Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm

Ina Maryani [1], Dwiza Riana [2], Rachmawati Darma Astuti [3], Ahmad Ishaq[4], Sutrisno [5], Eva Argarini Pratama[6]
[1,2,3]STMIK Nusa Mandiri Jakarta, [4,5,6] Universitas Bina Sarana Informatika
ina.maryani@nusamandiri.ac.id, dwiza@nusamandiri.ac.id, rachmawati.rcd@nusamandiri.ac.id, ishaq@bsi.ac.id, sutrisno.stz@bsi.ac.id, eva.eap@bsi.ac.id

***Abstract*- Every day there is a transaction process performed by Customer. The process generates a lot of data where there are 82,648 transactions from the month of January-December 2017. This study aims to perform customer segmentation on Nine Reload Credit by utilizing data mining process based on RFM model and by using techniques Clustering. The algorithm used for cluster formation is K-Means algorithm. K-Means produces a visual cluster model with the Rapidminer 5.2 tools that represent the number of customers in each cluster by using RFM (Recency, Frequency, and Monetary) attributes. From 82,648 transactions that were then processed, based on RFM model it resulted in 102 Customers. Furthermore, we analyzed cluster by using K-Means algorithm with the result of 63 Customers in Cluster 1 and 39 Customers in Cluster 2. The result of this research can be used by company to know customer category, and then the company will know how to maintain the customer owned.**

Keywords—Data Mining; RFM Model; Cluster Analysis; Customer Segmentation; K-Means Algorithm.

I. INTRODUCTION

In today's business competition, customers are the main focus of the company to maintain its excellence. Companies must plan and use clear strategies in serving customers [1]. The company's primary focus is not on how to get new potential customers but how to sell more products to the existing customers because the cost that companies must incur to acquire new customers is much more expensive than to retain existing customers [2]. In the credit business, the data can be obtained based on historical data, so the data will increase continuously such as the transaction data from each agent. The transaction process of agents in a credit server generates abundant data in the form of profiles of transactions that the agent performs. This will happen repeatedly to the credit business. Agent transaction data cumulation will slow down the search for information on that data [3]. This data can be called as data mining. Data mining is a part of knowledge discovery data which is an information extraction process that is useful, not known before, and hidden from data [4]. Based on the number of available agent transaction data, the unknown or hidden information can be known by processing the data so that it is useful for the credit business agent [4], for example in which information on the grouping of agent data has the potential to give the most profit to the company which will help companies to make decisions in product marketing. The model used by the researcher is RFM (Recency, Frequency, Monetary) commonly used to perform the last visit time grouping, visit frequency, and revenue obtained by the company [5]. The reason why continuing to use the RFM model is that it is easy to use and quickly implemented in companies, and in addition RFM is easily understood by managers and marketing decision makers [6].
The results of this study can be used as a decision support system in the credit business to map customers and to know potential customers.

II. LITERATURE REVIEW OF RFM MODEL

Some previous studies used RFM to analyze sales data as performed by [8] where in the research, online sales (e-commerce) was analyzed so that it obtained the results into 8 clusters. From the whole cluster, cluster 7 is the cluster with the highest RFM value compared to other clusters. What was performed by [7] provides information for e-commerce entrepreneurs, so they can know from each category of customer. Then [8] also used RFM to know customer value at airlines customer. From the result of the research, there are 4 customer categories that demand company to give different service to customer.

Furthermore the study [1] also used RFM to process the transaction data of exhaust sales which were then clustered to categorize the customer type of the company.

RFM technique is based on three simple customer attributes, namely Recency of purchase, Frequency of purchase, and Monetary value of purchase. The purpose of RFM is to predict future consumer behavior (directing better segmentation decisions) [9]. Therefore, it is necessary to translate consumer behavior in “number” so that it can be used all the time. In this case the researcher intended to do the test by using RFM Variable on the dataset of credit sale

transaction where the amount of the data is very much. Every month, there are thousands of transactions. The total number of transactions for a year is 82,648 times collected from January-December 2017. After the data is mapped by using RFM variable, it will be combined with K-Means algorithm to categorize from each customer so that from the process the company will be able to know the category of each customer.

III. Review Of Cluster Analysis

Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases. Data mining is a part of knowledge discovery data which is a useful, unknown, and hidden information extraction process from data [4].

Data mining aims to obtain a relationship or pattern that may provide useful indications [10]. The relationship sought by data mining is a relationship between two or more in one dimension [10].

This research using K-means to grouping data transaction with consideration, such as:

1. Could not specified the number of manual data cluster.
2. Unknown a cluster central point of data.
3. Difficult to grouping the customer types with the amount of data 82.648

Besides K-means also having an axcess, such as :

1. Easy to be implemented and used.
2. Takes the fairly quickly time to execute this learning
3. Easy to adapted.
4. Commonly used.

The K-Means algorithm is a distance-based clustering method that partitions data to a number of groups and works on numeric attributes [11].

Here are the steps to calculate K-Mean Algorithm [12]:

a. Determine the number of k-clusters to be formed.
b. Generate k-centroid (cluster center point) randomly.
c. Calculate the distance of each data to each centroid. The formula used is Euclidean distance with the equation (1) as follows:

$$D(xi, \pi i) = \sqrt{\sum (n\, i = 1\, xi - \pi i)2} \quad (1)$$

Where $d(xi, \mu\, i)$ is the distance between the cluster $n\, x$ with the center of cluster μ in the i-th word. xi is the i-th word weight of the cluster whose the distance will be searched for. μi is the weight of the i-th word at the center of the cluster.

d. Group the data by the closest distance between data with centroid.

IV. A Case Study

The dataset used in this case study is credit sales data on Nine Reload Credit Server. At the company there is a lot of data stacking, thousands of transactions every month. You can imagine how difficult it would be if you had to analyze the data manually one by one. The researchers tried to analyze the data as much as 82,648 customer transactions. The model proposed in determining the profitable customer is described in Figure 1 which shows the steps to determine the profitable customer.

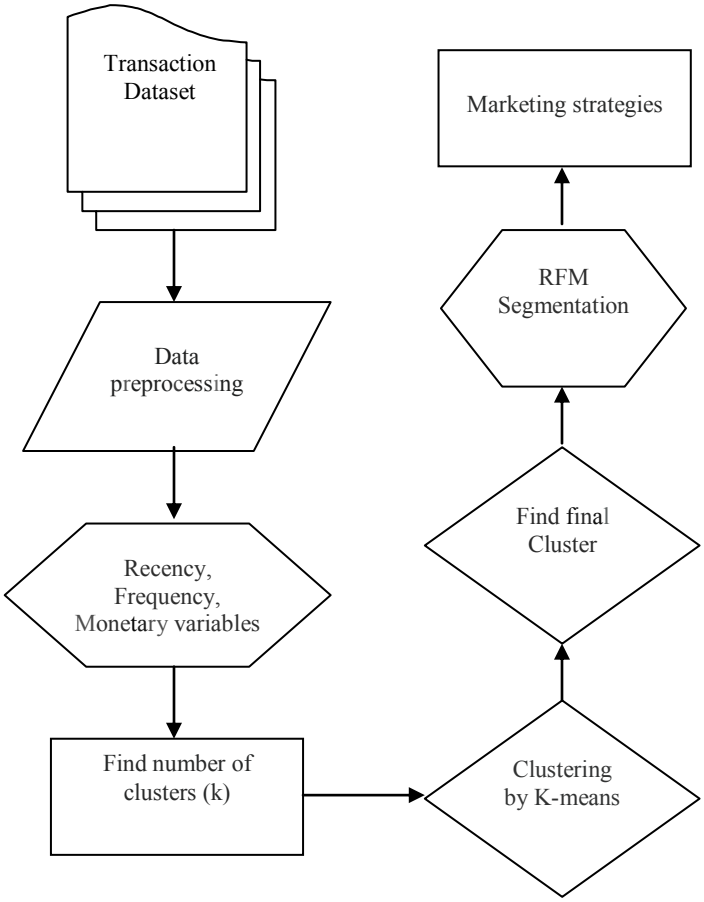

Fig 1. Framework for Customer Segmentation based on RFM model and Clustering Techniques

In this study the database used is the data collected from the transaction as much as 82,648 sales transactions. Table 1 is an example of a sales transaction database.

Table 1. TRANSACTION DATASET

Id	Date	Name	Hp	Id Member	Product	To	Status	Suplier	Price	Repeat	Des	Sn	Profit	beginning balance	ending balance	Com	Ip	price package	Type
22370	31-01-201	SAHAL CELL	6.2857E+12	SR00038	I5	8.5703E+10	sukses	SEMBILAN R	5,675	0	22-07-0	7.61E+17	225	155,671	149,996	207	127.0.0.	HARGA M	Saldo
22369	31-01-201	CAHAYA CELL	6.2857E+12	SR00110	PLN20	4.5002E+10	sukses	SEMBILAN R	20,550	0	21-43-3	1224-794	675	30,905	10,355	207	127.0.0.	HARGA M	Saldo
22368	31-01-201	TARI	6.2877E+12	SR00134	XI00	8.7879E+10	sukses	SEMBILAN R	100,150	0	21-37-0	6.65E+13	2,400	338,550	238,400	207	127.0.0.	HARGA M	Saldo
22367	31-01-201	IBU CELL	6.2877E+12	SR00008	XS	8.1809E+10	sukses	SEMBILAN R	5,700	0	21-33-1	1.7E+13	200	193,500	187,800	207	127.0.0.	HARGA M	Saldo
22366	31-01-201	ADI CELL	6.2856E+12	SR00017	I10	8.5648E+10	sukses	SEMBILAN R	10,675	0	21-17-4	7.61E+17	225	304,160	293,485	207	127.0.0.	HARGA M	Saldo
22365	31-01-201	ADI CELL	6.2856E+12	SR00017	SN20	8.2135E+10	sukses	SEMBILAN R	20,300	0	20-59-0	4.1E+13	500	324,460	304,160	207	127.0.0.	HARGA M	Saldo
22364	31-01-201	ADI CELL	6.2856E+12	SR00017	SN50	8.2135E+10	sukses	SEMBILAN R	49,800	0	20-58-1	4.1E+13	1,650	374,260	324,460	207	127.0.0.	HARGA M	Saldo
22363	31-01-201	HUYA CELL	6.2859E+12	SR00097	I5	8.5889E+10	sukses	SEMBILAN R	5,775	0	20-51-1	7.62E+17	325	232,159	226,384	207	127.0.0.	HARGA M	Saldo
22362	31-01-201	HUYA CELL	6.2859E+12	SR00097	IS10	8.5641E+10	sukses	SEMBILAN R	10,800	0	20-50-3	7.61E+17	300	242,959	232,159	207	127.0.0.	HARGA M	Saldo
22357	31-01-201	LIDA CELL	6.2856E+12	SR00039	I10	8.5871E+10	sukses	SEMBILAN R	10,775	0	20-37-3	7.61E+17	325	102,074	91,299	207	127.0.0.	HARGA M	Saldo
22356	31-01-201	LIDA CELL	6.2856E+12	SR00039	SN10	8.5327E+10	sukses	SEMBILAN R	10,700	0	20-36-3	4.1E+13	475	112,774	102,074	207	127.0.0.	HARGA M	Saldo
22353	31-01-201	INA CELL	6.2838E+12	SR00123	XI0	8.7803E+10	sukses	SEMBILAN R	10,800	0	20-26-3	1.7E+13	300	27,870	17,070	207	127.0.0.	HARGA M	Saldo
22352	31-01-201	FAIZAL CELL	262478663	SR00098	I5	8.5718E+10	sukses	SEMBILAN R	5,675	0	20-12-1	7.62E+17	225	50,825	45,150	207	127.0.0.	HARGA M	Saldo
22351	31-01-201	AJENG CELL	241838947	SR00109	I5	085726371	gagal	SEMBILAN R	5,825	0	19-55-0	7.62E+17	375	57,447	57,447	207	127.0.0.	HARGA M	Saldo
22350	31-01-201	HUYA CELL	6.2859E+12	SR00097	IS5	8.5866E+10	sukses	SEMBILAN R	5,800	0	19-54-1	7.62E+17	300	48,759	42,959	207	127.0.0.	HARGA M	Saldo
22349	31-01-201	ANES CELL	6.2877E+12	SR00107	T10	8.9538E+11	sukses	SEMBILAN R	10,300	0	19-23-2	1.31E+17	430	1,300,220	1,289,920	207	127.0.0.	HARGA M	Saldo
22348	31-01-201	HILYA CELL	6.2852E+12	SR00015	SN10	8.2345E+10	sukses	SEMBILAN R	10,600	0	19-21-0	4.1E+13	375	159,575	148,975	207	127.0.0.	HARGA M	Saldo
22347	31-01-201	TASY CELL	6.2857E+12	SR00077	SN10	8.2227E+10	sukses	SEMBILAN R	10,600	0	19-19-4	4.1E+13	375	74,827	64,227	207	127.0.0.	HARGA M	Saldo
22346	31-01-201	INA CELL	6.2838E+12	SR00123	I10	8.5817E+10	sukses	SEMBILAN R	10,775	0	19-15-4	7.61E+17	325	38,645	27,870	207	127.0.0.	HARGA M	Saldo
22345	31-01-201	AJENG CELL	241838947	SR00109	I5	8.5726E+10	sukses	SEMBILAN R	5,825	0	19-15-3	7.62E+17	375	63,272	57,447	207	127.0.0.	HARGA M	Saldo

Data Preparation

At this stage the database structure will be prepared so as to simplify the mining process. The preparation process includes three main things: selection, pre-processing, and transformation data. This process also carries out the selection of attributes that are adjusted to the data mining process. The attributes used can be seen in Table 2.

Table 2. ATTRIBUTES USED

Field	Information
Agent Name	Used to specify the customer code
Date	The date of the customer's purchase transaction is used to model Recency and Frequency. Recency, within a year when the last customer made a transaction with Nine Reload. Frequency is the number of transactions conducted by the customer within a period of one year.
Price	To model the Monetary attribute, that is by summing up all customer's transactions in one year.

The overall data available on the transaction dataset must be selected first to determine the data that can be used in accordance with the RFM variable. The total of 82,648 transactions are then selected by RFM variable to be 102 Customer. Table 3 shows the dataset in accordance with the Recency, Frequency, and Monetary variables.

Table 3. The Description of Recency, Frequency and Monetary

Agent Code	R	F	M
C001	31-12-2017	2035	Rp22,909,504.00
C002	18-06-2017	339	Rp 5,878,306.00
C003	04-11-2017	352	Rp 4,525,250.00
C004	31-12-2017	36	Rp 526,250.00
....	....	....	....
C102	25-01-2017	28	Rp 231,375.00

This study collected data in the form of sales transaction history dataset on the credit business of 82.648 transactions which prformed the determination of criteria weighting first based on recency, frequency, and monetary variable. The weighting was divided into 5 scales/ scores as listed in Table 4.

Table 4. DECISION TABLE AFTER DIGITAL

Weight	R (*Recency*)		F (*Frecuency*)		M (*Monetary*)	
5	Shortest	<1 Month	Highest	>15000	So Many	>300 Million
4	Short	1-3 Month	High	8000 - 15000	Many	150 – 200 Million
3	Reguler	3-5 Month	Reguler	5000 - 8000	Normal	100 - 150 Million
2	long	5-8 Month	low	2000 - 5000	Few	50 - 100 Million
1	longest	>8 Month	lower	<2000	fewer	<50 Million

Once the scale is determined, the next step is to transform its data on the existing scale. Table 5 shows the sample data transformed.

Table 5. EXAMPLE R-F-M VALUES OF SOME CUSTOMERS AFTER DATA PREPROCESSING

Agent Code	R	F	M
C001	5	2	1
C002	1	1	1
C003	1	1	1
C004	5	1	1
....	....	....	....
C102	1	1	1

After all transaction data is transformed into numeric form, then the data have been able to be grouped by using K-means algorithm. To be able to group these data into several clusters needs to do some steps (Rahman, 2017):

1. In this study the existing data will be grouped into four clusters.
2. In this study the initial center point was determined randomly, and it obtained the central point of each cluster which can be seen in Table 6.

Table 6. Initial Center Point

Agent Code	R	F	M
C005	5	2	1
C061	1	1	1

3. In this research k-means method was used to allocate each data into a cluster, so the data will be entered in a cluster that has the closest distance to the center point of each cluster. To find out which cluster is closest to the data, it is necessary to calculate the distance of each data with the center point of each cluster.

Table 7. CALCULATION RESULT OF EACH DATA

CUSTOMER CODE	R	F	M	C1	C2	Closest Distance
C001	5	2	1	0.985150517	3.669114335	0.985150517
C002	1	1	1	3.527989798	0.471593045	0.471593045
C003	1	1	1	3.527989798	0.471593045	0.471593045
C004	5	1	1	0.50619742	3.564042648	0.50619742
C005	5	1	1	0.50619742	3.564042648	0.50619742

4. After all the data is placed into the closest cluster, then recalculate the new cluster center based on the member average in the cluster.
5. After obtaining a new center point for each cluster, repeat the third step until the center point of each cluster is fixed, and no data moves from one cluster to another.

From the results of data processing performed, based on the customer transaction dataset using K-Means through 4 iterations in the form of clusters as shown in Figure 2, shows that the clustering results obtained 63 members of cluster 1, 39 members of cluster 2.

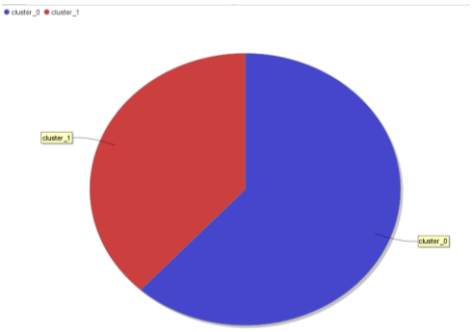

Fig 2. Graph of Cluster Analysis results

In Table 8 and Table 9, There are a number of agent names that are in Cluster 1 and Cluster 2 in which the data can be utilized by the Company.

Table 8. Customer Names in Cluster 1

NO	CUSTOMER CODE	AGENT NAME
1	C001	ADAM CELL
2	C004	ADIN TRONIK
3	C005	ADITIYA CELL
4	C006	AIS ALL CELL
5	C007	ANIDATUL CELL
6	C008	AQILA
7	C009	ARA CELL
8	C010	ASIH
9	C011	ASNEY TRONIK
10	C012	ATIKA CELL
11	C014	AYTHA CELL
12	C015	BARLI TRONIK
13	C016	BOYOUT21
14	C017	CAHAYA CELL
15	C018	DEZTI CELL
16	C019	DIA TRONIK
17	C020	EB TRONIK
18	C021	ERNI CELL
19	C022	FAIT CELL
20	C023	FITRI CELL
21	C024	FITRI POJOK CELL
22	C025	GRISELDA CELL
23	C026	HERA CELL
24	C027	HESTI CELL
25	C028	HILYA CELL
26	C029	IBU CELL
27	C030	LIA CELL
28	C031	LIDA CELL
29	C032	MUJI ASTUTI
30	C033	MUSTIKA
31	C034	NABIL CELL
32	C035	NDARI CELL
33	C036	ONDLENK CELL
34	C037	PUJI CELL
35	C038	QORY CELL
36	C039	RARA CELL
37	C041	RASITO
38	C042	RISWATI CELL
39	C043	RIZA CELL
40	C044	RIZKY CELL
41	C045	ROKHIM KOMPUTER
42	C046	SAHAL CELL
43	C047	SEMBILAN RELOAD
44	C049	SUSI TRONIK
45	C050	TARI
46	C051	SUKRON
47	C054	TOINK CELL
48	C056	UTAMA CELL
49	C059	WAHYONO CELL
50	C060	YANI CELL
51	C062	YUNITA CELL
52	C063	AJENG CELL

53	C064	ARRASYID RELOAD
54	C068	DEDERIZKY CELL
55	C069	FAIS CELL
56	C070	TASY CELL
57	C071	ADIVA CELL
58	C073	FAIZAL CELL
59	C076	FATH CELL
60	C077	LUCAS TRONIK
61	C079	LULU CELL
62	C088	UNYIEL
63	C090	YUNITA CELL

29	C092	AGUSTIN CELL
30	C093	FAKIH CELL
31	C094	TABALONG-RELOAD
32	C095	MEI-TRONIK
33	C096	KALILLA CELL
34	C097	DELTRA TRONIK
35	C098	DWI
36	C099	AJENG JKT
37	C100	AYU
38	C101	DWI CELL
39	C102	EGA CELL

Table 9. Customer Names in Cluster 2

NO	CUSTOMER CODE	AGENT NAME
1	C002	ANES CELL
2	C003	HAFI CELL
3	C013	HERI
4	C040	HUYA CELL
5	C048	IMA CELL
6	C052	JUJU CELL
7	C053	INA CELL
8	C055	JM IRS
9	C057	IBU KECE
10	C058	RAFKA RELOAD
11	C061	SAMSITI CELL
12	C065	SIMPLE PAY
13	C066	SUPRI CELL
14	C067	TAKIM
15	C072	TRIDAYA RELOAD
16	C074	ULFA CELL
17	C075	SEMBILAN CELL
18	C078	SITRIADI CELL
19	C080	SOLIH TRONIK
20	C081	TRANSZHEN
21	C082	NASYAH PULSA
22	C083	ADI CELL
23	C084	EKA CELL
24	C085	ELLA CELL
25	C086	WAHYU CELL
26	C087	INCES
27	C089	KHAYLA CELL
28	C091	MUNDRI CELL

V. Conclusion

The main purpose of this research was to segment the customers from the transaction data of 82,648 based on RFM model, and furthermore clustering analysis was performed by using K-Means.

The result of this research is 102 customers. 63 customers are in Cluster 1, and 39 customers are in Cluster 2. Cluster 1 has higher average of RFM value than Cluster 2.

By knowing the categories of each Customer, it is expected that the company will be able to take the right decision in marketing strategy.

Acknowledgment

We would like to thank Nine Reload Credit which is the business of selling credit which has provided data for us.

References

[1] Maryani, Ina, and Dwiza Riana. 2017. "Clustering and Profiling of Customers Using RFM for Customer Relationship Management Recommendations." *2017 5th International Conference on Cyber and IT Service Management, CITSM 2017*, 2–7. https://doi.org/10.1109/CITSM.2017.8089258.

[2] Tama, Bayu Adhi. 2010. "Penetapan Strategi Penjualan Menggunakan Association Rules Dalam Konteks CRM." *Jurnal Generic* Vol. 5 (No.1):35–38.

[3] Hand, David J. 2007. "Principles of Data Mining." *Drug Safety* 30 (7):621–22. https://doi.org/10.2165/00002018-200730070-00010.

[4] Ramamohan, Y, K Vasantharao, C Kalyana Chakravarti, and a S K

Ratnam. 2012. “A Study of Data Mining Tools in Knowledge Discovery Process.” *International Journal of Soft Computing and Engineering* 2 (3):191–94.

[5] Wongchinsri, Pornwatthana, and Werasak Kuratach. 2016. “A Survey -Data Mining Frameworks in Credit Card Processing.” *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016*. https://doi.org/10.1109/ECTICon.2016.7561287.

[6] Peiman Alipour Sarvari, Alp Ustundag, and Hidayet Takci. 2014. “Performance Evaluation of Different Customer Segmentation Approaches Based on RFM and Demographics Analysis.” *Kybernetes* 43 (8):1209–23. https://doi.org/10.1108/K-01-2015-0009

[7] Rachid, et al. 2015. “Combining RFM Model and Clustering Techniques for Customer Value Analysis of a Company selling online.” *2015 12th International Conference of Computer Systems and Applications (AICCSA) 2015,1-6.*

[8] Liu Jiali and Du Hyung. 2010. “Study on Airline Customer Value Evaluation Based on RFM Model (2010).” *2010 International Conference On Computer Design And Appliations (ICCDA 2010) ,278-281*

[9] Aviliani, U. Sumarwan, I. Sugema, and A. Saefuddin. 2011. “Segmentasi Nasabah Tabungan Mikro Berdasarkan Recency, Frequency, dan Monetary : Kasus Bank BRI.” *Finance and Banking Journal* 13 (1):95–109.

[10] Kusrini Luthfi, Ema Taufiq. 2009. *Algoritma Data Mining*. Edited by Theresia Ari Prabawati. Yogyakarta: C.V Andi OFFSET. https://books.google.co.id/books?id=Ojclag73O8C&pg=PA3&dq=data+mining+adalah&hl=id&sa=X&ved=0ahUKEwijrefgpYnZAhXBPY8KHWeJCQ4Q6AEIKzAA#v=onepage&q=data mining adalah&f=false.

[11] Lubis, Abdul Haris. 2016. “Model Segmentasi Pelanggan Dengan Kernel K-Means Clustering Berbasis Customer Relationship Management.” *Jurnal & Penelitian Teknik Informatika* 1:36–41.

[12] Rahman, Aulia Tegar; Wiranto ;Rini Anggrainingsih. 2017. “Coal Trade Data Clustering Using K-Means (Case Study PT . Global Bangkit Utama)” 6 (1):24–31.