPOST	WARS1
INTSPD	DRLIM	SEQNO	SWCIN	CINTR
CONVSW	CONVDT	ACCOLD	ACCNEW	ACSTA
YYMMDD	SEQBAL			

Analyzing this variable produces several important variables that can be used for modeling, as presented in Table IV: "Used Variables in Savings Data".

TABLE IV. USED VARIABLE AFTER DATA CLEANING SAVING

Variable	Description
ACNO	Account Number
LSTRN	Updated Date
MVCRED	Daily Credit Total
MVDBTD	Daily Debet Total
MVCREM	Montly Credit Total
MVDBTM	Montly Debet Total
BLCUR	Balance
BQCUR	Balance For Foreign Exchange
BLPRV	Previous Daily Balance
AVBLM	Previous Monthly Balance
CSSTA	Customer Status

Each data point will undergo data cleaning to remove transaction variables that cannot be used in the clustering process. The list of deleted variables is shown in Table V. The process of analyzing the remaining variables yields several important variables for modeling, which are presented in Table VI as the used variables in transactions data.

TABLE V. LIST OF REMOVED VARIABLE DATA TRANSACTION

LIST OF REMOVED VARIABLE TRANSACTIONS		
TRIDT	TRCON	POST
MODULE	TRCHAR	BRCOD
DPCOD	BTCOD	TRNO
SGBUT	SGIDT	GLIND
PDID	DPCODC	MMDDC
BTCODC	TRNOC	TRNOB
ACNOB	TRDAT	VLDAT
CQDATT	CQNUM	BKMNETT
BKCOD	CCCODT	RTX
AQTR	AMOD	AMPRK
TRDES2	TRDES3	INDAT
INTIM	INTRM	INUID
MNDAT	MNTIM	MNTRM
MNUID	AUDAT	AUTIM
AUTRM	AUUID	RLDAT
RLTIM	RLTRM	RLUID
ULOC	UBRANT	ULOCT
UBRANC	ULOCC	MODULC
ACNOO	SGBUC	SGIDC
BRCODB	MMDDB	BTCODB

TABLE VI. USED VARIABLES AFTER DATA CLEANING TRANSACTION

Variable	Description
TRST	Transaction Status
ACNOT	Debet Account
AMTR	Transaction Amount
MMDD	Date Transaction
ACNOC	Credit Account
TRDES1	Transaction Description

Similarly, each data point will also undergo data cleaning to remove loans variables that cannot be used in the clustering process. The list of deleted variables is shown in Table VII. The process of analyzing the remaining variables yields several important variables for modeling, which are presented in Table VIII as the used variables in loans data.

TABLE VII. LIST OF REMOVED VARIABLES AFTER DATA CLEANING TRANSACTION

LIST OF REMOVED VARIABLE LOANS				
ACNO	NOREG	SGBULX	SGIDLX	CRSEG
AOIDLX	CSNO	PDID	GLB	CRNAME
CRBRN	CRLOK	CCCOD	ACSTA	ACCLS
TARGET	HSLBNG	BLACM	BLYEAR	BGYEAR
BLOLD	BLOLDN	PARMTR	AKDDAT	DRPDAT
LNSDAT	GPLX	MTDBNG	RATBIR	SPREAD
RATEFF	RATKON	CODLK	STSBNG	REVIEW
DLRCD	AKDNO	CRTNGD	BLNPOK	BLNBNG
BLNDND	GLNUM	CRCOUN	CPRNO	SWRK1
SWRK2	RKNO2	CODRK1	CODRK2	YYMMDD
CTB08	CTB09	INDAT	INUID	MNDAT
MNUID	AUDAT	AUUID	BISFT	BIJNS
BIDEB	BIKRD	BISEK	BISBNG	BILOK
BILOK2	BIPJM1	BIPJM2	BIBNG	NOSRT
SPKE	TGLSP	OLDCOLL	CRTRS	RPROV
JWPROV	TGPROV	CRPROV	TGLUCOL	BATUCOL
KETUCOL	SKETUCOL	TGLURAT	BATURAT	CADANG

TABLE VIII. USED VARIABLES IN LOANS DATA

Variable	Description
CRLIMA	Loan Limit First
CRLIM	Loan Limit
BLCUR	Current Oustanding
MVDEBT	Debet Movement
MVCRED	Credit Movement
MASA	Period Loans (Month)
RATDND	Penalty Rate
CRTNGA	Installment arrears
CRTNGB	Interest arrears
DESCON	End of Month Balanced
JAMIN	Guarantee amount
BICOLL	Collectability
OSATHN	Oustanding early years
CLATHN	Early year limit
PKATHN	Main Balanced early years
BGATHN	Interest early years
DNATHN	Penalty early years
RTATHN	*Rate* Early Years
PBLATHN	Main month number
BBLATHN	Interest month number
DBLATHN	Penalty month number

The second stage in pre-processing involves merging the data points and generating new variables based on the RFM concept. The loans data is the main data for segmentation in the bill collectability process, and it will be combined with savings and transaction data.

The transaction data is divided into two categories: recipient (debit) and sender (credit). This division is used to gain more insight into the transaction behavior of credit customers. The date-related variables are converted into days, with December 31, 2021, serving as the reference point.

For example, the MMDD variable is converted into days, where smaller values indicate customers who are new to transactions and vice versa. This processing transforms the original variables into variables according to the RFM concept, providing information about novelty, frequency, and total purchases or transactions. Unlike ordinary RFM, this study produces multiple RFM variables to classify credit customer behavior. The results of RFM variable processing are shown in Table IX.

TABLE IX. USED VARIABLES AFTER DATA CLEANING TRANSACTION

Variable RFM
recency_savings
monetery_daily_debet
monetery_monthly_credit
monetery_monthly_debet
monetery_current_balance
monetery_current_balance_equi
monetery_prev_daily_balance
monetery_prev_monthly_balance
monetery_first_credit_limit
monetery_credit_limit
monetery_current_outstanding
monetery_debet_mutation
monetery_credit_mutation
monetery_bad_installment
monetery_interest_arrear
monetery_balance_last_month
monetery_collateral_amount
monetery_colectibility
monetery_beginning_year
monetery_limit_first_year
monetery_principal_first_year
monetery_interest_first_year
monetery_penalty_first_year
monetery_rate_first_year
monetery_number_ofMonth_principal
monetery_number_ofMonth_interest
monetery_number_ofMonth_penalty
frequency_credit
recency_sender_transaction
frequency_sender_transaction
monetery_sender_transaction
recency_sender_transaction_pending
frequency_sender_transaction_pending
monetery_sender_transaction_pending
recency_sender_transaction_success
frequency_sender_transaction_success
monetery_sender_transaction_success
recency_sender_transaction_reverse
frequency_sender_transaction_reverse
monetery_sender_transaction_reverse
recency_sender_transaction_desc_taspen
frequency_sender_transaction_desc_taspen
monetery_sender_transaction_desc_taspen
recency_sender_transaction_desc_interest
frequency_sender_transaction_desc_interest
monetery_sender_transaction_desc_interest
recency_receiver_transaction
frequency_receiver_transaction
monetery_receiver_transaction
recency_receiver_transaction_pending
frequency_receiver_transaction_pending
monetery_receiver_transaction_pending
recency_receiver_transaction_success
frequency_receiver_transaction_success
monetery_receiver_transaction_success
recency_receiver_transaction_reverse
frequency_receiver_transaction_reverse
monetery_receiver_transaction_reverse
recency_receiver_transaction_desc_taspen
frequency_receiver_transaction_desc_taspen
monetery_receiver_transaction_desc_taspen
recency_receiver_transaction_desc_interest
frequency_receiver_transaction_desc_interest
monetery_receiver_transaction_desc_interest

2) Feature engineering: This stage focuses on engineering variables that will be used for clustering effectively. Feature engineering involves several stages including identifying the quality of each variable, handling empty data, and changing the original value to the RFM score variable. Table IX presents 64 variables resulting from pre-processing, which help identify the quality of the variables. Additionally, Table X provides a summary of statistics for each variable.

TABLE X. RESULT OF SUMMARY STATISTICS FOR EACH VARIABLE

Variable RFM	Type	Total NA	Mode Value	Total Mode	Total Unique Value	Presentase NA	Presentase Mode	Presentase Value
recency_savings	double	0	30	1016	153	0%	16%	84%
monetery_daily_debet	double	0	0	6432	37	0%	99%	1%
monetery_monthly_credit	double	0	1219700	218	3729	0%	3%	97%
monetery_monthly_debet	double	0	0	26	5917	0%	0%	100%
monetery_current_balance	double	0	0	26	6413	0%	1%	99%
monetery_current_balance_equi	double	0	0	6461	1	0%	100%	0%
monetery_prev_daily_balance	double	0	20000	7	6448	0%	0%	100%
monetery_prev_monthly_balance	double	0	0	6461	1	0%	100%	0%
monetery_first_credit_limit	double	0	0	3237	2012	0%	50%	50%
monetery_credit_limit	double	0	0	3470	2957	0%	54%	46%
monetery_current_outstanding	double	0	0	3473	2954	0%	54%	46%
monetery_debet_mutation	double	0	0	6461	1	0%	100%	0%
monetery_credit_mutation	double	0	0	6451	11	0%	100%	0%
monetery_bad_installment	double	0	0	5985	435	0%	93%	7%
monetery_interest_arrear	double	0	0	6410	52	0%	99%	1%
monetery_balance_last_month	double	0	0	6380	81	0%	99%	1%
monetery_collateral_amount	double	0	0	6338	13	0%	98%	2%
monetery_colectibility	double	0	0	3238	23	0%	50%	50%
monetery_beginning_year	double	0	0	3293	3160	0%	51%	49%
monetery_limit_first_year	double	0	0	6461	1	0%	100%	0%
monetery_principal_first_year	double	0	0	6388	74	0%	99%	1%
monetery_interest_first_year	double	0	0	6399	63	0%	99%	1%
monetery_penalty_first_year	double	0	0	6428	34	0%	99%	1%
monetery_rate_first_year	double	0	0	6461	1	0%	100%	0%
monetery_number_ofMonth_principal	double	0	0	6388	13	0%	99%	1%
monetery_number_ofMonth_interest	double	0	0	6399	15	0%	99%	1%
monetery_number_ofMonth_penalty	double	0	0	6428	9	0%	99%	1%
frequency_credit	integer	0	0	3231	15	0%	50%	50%
recency_sender_transaction	double	0	30	4120	8	0%	64%	36%
frequency_sender_transaction	integer	0	1	5384	25	0%	83%	17%
monetery_sender_transaction	double	0	1219700	229	2959	0%	4%	96%
recency_sender_transaction_pending	double	6547	25	2	4	100%	0%	0%
frequency_sender_transaction_pending	integer	6547	1	4	2	100%	0%	0%
monetery_sender_transaction_pending	double	6547	300000	1	5	100%	0%	0%
recency_sender_transaction_success	double	0	30	4120	8	0%	64%	36%
frequency_sender_transaction_success	integer	0	1	5387	25	0%	83%	17%
monetery_sender_transaction_success	double	0	1219700	229	2959	0%	4%	96%
recency_sender_transaction_reverse	double	6455	23	2	5	100%	0%	0%
frequency_sender_transaction_reverse	integer	6455	2	6	2	100%	0%	0%
monetery_sender_transaction_reverse	double	6455	0	6	2	100%	0%	0%

recency_sender_transaction_desc_taspen	double	3399	30	3047	6	53%	47%	0%
frequency_sender_transaction_desc_taspen	integer	3399	1	3015	3	53%	47%	1%
monetery_sender_transaction_desc_taspen	double	3399	1219700	220	1011	53%	3%	44%
recency_sender_transaction_desc_interest	double	6077	23	68	9	94%	1%	5%
frequency_sender_transaction_desc_interest	integer	6077	1	308	8	94%	5%	1%
monetery_sender_transaction_desc_interest	double	6077	22602	19	265	94%	0%	6%
recency_receiver_transaction	double	0	30	2973	8	0%	31%	69%
frequency_receiver_transaction	integer	0	1	2471	40	0%	38%	62%
monetery_receiver_transaction	double	0	1000000	137	3947	0%	2%	98%
recency_receiver_transaction_pending	double	6454	29	3	5	100%	0%	0%
frequency_receiver_transaction_pending	integer	6454	1	7	2	100%	0%	0%
monetery_receiver_transaction_pending	double	6454	4000000	1	8	100%	0%	0%
recency_receiver_transaction_success	double	3	30	1974	9	0%	31%	69%
frequency_receiver_transaction_success	integer	3	1	2483	38	0%	38%	62%
monetery_receiver_transaction_success	double	3	1000000	137	3946	0%	2%	98%
recency_receiver_transaction_reverse	double	6383	25	18	9	99%	0%	1%
frequency_receiver_transaction_reverse	integer	6383	2	64	4	99%	1%	0%
monetery_receiver_transaction_reverse	double	6383	0	78	2	99%	1%	0%
recency_receiver_transaction_desc_taspen	double	6444	30	11	5	100%	0%	0%
frequency_receiver_transaction_desc_taspen	integer	6444	1	17	2	100%	0%	0%
monetery_receiver_transaction_desc_taspen	double	6444	1500000	2	17	100%	0%	0%
recency_receiver_transaction_desc_interest	double	5111	30	444	8	79%	7%	14%
frequency_receiver_transaction_desc_interest	integer	5111	1	1225	8	79%	19%	2%
monetery_receiver_transaction_desc_interest	double	5111	1000000	48	425	79%	1%	20%

Table X displays the amount of empty data, most frequent data, and corresponding percentages for each variable. Certain variables are deemed unqualified and are therefore cleaned up using some simple logic, including the following:

- Variables with empty data greater than 70 percent are deleted as they do not provide any useful information or insights.
- Variables with a mode percentage greater than 80 percent are deleted as they do not offer significant insights for grouping. For instance, Fig. 3 shows that the RFM variable will be deleted.

```
[1] "monetery_daily_debet"                  "monetery_current_balance_equi"         "monetery_prev_monthly_balance"
[4] "monetery_debet_mutation"               "monetery_credit_mutation"              "monetery_bad_installment"
[7] "monetery_interest_arrear"              "monetery_balance_last_month"           "monetery_collateral_amount"
[10] "monetery_limit_first_year"            "monetery_principal_first_year"         "monetery_interest_first_year"
[13] "monetery_penalty_first_year"          "monetery_rate_first_year"              "monetery_number_ofMonth_principal"
[16] "monetery_number_ofMonth_interest"     "monetery_number_ofMonth_penalty"       "frequency_sender_transaction"
[19] "recency_sender_transaction_pending"   "frequency_sender_transaction_pending"  "monetery_sender_transaction_pending"
[22] "frequency_sender_transaction_success" "recency_sender_transaction_reverse"    "frequency_sender_transaction_reverse"
[25] "monetery_sender_transaction_reverse"  "recency_sender_transaction_desc_taspen" "frequency_sender_transaction_desc_taspen"
[28] "recency_sender_transaction_desc_interest" "frequency_sender_transaction_desc_interest" "monetery_sender_transaction_desc_interest"
[31] "recency_receiver_transaction_pending" "frequency_receiver_transaction_pending" "monetery_receiver_transaction_pending"
[34] "recency_receiver_transaction_reverse" "frequency_receiver_transaction_reverse" "monetery_receiver_transaction_reverse"
[37] "recency_receiver_transaction_desc_taspen" "frequency_receiver_transaction_desc_taspen" "monetery_receiver_transaction_desc_taspen"
[40] "recency_receiver_transaction_desc_interest" "frequency_receiver_transaction_desc_interest" "monetery_receiver_transaction_desc_interest"
```

Fig. 3. Removed RFM variable.

After conducting quality identification on the 21 variables, any empty data is changed using the median or middle value of each variable. Table XI lists the variables along with their median values.

TABLE XI. EMPTY VARIABLES RFM WITH THEIR MEDIAN VALUE

Variable	Median
monetery_sender_transaction_desc_taspen	3185000
recency_receiver_transaction_success	28
frequency_receiver_transaction_success	2
monetery_receiver_transaction_success	1136039

In the feature engineering process of this research, the original data is converted into RFM scores. The variable values are sorted into five groups based on the data distribution quantiles. Fig. 4 illustrates the result of dividing the interval from the recency_savings, which denotes the last date of the customer's saving data update.

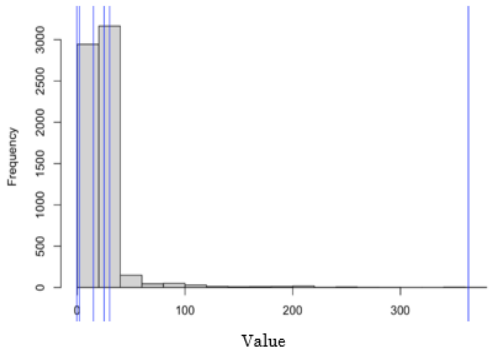

Fig. 4. Recency_savings variable bin formation interval.

The results of this bin formation will produce the RFM score variable, where the distribution of each score is shown in Fig. 5.

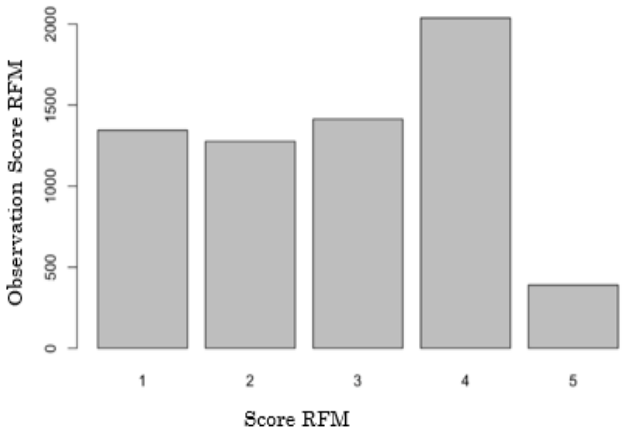

Fig. 5. Distribution of score_recency_tabungan.

The distribution of the Score_recency_saving variable indicates that the smaller the score, the more recently the customer carried out an activity related to savings. A score of 5 means that the customer has sustained long-term activities related to savings, and their distribution is quite small.

Fig. 6. Distribution of score_monetery_monthly_credit.

Fig. 6 shows the distribution of the Score_monetary_monthly_credit variable, where the difference is not significant. Fig. 7 shows the distribution of the Score_monetary_current_balanced variable.

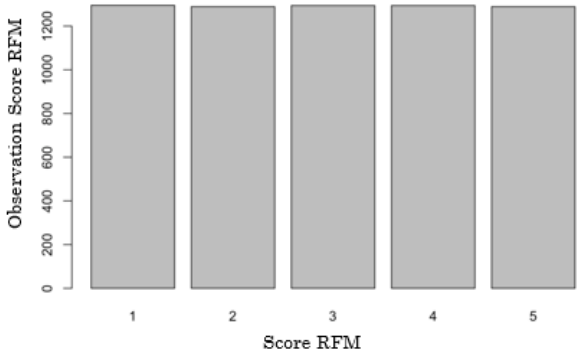

Fig. 7. Distribution of score_monetery_current_balanced.

3) Modelling: The modeling process is the final stage in the modeling process. This process involves grouping the RFM score variables using K-Means. First, the data will be randomly divided into two data sets, namely the training data and testing data, each comprising 80% and 20% of the data, respectively. The training data will be used to build a clustering model, which will be tested later using the testing data.

The first step is to remove variables that have a correlation greater than 0.7, as two or more highly correlated variables will not significantly affect the clustering process. Since there are 11 variables that have a high correlation, the clustering process uses 10 score variables.

The development stage of the K-Means model has been completed after testing the number of clusters. Testing the number of k clusters is an important step in the grouping process to obtain a number of clusters that is close to the ideal and can meet the need for customer collectibility.

The Calinski-Harabasz index is used to test how well the number of clusters is formed. Testing is done for the number of clusters ranging from 2 to 25, and the results of the Calinski-Harabasz index for each cluster are shown in Fig. 8.

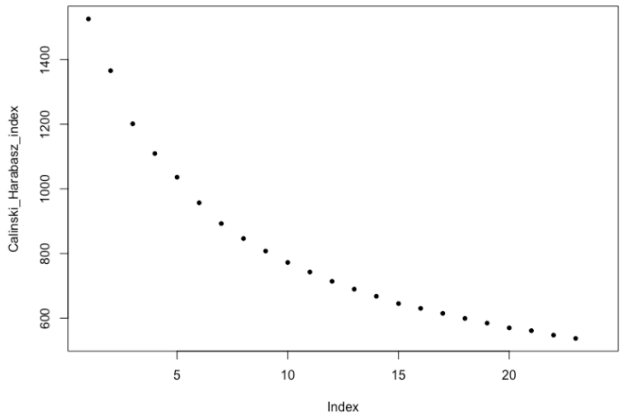

Fig. 8. Calinski-harabasz index results.

The Calinski-Harabasz (CH) Index can be used to evaluate the clustering model when ground truth labels are not known. It is used to test how well the clustering model has been created using quantities and features inherent to the dataset. The results of the Calinski-Harabasz index in Fig. 8 show that the number of clusters 2 has the highest value, and the larger the number of k clusters used, the smaller the Calinski-Harabasz index value. The CH index is a measurement of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

In addition to testing the Calinski-Harabasz index, the results of plotting all the variables on the clusters that are formed are also checked. Fig. 9 is a plot of the number of clusters ranging from 2 to 7.

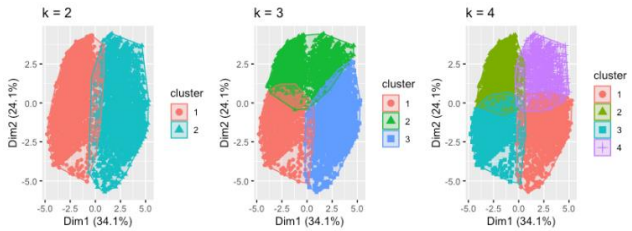

Fig. 9. Cluster plotting results k <4.

Cluster 2 shows a significant difference in the distribution of the two customer groups, indicating that it is quite good at clustering. However, there is a small number of customers who intersect each other.

The same is true for clusters 3 and 4, but they have a larger number of intersecting customers. In contrast, cluster 5 has one customer group whose characteristics are quite similar to the other groups, namely customer group 4, which is mostly similar to customer group 5.

Fig. 10 shows that as the number of clusters increases to 6 and 7, more and more customer groups have overlapping characteristics. This is in line with the results of the Calinski-Harabaszindeks index, where an increasing number of clusters results in a smaller value, indicating that the clustering results performed by K-Means are less optimal.

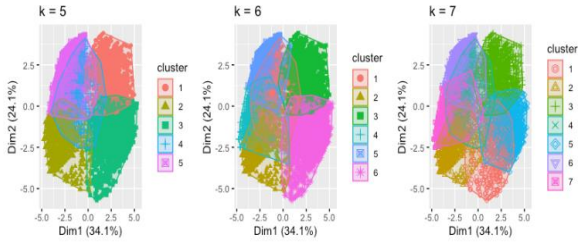

Fig. 10. Cluster plotting results k > 4.

This study chose to use four clusters because K-Means can divide customer groups quite well, with each group having different characteristics. There are some customers whose characteristics intersect with each other but are still understandable. Fig. 11 shows the result of clustering with four clusters.

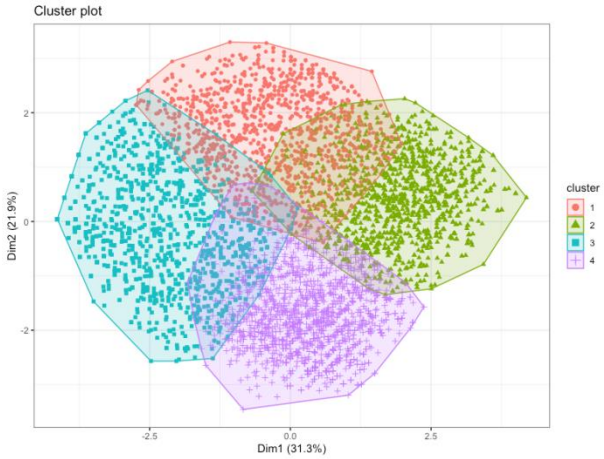

Fig. 11. Plotting results 4 clusters.

The results of clustering using four clusters show that there are only a few customer data that have intersections with other groups. Therefore, it can be concluded that four clusters are the most optimal number. Fig. 12 shows the result of grouping customers on test data.

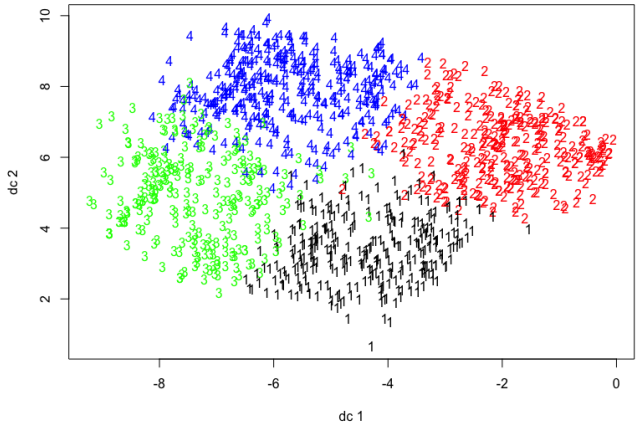

Fig. 12. Test data clustering.

This test step used training data and test data, each of which had a composition of 80% and 20%, respectively. The training data was used to build a clustering model, which was then applied to the test data. Clustering results on the test data also provide a good grouping of customers. There are clear differences between clusters, and only a few customers occur in cluster slices, especially in groups 3 and 4, which have more slices than the others.

4) Clustering result interpretation: Fig. 13 shows that the smaller the score, the better it is for the customer who has the most recent activity on their savings. Cluster 3 is the group where most customers carry out updates on savings data, followed by cluster 1. Clusters 2 and 4 are customer groups whose average is not too updated on savings activity, but cluster 2 is less updated than cluster 4. This can be an important record for the collectibility process. Fig. 14 shows a plot between the score_monetary_monthly_credit variable and the clustering results.

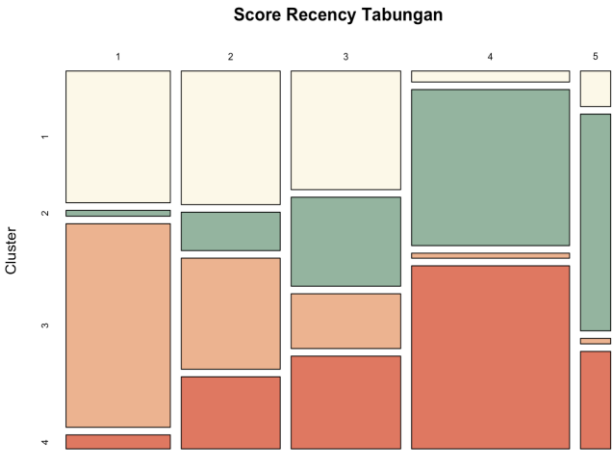

Fig. 13. Ploting variabel score_recency_tabungan and cluster.

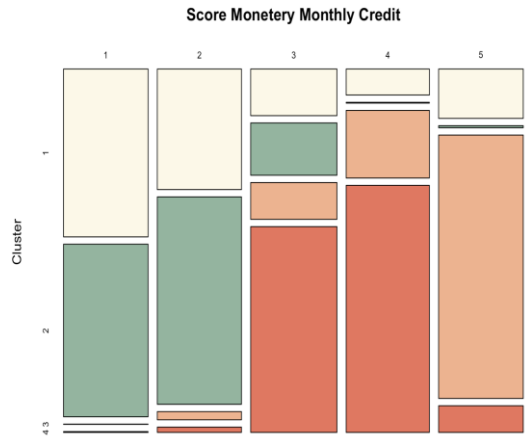

Fig. 14. Ploting variable score_monetery_monthly_credit and cluster.

This variable measures the average amount of money that comes out of the bank account of a customer each month. The higher the variable score, the higher the amount of money coming out of the customer's account. Cluster 3 is the group of customers who withdraw money from large nominal accounts, followed by cluster 4. Cluster 1 has an average score of 1 and a portion of the score 2. Cluster 2 has most customers in score 1, followed by score 2. Fig. 15 shows a plot of the variable score_monetary_current_balance and the clustering results.

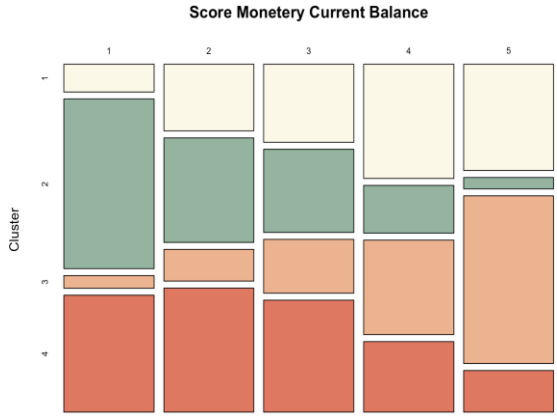

Fig. 15. Plotting variable score_monetary current balance and clusters.

The plotting results show that cluster 3 has the largest current account balance, followed by cluster 1. Meanwhile, cluster 4 has the average account balance, and cluster 2 has the smallest current balance.

The analysis that has been done can be summarized in Table XII, which ranks each variable.

TABLE XII. CONCLUSION OF CLUSTERING RESULTS ANALYSIS

Variable	Analysis
Score_recency_tabungan	The most frequent sequence of activities on savings is cluster 3, 1, 4, 2
Score_monetery_monthly_credit	The order of the most money from the account is cluster 3, 4, 1, 2
Score_monetery_current_balance	The current balance order of lots is clusters 3, 2, 1, 4
Score_frequency_credit	The order of the frequency of cash out is cluster 4, 2, 3, 1
Score_recency_sender_transaction	The sequence of most outgoing transaction updates is cluster 1, 3, 2, 4
Score_monetery_sender_transaction	The order of the most outgoing transactions is cluster 4, 3, 2, 1
Score_monetery_sender_transaction desc_taspen	The order of the most outgoing transactions for pension funds is cluster 4, 1, 3, 2
Score_recency_receiver_transaction	The most updated sequence of incoming transactions is cluster 3, 1, 2, 4
Score_frequency_receiver_transaction	The order of the most incoming transactions is cluster 3, 1, 4, 2
Score_monetery_receiver_transaction	The order of the largest nominal incoming transactions is 3, 4, 2, 1

Based on the analysis of each score variable in the table above, it can be concluded that there is no dominant cluster that always ranks first, and the results are quite dynamic. The determination of cluster priority uses a point system, with each variable carrying the same weight.

The rule is that the first rank will receive the most points, and the lower the rank, the lower the points will be. Each rank starting from the first rank will receive 5, 3, 2, and 1 point. Table XIII provides detailed results of the point calculations for each score variable in the cluster.

TABLE XIII. CALCULATION OF POINT FOR EACH SCORE CLUSTER VARIABLES

Name of Variables	Point			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Score_recency_tabungan	3	1	5	2
Score_monetery_monthly_credit	2	1	5	3
Score_monetery_current_balance	2	3	5	1
Score_frequency_credit	1	3	2	5
Score_recency_sender_transaction	5	2	3	1
Score_monetery_sender_transaction	1	2	3	5
Score_monetery_sender_transaction desc_taspen	3	1	2	5
Score_recency_receiver_transaction	3	2	5	1
Score_frequency_receiver_transaction	3	1	5	2
Score_monetery_receiver_transaction	1	2	5	3
Total Point	**24**	**18**	**40**	**28**

Based on the results of the point calculation shown in Table XIII, we can conclude the priority of cluster collectibility, which is as follows:

- Cluster 2 is a group of customers with very low scores in activity, frequency, and nominal transactions of money going out or coming in. Additionally, most customers have very low balances. Therefore, this cluster is the priority in the collectibility process due to the high potential for default.
- Cluster 1 is a group of customers with infrequent activity, the majority of whom have low balances, and the nominal amount and frequency of incoming money are also quite low. Therefore, this customer group is the second priority for the collectibility process because of its high potential for default.
- Cluster 4 is a group of customers with average savings activity, average cash-out transactions, most sufficient balances, and the largest outgoing transactions. Therefore, this group is included in the medium category, meaning that the potential for default is quite small.
- Cluster 3 has many groups of customers with new activities. Most customers have the largest balances, the most updated incoming money activity with the highest frequency, and the largest total nominal value. Therefore, this group is the one with the least potential for default.

V. Conclusion

The RFM Model can effectively form the score variable, allowing for an efficient clustering process using K-Means. The resulting cluster interpretation is easy and can provide solutions to problems. By implementing the Calinski-Harabasz index, the number of clusters used can be evaluated. This is an initial step towards determining the optimal number of clusters for the financial industry in banking data. The K-Means clustering results in well-formed groups, with significant customer grouping and no overlap between clusters. The resulting customer grouping can be useful for the financial industry in the process of collecting credit customers.

The next step in this study is to include additional variables, such as credit limits and credit card transactions, to provide payment options. The limitation of this approach is the lack of a Customer Relation Management system, which could provide a better understanding of customer perspectives, describe customer value, and improve the selling or cross-selling of various banking products and programs.

Acknowledgment

The authors acknowledge that Binus University Jakarta provided financial support and conducted a plagiarism check for this study.

References

[1] P. Indonesia, "Undang-Undang Republik Indonesia Nomor 10 Tahun 1998 Tentang Perubahan Atas Undang-Undang Nomor 7 Tahun 1992 Tentang Perbankan.," 1998.

[2] H. Y. Y. &. T. Z. Song, "How different types of financial service providers support small- and medium- enterprises under the impact of COVID-19 pandemic: from the perspective of expectancy theory," pp. Front. Bus. Res. China 14, 27, 2020.

[3] B. Indonesia, "GROUP OF BANKS AND TYPE OF LOANS," Bank Indonesia, Indonesia, 2022.

[4] A. Subrahmanyam, "Big data in finance: Evidence and challenges," Borsa Istanbul Review, pp. vol. 19, no. 4, pp. 283-287, 2019.

[5] S. M. J. Umarani, "Implementation of Data Mining Concepts in R Programming," International Journal of Trendy Research in Engineering and Technology, pp. ISSN NO 2582-0958, 2020.

[6] S. M. e. al., "Analysis for customer lifetime value categorization with RFM model," Procedia Computer Science, vol.161, pp.834-840, 2019.

[7] U. F. &. D. N. Utama, "DEVELOPMENT OF BANK'S CUSTOMER SEGMENTATION MODEL," ICIC International c 2021 ISSN 2185-2766, pp. pp. 17-26, 2021.

[8] M. N. A. R. A. M. R. Md. Zakir Hossain, "A dynamic K-means clustering for data mining," Indonesian Journal of Electrical Engineering and Computer Science, pp. Vol. 13, No. 2, February 2019, pp. 521-526 ISSN: 2502-4752, 2019.

[9] E. Schubert, "Stop using the elbow criterion for k-means and how to choose the number of clusters instead," p. arXiv preprint arXiv:2212.12189. , 2022.

[10] C. J. H. K. J. Z. W. Chang Yin, "Evaluating the credit risk of SMEs using legal judgments," Decision Support Systems, 2020.

[11] T. S. Farida Gultom, "PREDIKSI TINGKAT KELANCARAN PEMBAYARAN KREDIT BANK DENGAN MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBOR," Jurnal Manajemen Informatika & Komputerisasi Akuntansi Vol. 4 No. 2 ISSN: 2598-8565, 2020.

[12] L. S. L. Y. X. N. Y. L. X. C. S.-B. T. a. Y. Z. Jun Wu, "User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Analysis," Wireless Communications and Mobile Computing, pp. Volume 2021, Article ID 9982484, 8 pages, 2021.

[13] A. U. L. P. a. A. N. A. Joy Christy, "RFM ranking – An effective approach to customer segmentation," Journal of King Saud University - Computer and Information Sciences, 2018.

[14] L. S. a. S. J. Rahul Shirole, "Customer Segmentation using RFM Model and K-Means Clustering," International Journal of Scientific Research in Science and Technology, pp. Volume 8, Issue 3 Page Number : 591-597, 2021.

[15] L. S. W.-P. L. S.-B. T. Y. L. L. Y. a. G. X. Jun Wu, "An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm," Mathematical Problems in Engineering, pp. Volume 2020, Article ID 8884227, 7 pages, 2020.

[16] A. a. B. K. Boyaci, "Data mining application in banking sector with clustering and classification methods," Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management, Dubai, United Arab Emirates (UAE), 2015.

[17] Dedi et al., "Customer segmentation based on RFM value using k-means algorithm," Proc. of International Conference on Informatics and Computing (ICIC), Semarang, pp. pp.1-14, 2020.

[18] S. R. M. a. A. S. Nawrin, "Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System," International Journal of Advanced Computer Science and Applications, pp. Vol. 8, No. 3, 2017.

[19] M. C. H. a. I. P. G. H. Suputra, "ustomer Segmentation Using RFM Model," Jurnal Elektronik Ilmu Komputer Udayana, pp. Vol. 8, no. 2, pp. 153-161, 2019.

[20] J. Sessa and D. Syed, "Techniques to deal with missing data," 5th international conference on electronic devices, systems and applications (ICEDSA), pp. 1-4, December 2016.