Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation

Firman Pradana Rachman, Handri Santoso, Arko Djajadi
Master of Information Technology Department, Faculty of Science and Technology
Pradita University, Tangerang, Indonesia

***Abstract*—An effective marketing strategy is a method to identify the customers well. One of the methods is by performing a customer segmentation. This study provided an illustration of customer segmentation based on the RFM (Recency, Frequency, Monetary) analysis using a machine learning clustering that can be combined with customer segmentation based on demography, geography, and customer habit through data warehouse-based business intelligence. The purpose of classifying the customers based on the RFM and machine learning clustering analyses was to make a customer level. Meanwhile, customer segmentation based on demography, geography, and behavior was to classify the customers with the same characteristics. The combination of both provided a better analysis result in understanding customers. This study also showed a result that minibatch k-means was the machine learning model with the rapid performance in clustering 3-dimension data, namely recency, frequency, and monetary.**

Keywords*—Customer segmentation; machine learning; business intelligence; data warehouse

I. Introduction

Knowing customer needs is a way to win the competition in the market and increase company profits. By knowing what customers want, companies can create effective marketing strategies. Every customer has different needs and expectations, but some have similar or the same characteristics. One way to group several customers who have the same characteristics is to create customer segmentation. Customer segmentation is also the key to improving customer relationships. The process of information analysis to understand the market and customer is a part of a marketing strategy known as marketing intelligence.

Previous research used machine learning for customer segmentation. The type of machine learning used is unsupervised machine learning. One of them is using k-means [1] or using Hierarchical Clustering which is combined with PCA (Principal Component Analysis) technique [2]. Several other studies combine RFM (Recency, Frequency, Monetary) Analysis with K-means to determine customer ratings [3]. Some of these studies only make segmentation based on numerical values or predictive numbers generated by machine learning such as annual income and spending scores [1] or RFM Score [3], but do not grouping them with categorical and descriptive data. When there is a question, which city does the customer group live in with the highest annual income? So, to get the answer, we must explore the data further.

Another problem is how to make the data well integrated. Well-organized data will facilitate analysis and report generation better. The quality of the data must also be considered. Problems in data such as duplication, different formats, incomplete and dirty data are things that must be overcome for better data governance [4].

Based on that fact, a research idea emerged that utilizes machine learning and business intelligence to create customer segmentation in a data warehouse platform. The information technology advancement enables the data to be processed and analyzed better. One of the examples is machine learning and business intelligence technology utilization. The use of machine learning can display predictive data, while business intelligence displays descriptive data. The integration and the combination of both will provide knowledge for a company in making an accurate business need. Data warehouse will make the data well integrated, stored for a long time and not interfere with data in the main system or transaction operations. Data quality can be handled well through the ETL (Extract, Transform, Load) process in the data warehouse [5].

This study discussed the utilization of machine learning and business intelligence built in the data warehouse platform using SQL Server. The outcome can be analyzed by marketing through dashboard and business intelligence reports. The data used here were the credit card transaction data for three months from banking companies in Indonesia. A machine learning model that can be used is unsupervised learning known as clustering. This study also tested some machine learning models of clustering to find a model with a rapid performance.

II. Theoretical Framework

Banking is any kind of activity in banks, including organizational business activities and the process. Meanwhile, a bank is a business entity operating the business activity. The function of a bank is collecting funds and also a distributor of credit to both individuals and business entities [6]. A bank has several types of loan products, such as working capital loans, investment loans, and consumer loans. A credit card is a part of consumer credit given to an individual in the form of a card that can be used for purchasing goods and services in shops, supermarkets, restaurants, etc.

Marketing Intelligence is a process of analyzing information to understand consumers, attitudes, and market behavior for accessing changes in a business and industrial environment to support the decision-making process [7]. Marketing intelligence consists of two parts, namely marketing research and customer relationship marketing database. Marketing research focuses more on the process of marketing planning, analyzing a situation, and building a strategy, while customer relationship marketing database focuses on data processing in a database.

Customer segmentation consists of a group of customers having the same needs and wants [8]. A segmentation group can be divided into 4 parts, namely.

1) Geography. It divides a market based on the location of the domicile, for instance, country, province, and city.

2) Demography. It divides a segmentation based on age, family, income, occupation, education, religion, etc.

3) Psychography. It is a part of psychological and demographic science in understanding consumers better, such as lifestyle or the value of life.

4) Behavior. This segmentation divides customers into several groups based on their habits, knowledge, or responses to a certain product.

Customer segmentation can also be performed based on Cost to Serve, Net Price, and relationship value [9]. Net Price and Cost to Serve are types of costs that can be measured. Relationship value has qualitative characters, and the value is determined intuitively by managers. This matrix is able to provide variability in making customer segmentation.

Another method in making customer segmentation is by using the RFM (Recency, Frequency, and Monetary) analysis. The purpose is to determine the customer level based on their purchase history [10]. This method consists of 3 dimensions, namely.

1) Recency. The last time a customer does a transaction. The Recency value is calculated from the difference of total days by subtracting the last date of transaction from the date of the current process. The lower value will be better.

2) Frequency. The frequency of a customer does a transaction. The higher value will be better.

3) Monetary. The amount of money spent. It is equal to Frequency that the higher value will be better.

A data warehouse is separate data storage from the primary application operating an operational transaction process. In this process, the data are transformed into information that can be analyzed by consumers [11]. The data process starts from data integration that has been extracted from the primary operational application and is transformed and loaded into a format that is appropriate to the data structure in the data warehouse. This process is known as ETL (Extract Transform Load).

The ETL process contains data cleaning, filtering, aggregate, and types of transformation. After the data are collected orderly, the data can be used for data mining or business intelligence. Business intelligence (BI) is defined as data presentation to entrepreneurs to be used for gaining knowledge or making a business decision [11]. Business intelligence is an important part of business analytics because it produces effective analysis [12].

In making a business intelligence model, two tabulated models are generally used, namely.

1) A fact table is a table containing a transaction table consisting of numeric data that can be changed every day. For example, sales, purchase, finance, etc.

2) A dimension table is a table containing data category that generally the content rarely changes, and it will be used for data classification and aggregation contained in the fact table. For example, the Customer dimension table contains customer id, customer name, date of birth, address, and the like.

A schema in business intelligence is a group of tables consisting of dimension, fact, and attribute designed in differently according to the necessity. A schema consists of several types. They are as follows.

1) Star Schema. This schema puts the fact table in the center and the dimension table is placed around the fact table forming a star pattern. In the star schema, the dimension table with the same hierarchical data structure is placed in one table.

2) Snowflake Schema. This schema is different from the star schema model, whereby the dimension table is separated for the main category. For example, in the star schema, the data in the product table for product and product group is merged, while in the snowflake schema, both data are separated. Thus, in the end, the snowflake schema model looks like a snowflake.

Machine learning is a part of Artificial Intelligence (AI) allowing a system to learn from data rather than by explicit programming. Machine learning uses several repetitive algorithms by learning from data to improve, describe data, and predict results [13].

Unsupervised learning is a process of grouping unlabeled data. One of the utilizations is for clustering. The machine learning models for clustering are as follows.

1) Hierarchical Clustering. It works by forming a hierarchy or based on a certain level to appear like a tree structure. Thus, the clustering process is performed according to the level or step by step. Hierarchical clustering consists of two clustering, namely Agglomerative (bottom-up) and Divisive (Top-Down).

2) Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). It is an algorithm that can cluster big data by making a small and brief summary at first and storing information as much as possible. The smaller and brief summary is then grouped as a substitute for the larger data cluster. The mechanism of the first BIRCH algorithm is summarizing a group of big data into a smaller one that is known as Clustering Feature (CF) tree. Each node of this tree consists of some Clustering features (CF). Then, each node,

including a leaf node, has some CF; besides, the internal node CF has a pointer to sub-node, and all leaf nodes are linked by a doubly linked list.

3) K-means. This clustering algorithm is one of the non-hierarchical clustering methods that try to make partitions for the existing objects into one or more clusters or object groups according to the characteristics. Thus, the object with the same characteristics is grouped in one cluster and the object with different characteristics is grouped in another cluster.

4) Mini Batch K-means. This algorithm forms a minibatch consisting of a collection of small randomized data with a constant size enable to be stored in a memory. The mechanism is that the sample is taken randomly from the dataset to form a minibatch, and then it is assigned to the nearby centroid. In the second step, the centroid is updated and so on.

One of the ways to find the optimal total cluster is by using an elbow method. It is done by seeing the percentage of the comparison between the total clusters that will form an elbow in a certain point. This method can be illustrated through a line plot between SSE (Sum of Squared error) compared to the total cluster and finding a point that represents 'an elbow point" (the point after SSE or inertia starts decreasing in a linear fashion). Elbow method is often used in previous studies for determining the optimal number of clusters [14], [15] , in addition to the silhouette coefficient method [16].

III. Related Work

A study on Customer segmentation using a machine learning method was the Fuzzy C-Means Clustering utilization for Customer Relationship Management (CRM) database on an online shop, namely tokodiapers.com [17]. Subsequently, another study focused on the implementation of k-means clustering on Recency-Frequency-Monetary (RFM)-based customer segmentation [18], [19]. Customer segmentation using PCA was combined with machine learning to make clustering [20]. The combination of K-means and ANN methods used SOM [21], [22]. Some other studies compared the clustering models between k-means, fuzzy c-means, Repetitive median K-Means [10], and between k-means, k-medoids, and DBSCAN [23].

Previous studies showed that business intelligence can be used for descriptive data in marketing strategy [24], social media analysis [25], travel companies [26], and can also be implemented in small-scale companies [27]. The business intelligence implementation can be combined with the data warehouse implementation. For example, business intelligence implementation using Higher Education data in Iran [28].

Based on literature studies and previous research, this study offers a complete and different solution for customer segmentation. First, customer segmentation based on RFM analysis creates customer levels combined with Geographic, Demographic, Psychographic and Behavioral to classify customers with the same characteristics. Second, the data is presented in business intelligence reports and is based on a data warehouse. Finally, the research will test several clustering models to find the fastest model.

IV. Research Methodology

This study was conducted through several phases as shown below.

A. Understanding the Business Process

In this phase, it was done by seeking information on how customer segmentation was implemented. There were two methods. They are as follows.

1) Literature study. It was performed by reading relevant books and journals.

2) Conducting a field observation and interview with users.

Based on these processes, it can be inferred that 2 methods of customer segmentation will be implemented. First, customer segmentation was ranked based on the RFM (Recency, Frequency, Monetary) analysis [10], and the second method was customer segmentation based on Geography, Demography, Psychography, Behavior [8].

B. Analysis Data

The study used secondary data taken from the credit card transaction history in bank XYZ in Indonesia for three months, namely October to December 2020. There are five CSV (Comma Separated File) format files that will be used, namely.

1) Cc_transaction.csv is credit card transaction data whose information consists of customer id, category id, transaction date, amount in foreign currency, currency, card number, payee account and payee name.

2) Category.csv is shopping category data whose columns consist of category id, category name and group category.

3) Currency.csv contains a list of currencies consisting of the following columns currency id, currency code and currency name.

4) Customer.csv contains data from customer profiles consisting of customer id, customer name, gender, marital, grade, profession, address1, address2, postal code, open date, birthday, city, and province.

5) Rate.csv contains exchange rate information for all currencies consisting of currency id, date and rate.

C. Designing a Data Architecture and Data Flow

This phase consisted of designing tables, data architecture, and the data flow. The column structure of the created table must match the format and content of the data.

D. Preparation Process in the Data Warehouse

This phase consisted of an ETL (extract, transform, and load) process. The data in this phase were processed through cleaning, filtering, and normalization, thus, the data entering the database were neat and clean data. The data from the ETL process were imported into a staging table and then processed into a fact and dimension table for the needs of business intelligence and a table of RFM analysis for the machine learning process.

E. The Process of Machine Learning

It consisted of several stages as seen below.

1) Feature selection. The data were taken from the table of RFM analysis that was made in the preparation process in the data warehouse.

2) The production of a machine learning model. This stage consisted of several tests for machine learning clustering to find the rapid model. The models being tested here were agglomerative hierarchical clustering, Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH), K-means, and Mini Batch K-means.

3) Optimizing the total clustering. This stage was done to find the optimum total clusters using an Elbow method.

4) Implementation of clustering model. The machine learning model selected is the fastest from the previous testing process. The number of clusters is in accordance with the recommendations of the elbow method.

F. Business Intelligence Process

It consisted of several steps as follows.

1) The tables of load from the machine learning process and ETL at data warehouse.

2) Making a dimension model and the relationship between those tables.At this stage, it will be decided to use the star schema or snowflake model.

3) Design and dashboard visualization. This stage aims to design and present data in a business intelligence portal. The data displayed is a summary of the clustering results of machine learning and demography, geography, and customer habits. On the other hand, it displays detailed data from transaction history and customer profiles. History data is also available for each transaction and customer grouping. All data can be selected and filtered based on the results of machine learning clustering, year and month.

V. Result and Discussion

A. The Process in the Data Warehouse

The result from the observation and analysis was an illustration of how to design the data architecture and data flow. The following is an illustration of the data architecture and data flow.

From the Fig. 1 we can see that the first process is data processing from the server of core banking and then it is integrated into SQL server database as data warehouse through an ETL (Extract, Transform, and Load) process using SQL commands (bulk insert method). The result of the ETL process was stored in the staging table. Subsequently, two processes were done; first, the data from the staging table were processed into a fact table and dimension table through the SQL demand. Second, the data from the staging table formed an RFM _analysis table that would be used as a feature selection for a clustering process in machine learning. Processes in machine learning using Python with the Scikit-learn library. The result obtained from the machine learning process was exported into a dimension table and a fact table.

The outcome of the machine learning process and the ETL process in the data warehouse was forming a dimension table and a fact table used for a modeling process in the business intelligence. The detail of processes and data structure can be seen in Fig. 2.

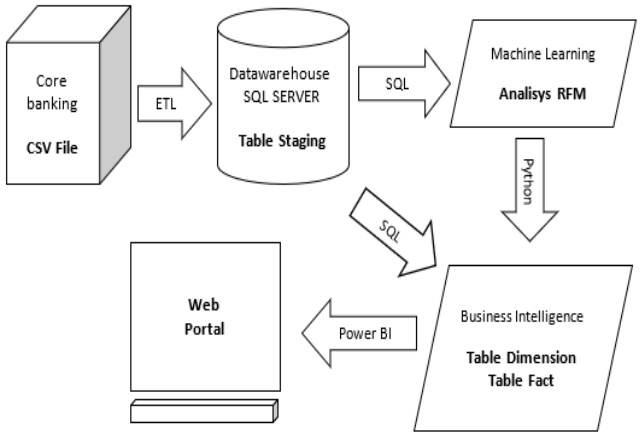


Fig. 1. The Design of Data Architecture and Data Flow.

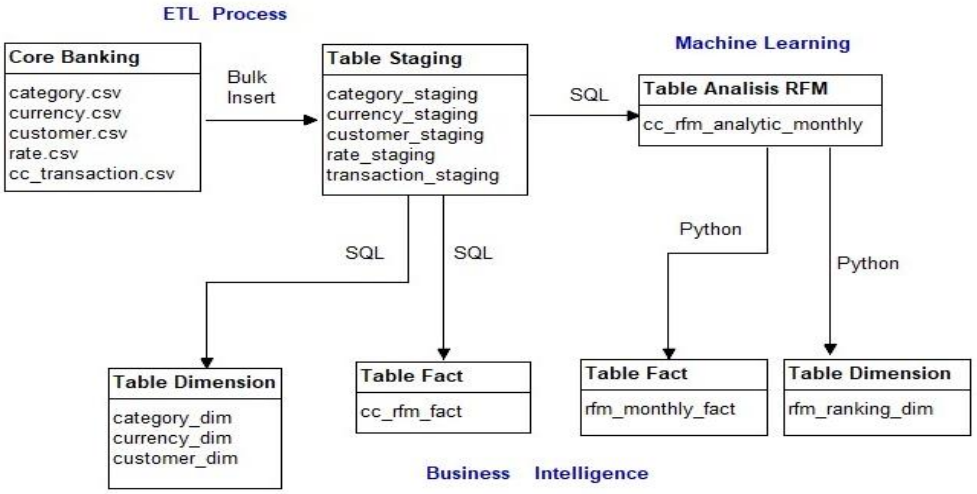


Fig. 2. Detail of Data Flow.

B. Machine Learning Process

The data used in this process was an RFM analysis table that was made in the ETL process previously. The content consisted of the summary of recency, frequency, and monetary values based on customer per month and per year as shown in Table I.

TABLE I. CC_RFM_ANALYTIC_MONTHLY TABLE

Customer id	Recency	Frequency	Monetary	Month	Year
190003214887	22	1	79000	12	2020
190000291053	1	4	845800	12	2020
190001424933	1	2	831608	12	2020
190002882666	3	1	540000	11	2020
190001940395	8	3	3317177	10	2020
190003283225	7	5	1611053	12	2020
190003229073	11	1	79000	12	2020

The next process was testing the machine learning clustering model aiming at seeking a clustering model with a rapid performance. The first trial test was conducted using the data with a range of 3 months from the staging table. Overall, the total data was 46.079 rows. The test was conducted using a trial test on 2 up to 5 clusters. The computer specification used here was Intel Core i3 6006U 2 GHz, Memory 12 GB, Hard disk SSD 512 GB, and VGA Nvidia GeForce 940MX with 2GB dedicated VRAM. The Table II illustrates the detail of the test result showing the speed of the clustering process done in a second.

TABLE II. THE SPEED DIFFERENCES BETWEEN CLUSTERING MODELS

Model	2 Clusters (second)	3 Clusters (second)	4 Clusters (second)	5 Clusters (second)
Aggromerative	Error Memory	2856	2700	2239
BIRCH	31	20.36	20.03	20.01
Kmeans	1.32	1.45	1.76	2.24
Minibatch Kmeans	1.47	1.40	1.18	1.08

From the Table II, agglomerative clustering has the longest clustering process; even when trying to perform a clustering process with 2 clusters, an error message appears stating that not enough memory. BIRCH occupies the third rank for better performance than agglomerative clustering. Meanwhile, K-means reaches a far better performance than BIRCH and agglomerative clustering. Nevertheless, in general, the performance of minibatch k-means is faster with the longest duration of 1.47 seconds for 2 Clusters.

The experimental results in Table II show that when the number of clusters increases, the clustering process becomes faster. This happens for clustering using agglomerative, birch and minibatch kmeans, but for kmeans the opposite happens. The unique thing is that when using 2 clusters, the kmeans algorithm is better than the mini batch kmeans. However, when trying 3 or more clusters, the minibatch kmeans performance is superior.

The machine learning Mini Batch K-Means model is finally chosen because it is the rapid model in performing a clustering. The next phase was selecting the optimal number of clusters from the data. The method used here was the elbow method. The result shows that the total optimal cluster in each month is 4 clusters. It is shown from the intersection of lines forming a perpendicular line as a visualization of an elbow method in Fig. 3.

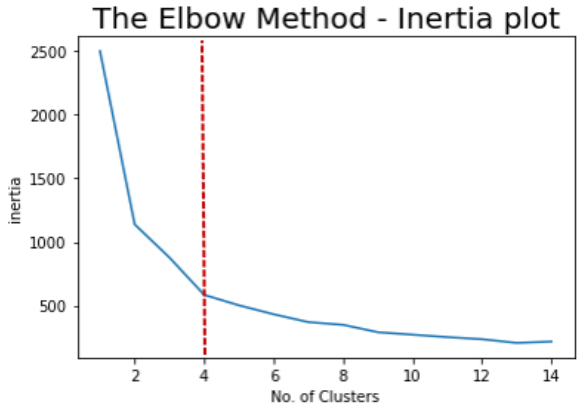


Fig. 3. A Graph of an Elbow Method in December 2020.

Before the data modeling, it was normalized using a Min-Max score. This process was aimed at making the data have the same range from 0 to 1. Therefore, data visualization in the form of a graphic becomes more precise. The formula used in this Min-Max method as shown in (1).

$$Xnew = \frac{(Xold - Xmin)}{(Xmax - Xmin)} \quad (1)$$

Xold is the former score. Xmin is the minimum score and Xmax is the maximum score in the data range.

The clustering prediction was done using minibatch k-means with a total of 4 clusters. After grouping the data using machine learning, the recency, frequency, and monetary scores of each datum were calculated. The scoring was done by ranking the data. The best one was scored 4 and the lower one was scored 3 and so on. The following is the calculation result of the RFM score for October 2020 in Table III.

TABLE III. THE CALCULATION RESULT OF RFM TABLE IN OCTOBER 2020

Cluster	R_means	R_score	F_means	F_score	M_means	M_score	Score	Segment	Label
0	0.81	1	0.01	1	0.08	3	5	113	At Rsik
1	0.46	2	0.02	2	0.08	1	5	221	Keep
2	0.16	3	0.06	3	0.37	4	11	344	Royal
3	0.13	4	0.03	4	0.08	2	9	432	Potensial

From the Table III, it can be inferred that the highest score is cluster 3 with a score of 11 points. Even though the score is the highest, cluster 3 does not have the highest recency score since the best recency score is occupied by cluster 1. Meanwhile, the lowest score is occupied by clusters 0 and 2 with the same score of 5. However, cluster 0 is considered the lowest since it has the lowest recency and frequency scores. Subsequently, each of these clusters is labeled according to their class. The highest cluster is cluster 3 labeled with Royal; the cluster below cluster 3 is cluster 1 labeled with Potential. Then, clusters 2 and 0 are labeled with Keep and At-Risk respectively.

Clusters with the Royal label are the most important and loyal customers because the number of shopping transactions is the largest and the shopping frequency is the most frequent. Meanwhile, the group of potential clusters are customers who have the potential to become loyal credit card users, because the recency value is higher even though the amount and frequency of shopping are smaller. It is possible that the group from this cluster is a new customer. Clusters with the labels At Risk and Keep must be considered, because in this group they rarely shop in large quantities. The same grouping was also done for the data in November and December 2020. The following is the 3D visualization of customer clustering in October 2020 in Fig. 4.

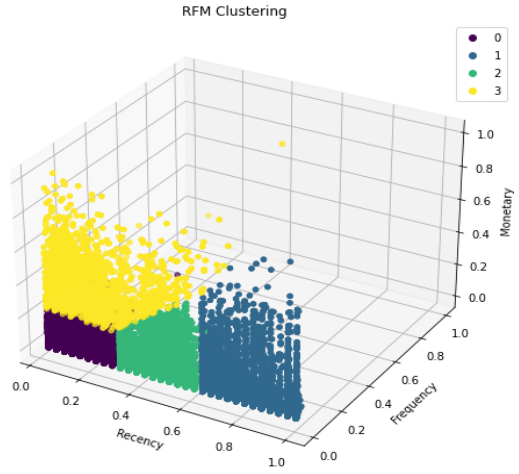


Fig. 4. 3D Visualization of Clustering for the Data in October 2020.

C. Business Intelligence Process

The table containing the result of ETL and Machine Learning processes is loaded into the business intelligence model. The modeling schema is made using a Star Schema by placing two fact tables, namely cc_rfm_fact and rfm_monhtly_fact. Meanwhile, the other dimension tables are around the two fact tables as shown in Fig. 5.

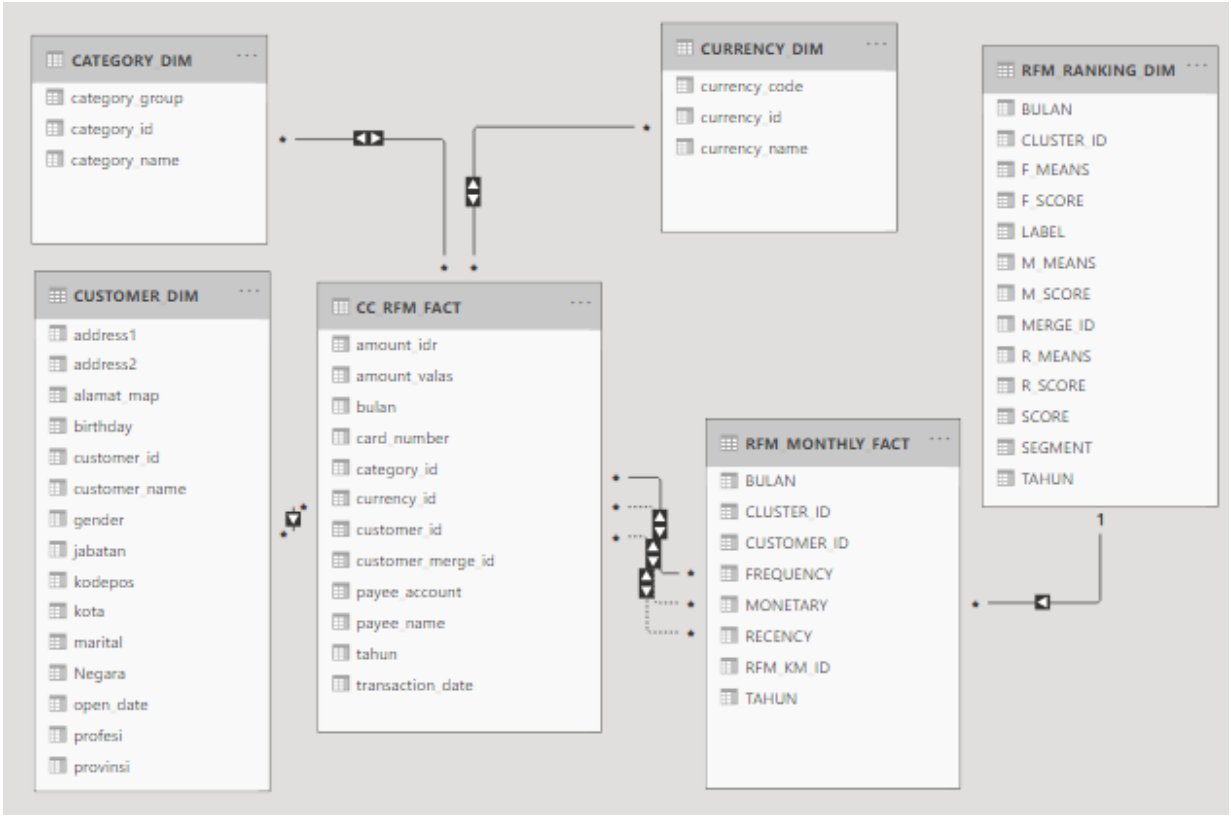


Fig. 5. Dimensional Modeling with a Star Schema.

The dashboard in the business intelligence consists of several parts as follows.

1) Main Dashboard. It displays a summary and aggregate data from all data. The top side has filtering based on year, month, and cluster choices. Besides filtering, there is a scorecard containing segmentation information of cluster, RFM score, and total customer, average recency score, average frequency score, and average monetary. score. Meanwhile, the central side has a bar graph showing data on shopping habits based on the shopping category. The demographic data, such as profession, sex, and marital status, are in the form of a pie chart or doughnut chart. On the right side, there is information about transaction amount based on the currency. The complete illustration can be seen in Fig. 6.

2) Distribution map. It shows the number of customer distribution based on the province and regency. The data are presented in the form of an Indonesian map. There is filtering on the top side according to customer segmentation, year, and month as shown in Fig. 7.

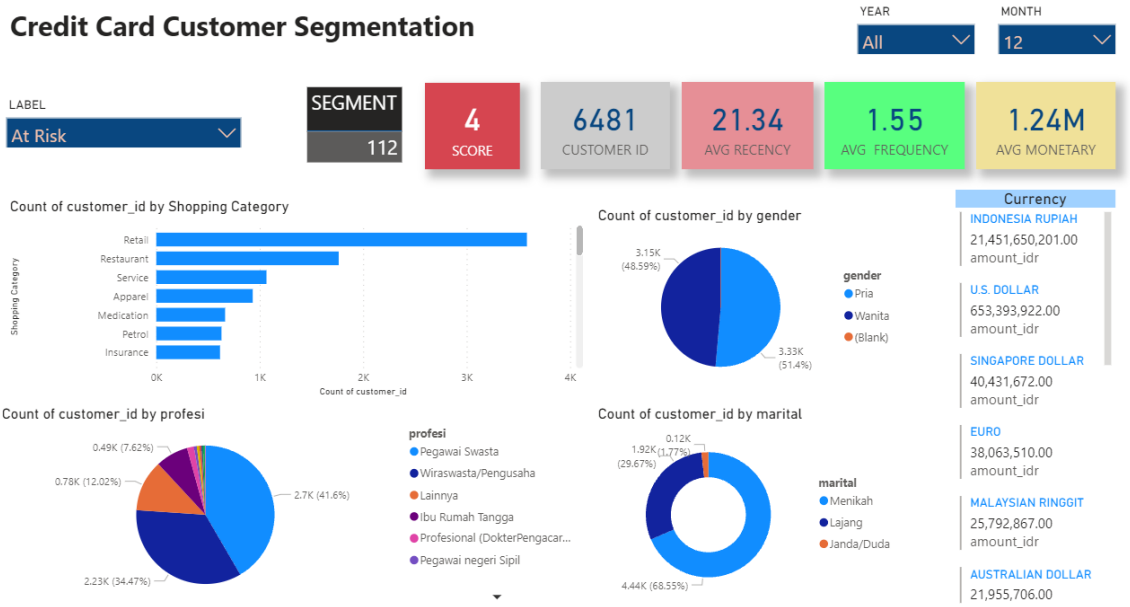
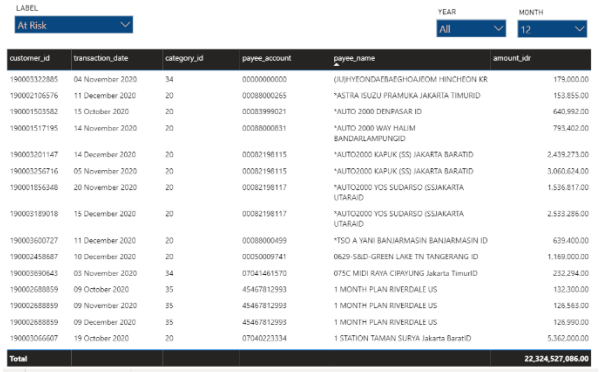


Fig. 6. The Dashboard Main Menu of Business Intelligence Portals.

Fig. 7. Customer Distribution in the Form of a Geographic Map.

3) Transaction detail. It displays detailed data of each customer transaction. The column consists of customer-id, transaction date, category id, payee account, payee name, and the amount in Indonesian rupiah (IDR). This information can be filtered based on the label, year, and month through a slicer on the top side as shown in Fig. 8.



LABEL: At Risk | YEAR: All | MONTH: 12

customer_id	transaction_date	category_id	payee_account	payee_name	amount_idr
190003323885	04 November 2020	34	00000000000	(U)HYEONDAEBAEGHOA(EOM HINCHEON KR	179.000.00
190002106576	11 December 2020	20	00088000265	*ASTRA ISUZU PRAMUKA JAKARTA TIMURID	153.855.00
190001503582	15 October 2020	20	00083999021	*AUTO 2000 DENPASAR ID	640.992.00
190001517195	14 November 2020	20	00088000831	*AUTO 2000 WAY HALIM BANDARLAMPUNGID	793.402.00
190003201147	14 December 2020	20	00082198115	*AUTO2000 KAPUK (SS) JAKARTA BARATID	2.439.273.00
190003256716	05 November 2020	20	00082198115	*AUTO2000 KAPUK (SS) JAKARTA BARATID	3.060.624.00
190001856348	20 November 2020	20	00082198117	*AUTO2000 YOS SUDARSO (SS)JAKARTA UTARAID	1.536.817.00
190003189018	15 December 2020	20	00082198117	*AUTO2000 YOS SUDARSO (SS)JAKARTA UTARAID	2.533.286.00
190003600727	11 December 2020	20	00088000499	*TSO A YANI BANJARMASIN BANJARMASIN ID	639.400.00
190002458687	10 December 2020	20	00050009741	0629-SBD-GREEN LAKE TN TANGERANG ID	1.169.000.00
190003690643	03 November 2020	34	07041461570	075C MIDI RAYA CIPAYUNG Jakarta TimurID	232.294.00
190002688859	09 October 2020	35	45467812993	1 MONTH PLAN RIVERDALE US	132.300.00
190002688859	09 November 2020	35	45467812993	1 MONTH PLAN RIVERDALE US	126.563.00
190002688859	09 December 2020	35	45467812993	1 MONTH PLAN RIVERDALE US	126.990.00
190003066607	19 October 2020	20	07040223334	1 STATION TAMAN SURYA Jakarta BaratID	5.362.000.00
Total					**22,324,527,086.00**

Fig. 8. The Detail of Credit Card Transaction.

4) Customer Detail. It displays the data of customer-id, customer name, sex, position, profession, marital status, zip code, and date of birth. As in the transaction detail, on the top side, there is a slicer to filter the data based on the segmentation, year, and month.

5) Forecasting Credit Transaction. It contains credit card transaction predictions in the next few days. Forecasting uses an Exponential Smoothing method that has been available in the Power BI application. Exponential Smoothing is a forecasting method of a moving average providing an exponential or graded weight in the latest data [29]. The graph can be seen in Fig. 9.

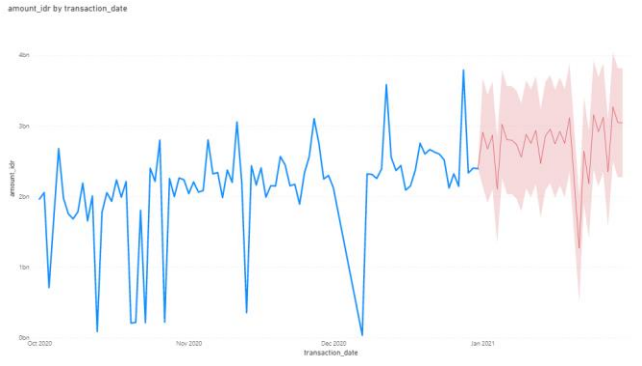


Fig. 9. Forecasting of the Amount of Credit Card Transactions.

6) Transaction History. Show in Fig. 10, displays the transaction history based on the shopping category per month in tabular form to the side part. The far-right column contains a difference in the amount in December takes away the amount in November. If the difference is positive, it shows a green color. Meanwhile, if the difference is negative, it shows red color. The top side has filtering according to year.

YEAR: 2020

category_name	October	November	December	Des - Nov
Apparel	3,075,197,591.00	4,254,841,188.00	4,236,944,676.00	-17,896,512.00
Audiovisual	118,202,603.00	342,195,575.00	311,534,295.00	-30,661,280.00
Beauty	1,199,893,787.00	1,508,490,744.00	1,243,242,324.00	-265,248,420.00
Cable TV	508,488,413.00	535,240,027.00	435,049,743.00	-100,190,284.00
Cafe	195,890,266.00	282,546,223.00	252,343,846.00	-30,202,377.00
Donation	82,429,750.00	85,649,935.00	40,145,798.00	-45,504,137.00
Education	162,083,988.00	178,586,153.00	190,599,805.00	12,013,652.00
Electronic	1,472,025,933.00	1,640,861,952.00	1,600,953,704.00	-39,908,248.00
Entertainment	897,517,521.00	725,555,122.00	477,125,279.00	-248,429,843.00
E-Wallet	2,655,443.00	4,944,338.00	10,524,936.00	5,580,598.00
Food & Drink	269,773,613.00	369,991,721.00	432,750,513.00	62,758,792.00
Gadget	46,924,224.00	31,810,850.00	31,070,300.00	-740,550.00
Games	451,596,632.00	448,844,763.00	385,277,388.00	-63,567,375.00
Gas & Electricity & Water	12,039,276.00	25,681,028.00	17,643,363.00	-8,037,665.00
Hobby	250,683,519.00	314,949,311.00	315,424,782.00	475,471.00
Insurance	4,737,119,517.00	6,383,239,183.00	4,938,243,182.00	-1,444,996,001.00
Lodging	1,028,829,942.00	1,403,979,509.00	1,594,240,926.00	190,261,417.00
Total	**53,169,212,314.00**	**67,723,124,723.00**	**62,164,839,832.00**	**-5,558,285,285.00**

Fig. 10. Monthly Credit Card Transaction History.

7) Customer History. The data show a monthly customer history shown in a tabular form. There is a column showing a gap between the amount in December and the amount in November. If the value is positive, it shows green; if the value is negative, it shows red. The variable of the data presented according to year can be selected, either frequency, recency, monetary, or RFM score. The illustration can be seen in Fig. 11.

YEAR: 2020 | VARIABLE: FREQUENCY

CUSTOMER_ID	OKTOBER	NOVEMBER	DECEMBER	DEC-NOV
190000000882	3	0	0	0
190000001432	1	1	2	1
190000001467	3	2	11	9
190000001950	1	1	2	1
190000002584	1	2	2	0
190000002884	5	0	0	0
190000003577	0	1	2	1
190000003672	1	4	3	-1
190000004082	1	0	0	0
190000004200	1	2	0	-2
190000004706	1	0	4	4
190000007237	2	0	0	0
190000007238	2	4	4	0
190000007606	0	0	2	2
190000007617	0	1	4	3
190000008250	0	1	0	-1
Total	**101445**	**121569**	**110791**	**-10778**

Fig. 11. Monthly Credit Card Customer History.

VI. Conclusion

This study makes a segmentation using a machine learning clustering model and business intelligence in the customers of data warehouse-based credit cards. The data warehouse concept utilization is used for constructing an integrated data management system that can handle a large amount of data. The machine learning clustering model is used for grouping customers according to a rank to know the most loyal customers and inactive customers. This clustering uses the RFM (Recency, Frequency, and Monetary) analysis as the measurement variable and feature selection in the clustering process. The RFM analysis is chosen because it can present customer loyalty based on their shopping behavior, such as the last time of shopping (recency), the frequency of shopping (frequency), and the amount of money spent for shopping (monetary).

The data from the machine learning clustering process are combined with other data and presented in the form of a business intelligence portal dashboard. The dimension modeling process uses a star schema because it is superior in terms of speed compared to the snowflake schema. The data and graphics displayed in the business intelligence portal present the segmentation data according to demography, geography, and behavior.

The combination of machine learning clustering segmentation for ranking customers with segmentation based on the demography, geography, and behavior provides complete and strong information as a support in a business decision for a marketing department. For example, the marketing department wants to increase the customers' credit card transactions by a limited promotional budget. First, he/she will see a group of customers having a relatively low RFM score. Then, from the group, the city with the highest number of customers is observed. Thus, the marketing department can effectively promote to a group of customers with the same characteristics.

This study also shows the minibatch k-means clustering algorithm has faster performance than that of agglomerative hierarchical clustering, BIRCH, and k-means algorithms. The result shows that out of 46.097 rows of data, the minibatch k-means method is superior to agglomerative clustering, BIRCH is superior with a thin margin to k-means. Some clusters are tested using an elbow method. The result shows that the best and optimal cluster is 4 clusters.

References

[1] M. Pradana, "Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering," J. Appl. Data Sci., vol. 2, no. 1, pp. 19–25, 2021, doi: 10.47738/jads.v2i1.18.

[2] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," J. City Dev., vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.

[3] J. Wu et al., "An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K -Means Algorithm," Math. Probl. Eng., vol. 2020, no. November 2017, 2020, doi: 10.1155/2020/8884227.

[4] M. McCaig and D. Rezania, "A Scoping Review on Data Governance," SSRN Electron. J., no. Icicnis, 2021, doi: 10.2139/ssrn.3882450.

[5] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. Ben Yahia, "Data quality in ETL process: A preliminary study," Procedia Comput. Sci., vol. 159, pp. 676–687, 2019, doi: 10.1016/j.procs.2019.09.223.

[6] M. Supriyono, Buku pintar perbankan. JOgjakarta: Andi Offset, 2011.

[7] A. Amborowati and M. Suyanto, "Studi Dukungan Marketing Intelligence pada Strategi Pemasaran," Semin. Nas. Inform. 2015, vol. 2015, no. November, pp. 49–53, 2015.

[8] P. Kotler and K. L. Keller, Marketing Management, 15th ed. Harlow: Pearson, 2016.

[9] P. Kolarovszki, J. Tengler, and M. Majerčáková, "The New Model of Customer Segmentation in Postal Enterprises," Procedia - Soc. Behav. Sci., vol. 230, no. May, pp. 121–127, 2016, doi: 10.1016/j.sbspro.2016.09.015.

[10] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," J. King Saud Univ. - Comput. Inf. Sci., 2018, doi: 10.1016/j.jksuci.2018.09.004.

[11] R. Sherman, Business Intelligence Guidebook From Data Integration to Analytics. Waltham: Elsevier Inc, 2015.

[12] Y. Zhao, "Transformation of Business Analytics from Business Intelligence," E3S Web Conf., vol. 253, pp. 3–6, 2021, doi: 10.1051/e3sconf/202125303013.

[13] J. Hurwitz and D. Kirsch, Machine Learning For Dummies IBM Limited Edition, 2018th ed. New Jersey: John Wiley & Sons, Inc, 2018.

[14] J. D'Silva and U. Sharma, "Unsupervised Automatic Text Summarization of Konkani Texts using K-means with Elbow Method," Int. J. Eng. Res. Technol., vol. 13, no. 9, pp. 2380–2384, 2020, doi: 10.37624/ijert/13.9.2020.2380-2384.

[15] H. Humaira and R. Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," 2020, doi: 10.4108/eai.24-1-2018.2292388.

[16] B. N. Sari, "Identification of Tuberculosis Patient Characteristics Using K-Means Clustering," Sci. J. Informatics, vol. 3, no. 2, pp. 129–138, 2016, doi: 10.15294/sji.v3i2.7909.

[17] L. Zahrotun, "Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering," Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017, vol. 2018-Janua, pp. 299–303, 2018, doi: 10.1109/ICITISEE.2017.8285515.

[18] R. D. F. Ruly, Purbandini, and E. Wuryanto, "Penerapan Clustering K-Means Pada Customer Segmentation Berbasis Recency Frequency Monetary (Rfm) (Studi Kasus : Pt . Sinar Kencana Intermoda Surabaya)," Semin. Nas. Mat. Dan Apl., pp. 418–427, 2017.

[19] J. Jamal and D. Yanto, "Analisis RFM dan Algoritma K-Means untuk Clustering Loyalitas Customer," Energy, vol. 9, no. 1, pp. 1–8, 2019.

[20] A. Aziz, "Customer Segmentation basedon Behavioural Data in E-marketplace," 2017, [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1145508.

[21] W. Qadadeh and S. Abdallah, "Customers Segmentation in the Insurance Company (TIC) Dataset," Procedia Comput. Sci., vol. 144, pp. 277–290, 2018, doi: 10.1016/j.procs.2018.10.529.

[22] J. Silva, N. Varela, L. A. B. López, and R. H. R. Millán, "Association rules extraction for customer segmentation in the SMES sector using the apriori algorithm," Procedia Comput. Sci., vol. 151, no. 2018, pp. 1207–1212, 2019, doi: 10.1016/j.procs.2019.04.173.

[23] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," Lontar Komput. J. Ilm. Teknol. Inf., vol. 11, no. 1, p. 32, 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[24] K. K. Halim, S. Halim, and Felecia, "Business intelligence for designing restaurant marketing strategy: A case study," Procedia Comput. Sci., vol. 161, pp. 615–622, 2019, doi: 10.1016/j.procs.2019.11.164.

[25] J. Choi, J. Yoon, J. Chung, B. Y. Coh, and J. M. Lee, "Social media analytics and business intelligence research: A systematic review," Inf. Process. Manag., vol. 57, no. 6, p. 102279, 2020, doi: 10.1016/j.ipm.2020.102279.

[26] P. L. Bourbonnais and C. Morency, "A robust datawarehouse as a requirement to the increasing quantity and complexity of travel survey data," Transp. Res. Procedia, vol. 32, pp. 436–447, 2018, doi: 10.1016/j.trpro.2018.10.054.

[27] C. D'Arconte, "Business intelligence applied in small size for profit companies," Procedia Comput. Sci., vol. 131, pp. 45–57, 2018, doi: 10.1016/j.procs.2018.04.184.

[28] V. Khatibi, A. Kheramati, and F. Shirazi, "Deployment of a business intelligence model to evaluate Iranian national higher education," 2020, [Online]. Available: www.elsevier.com/locate/ssaho.

[29] H. Yonar, "Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods," Eurasian J. Med. Oncol., no. April, 2020, doi: 10.14744/ejmo.2020.28273.