Enhancing Customer Segmentation Insights by using RFM + Discount Proportion Model with Clustering Algorithms

Victor Hugo Antonius, Devi Fitrianah
BINUS Graduate Program - Master in Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480

***Abstract*—In this digital era, the use of e-commerce has expanded and is widely adopted by society. One of the reasons why people use e-commerce platforms is because of their convenience and ease of use. However, the rapid growth of e-commerce has led to a substantial rise in transactions within the platform, involving various business entities. Therefore, it is crucial to perform customer segmentation to group them based on their purchasing behavior. The implementation of data mining techniques, such as clustering, is highly beneficial in this case. Clustering helps process datasets and transform them into useful information. In this study, transaction data obtained from one of the e-commerce stores, i.e. MurahJaya888 and followed by analysis using various clustering methods such as K-means, K-medoids, Fuzzy c-means, and Mini-batch k-means. We also proposed a new model that will become the attributes cluster, namely, RFM + DP (Discount Proportion). The Discount Proportion Rate will provide more insights for customer segmentation as it helps understand purchasing behavior that is more responsive to discount utilization. Implementing these four clustering methods with RFM + DP model resulted in four clusters based on the optimal elbow method. Furthermore, the evaluation and performance metrics for each clustering algorithm indicate that Mini Batch K-Means achieved the highest silhouette score of 0.50. Meanwhile, K-Means obtained the highest CH index value compared to the other algorithms, which was 1056.**

Keywords*—Clustering; RFM; discount proportion; customer segmentation; data mining

I. Introduction

The entire world has been shaken by big news in 2020, COVID-19 pandemic. This pandemic has taken many human lives and made a huge impact on all aspects of human life such as social, political, and even economic. In the economic field, everyone knows that businesses have to go through difficult times in the COVID-19 era to survive [1] [2]. Therefore, the pressure they received must be faced with the increasing adoption of technology 4.0. This reliance on technologies is crucial to not only survive but also to adapt and thrive effectively.

The rapid adoption of technology in this digital era has come up with opportunities to boost efficiency and expand business market presence. One effective way to achieve this is by using e-commerce platforms. e-Commerce is concerned with facilities such as marketing, selling products or services, delivering, and developing over the internet [3]. e-Commerce enhances companies' efficiency and reliability by automating transactions [4]. In this case, transactions can be done online, making it easier for people like customers who make purchases and use services.

The use of e-commerce in the business field has played big role across the world, especially in the context of Indonesia. Many businesses have switched their business from traditional offline business to selling online because the growth of online business and technologies [5]. Also, some of them have been implementing hybrid models to maintain their business continuity, entice to their own customers, and provide good value through products and services. According to latest statistical data in Indonesia collected by the BPS-Statistics Indonesia (2022), the survey results show that the number of e-commerce businesses in Indonesia in 2021 was 2,868,178 businesses which experienced an increase from 2020, although the growth is not too much. The same situation also occurred in 2022. With this volume of visitors, the use of e-commerce signifies convenience for consumers. There are many popular e-commerce sites or platforms in Indonesia such as Tokopedia, Shopee, Blibli, and many more. These platforms are frequently visited and have a high number of transactions due to the presence of various business entities and customers [6].

While it's true that e-commerce platforms process countless transactions daily, business entities and sellers need to evaluate their purchase behavior by doing customer segmentation. One of the most common ways is by using RFM model. RFM will help to understand the behavior and customers value based on their characteristics [7]. Even small businesses can significantly benefit from customer segmentation. Therefore, for processing transaction data in order to understand customer segmentation, a technique in data mining can be utilized to determine the relation of its data [8]. Hence, using the RFM model itself is felt to be insufficient for representing customer segmentation. General customer segmentation is divided into four types, namely, based on demographic (customer's personal information), geographic (location, population density), purchase behavior, and psychographic factors [9]. These can be used as an additional feature in the RFM model to gain new insights that help understand customer behavior better, enhance market targeting, and optimize marketing strategies. This proves that RFM model in customer segmentation can be adjusted according to business needs.

In data mining, there is unsupervised learning technique named clustering. This approach often used to group data objects based on their similarities. Clustering is a technique used in data mining analysis that groups of data objects into a set based on their similar characteristics [8] [10]. Particularly in customer segmentation, clustering techniques play a crucial role. By applying clustering algorithms to customer data, businesses can identify groups or segments with similar behaviors and preferences. This segmentation helps companies in tailoring marketing strategies and enhancing overall customer satisfaction.

Based on the findings outlined in [11] states that startup businesses can achieve quicker adaptation by thoroughly comprehending their market and customer base. Therefore, customer segmentation is an effective market strategy for grasping customer characteristics. In their approach, they employ clustering methods combined with the RFM (Recency, Frequency, Monetary) model, They use clustering techniques with the RFM model because of its ease of application to the market and give empirical of the better result. Clustering also allows to see hidden patterns based on customer behavior because the properties of the mining data are specific to finding pattern.

The purpose of this study is to compare four clustering algorithms in data mining such as K-means, K-medoids, Fuzzy c-means, and Mini batch k-means. The initial step of this research is to conduct literature to collect papers related to e-commerce, and customer segmentation using several models, and clustering algorithms. Additionally, the related dataset was obtained from a store called MurahJaya888 in Indonesia on the Shopee e-commerce platform in Indonesia. The next step involves using clustering techniques in the Python programming language to put into practice our proposed model that merges RFM (Recency, Frequency, Monetary) with the Discount Proportion attribute. The benefit of using the Discount Proportion attribute is to observe whether those who shop tend to frequently take advantage of discounts or not. It indicates whether their behavior involves frequent shopping, particularly highlighting if they are inclined to make purchases more frequently when discounts are available. Through this research, we aim to answer pivotal research questions such as how the integration of discount proportion with RFM clustering enhances customer segmentation accuracy, and how these refined segments can be effectively leveraged to optimize targeted marketing efforts in the retail sector.

II. Related Works

The following are summaries of previous research that are relevant to this study. Table I presents a summary of the related works, outlining key findings and insights gathered from prior studies in the field.

Based on all of this research, we understand that the use of clustering and an expanded RFM model can lead to deeper customer segmentation. However, clustering can become overly complex due to the additional variables introduced. For example, in the case of this study, we propose the RFM + Discount Proportion in percentage model to investigate several customer segments and find hidden patterns that are more responsive to discounts, whether they tend to shop during discount periods or not, and with significant discounts, they become loyal customers or only take advantage of discounts occasionally.

TABLE I. Summary of Related Works

Dataset	Methodology	Result
UK retail dataset from UCI machine learning repository	RFM model with various clustering algorithms.	The dataset has more than 540,000 records. We tried different methods like k-means, GMM, DBSCAN, BIRCH, and Agglomerative Clustering to analyze it. The best result came from using GMM (Gaussian Mixture Model) with PCA for reducing dimensions. It got a silhouette score of 0.80, which is the highest compared to previous studies [12].
Brazillian store dataset from Kaggle	K-means with RFM and create a website-based dashboard using Streamlit	With over 100,000 records, a clustering method with a k of 4 was employed. The clusters were visualized based on RFM values, yielding a silhouette coefficient of 0.47. Additionally, this research visualized the results using Streamlit, featuring three key components: overview, RFM Analysis, and page report, aimed at providing valuable insights [13].
UCI machine learning repository	RFM-D(Diversity) with various machine learning algorithms	Diversity, an attribute that complements RFM, measures the variety of products purchased by each customer. The dataset utilized the purchase history of 4,383 transactions. Various machine learning methods including SVM, Decision Tree, and K-Means were compared. However, the highest silhouette score accuracy of 0.98 was attained using K-Means [14].
Olist store dataset from Kaggle	RFMTS with K-means algorithm	T stands for Time and S for Satisfication score. Due to the utilization of five variables or attributes, the researchers opted for PCA to reduce the dimensionality of their dataset. Based on the elbow method, they determined the optimal cluster number to be 5. They did not explicitly focus on the clustering algorithm but discussed the results regarding customer statistics, which, if ignored, could have a negative impact on the company [15].
Bank dataset	RFM+ B(Balance) model with K-means	The B model, which is the amount of savings a customer has at the end of the data period, can be very helpful for grouping customers and is useful in developing marketing strategies. The data was obtained from a bank with over 147,000 entries and was processed using k-means clustering. The accuracy level achieved from grouping savings customers using the RFM+B model is 77.58% [16].

III. Research Methodology

The general process of the methodology of this study is shown in Fig. 1. The initial step entails collecting literatures to define the problem through a literature study. After defining the problem and establishing the goals of this research area, the subsequent steps involve collecting the dataset, preprocessing the data, implementing models such as K-Means, K-Medoids, Fuzzy C-Means, and Mini-Batch K-

Means, and evaluating them to determine the algorithm that comes out with the best performance. For a better understanding, let's take a look at the research stages Fig. 1:

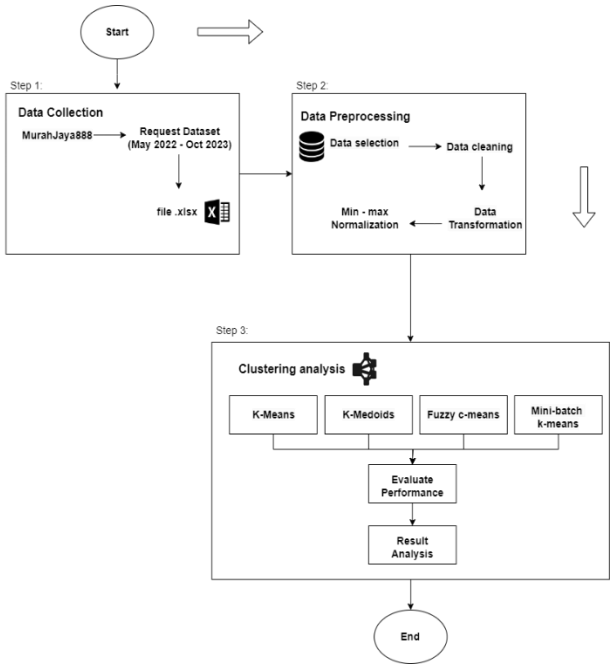

Fig. 1. Research Stages.

A. Data Collection

The dataset is collected and obtained from an online store called MurahJaya888 in Indonesia on the e-commerce platform. The dataset is provided based on a request from the shop owner as an example of a research subject in this e-commerce area. The data obtained is in .xlsx files format where there are 12 files (October 2021 – October 2023) with a total of 46 columns and 2497 rows. Before preprocessing the data, we must combine them into one file and convert it into data frame so that will help us to analyze more easily using Python programming language.

B. Pre-processing Data

After combining the data obtained into a dataframe, the next step is to process the dataset by pre-processing. Pre-processing is the process of preparing data to make it easier and feasible to analyze. Preprocessing is also use to enhance the quality of data, ensuring that it fulfills the requirements of the algorithms [17]. The preprocessing stages we use included data selection, data cleaning, and transformation.

Data selection is the first preprocessing step in this study to drop some columns that are irrelevant to be analyzed. Raw data that are received must be preprocessed before diving into the next step. By selecting the features that will be used later on in this research, we did feature selection first to prepare for the next important features to be analyzed [18]. Data cleaning is a method used to address issues or errors in the dataset that will be analyzed. This stage involves handling missing values, duplicates, renaming columns, and other tasks to prevent numerous outliers [19]. Moreover, data transformation involves altering raw data to make it suitable for analysis, such as creating variables or features to fulfill the requirements of the analysis purpose [20]. There are several techniques that can be applied, such as convert categorical variables into numerical types and attribute construction. Table II is the formula used in creating RFM + Discount Proportion attributes:

TABLE II. RFM + DP FORMULA

RFM Attribute	Calculation Formula
Recency(R)	$\sum_{i=1}^{n}(T_{now} - T_i)$
Frequency(F)	$\sum TP_i$
Monetary(M)	$\sum_{i=1}^{n} TAS_i$
Discount Proportion (DP)	$\frac{\sum_{i=1}^{n} T_{discount,i}}{\sum_{j=1}^{m} T_{payment,j}} * 100\%$

C. Normalization

Data normalization is a way of turning data into a series with the same range. In general, the normalization that is often used in research is Min-max normalization. The formula for Min-max normalization is as follows [21]:

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

In this formula, the attributes value will be on a scale of 0 to 1. So, the RFM attributes created will be normalized so that the attribute ranges are not significantly different from each other.

D. Cluster Analysis

Clustering is methods that can assist us in analyze our data by identifying similarities among data points. Each data point refers to each object or individual present in the data. Therefore, clustering groups data points based on the similarity of their attributes with the goal of discovering patterns of dataset [22].

The clustering method is a process grouping data objects that are similar to each other into the same cluster but different from objects in other clusters [10]. There are various types and algorithms of clustering. Here are the algorithms that will be explained along with their working steps:

1) K-means: K-Means is an unsupervised learning algorithm for grouping data objects based on the shortest distance between data points. K-Means algorithm steps are as following [23]:

- Determine the value of k as many as the number of clusters.
- Select the data that will be the center of the cluster or temporary centroid.
- The algorithm will calculate the distance between objects to the centroid and then group them.

$$d(x,y) = \sqrt{(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2)} \quad (2)$$

- Calculate the next centroid value with the average value of the data that has been obtained.
- Repeat until the condition is met or there is no more cluster changes.

2) K-medoids: K-medoids is an algorithm that is similar to K-means in its clustering process. However, in K-medoids, there are medoids that serve as the representatives of the clusters. The following is the mechanism of the K-medoids algorithm [24]:

- Initiate by choosing a temporary medoid as the initial medoid.
- In this step, each data object is assigned to the medoids that have the shortest distance (minimum distance).
- For each medoid, it's tested by swapping it with each non-medoid point one by one. If this swap reduces the total cost.
- Choose the lowest total cost.
- Repeat steps 2 to 4 until there is no medoid change so that clusters and the members of their corresponding clusters are acquired.

3) Fuzzy c-means: Fuzzy c-means is a clustering algorithm that is flexible and can classify a data object into two or more clusters based on its membership level. This can be referred to as soft clustering. The workings of the Fuzzy c-means (FCM) algorithm are as follows [25]:

- Determine the data to be grouped or clustered where X is a matrix of size i x j.
- Set the values needed for the FCM calculation, such as maximum iterations, expected number of clusters, epsilon value, objective function, and rank matrix partition (w).
- Initialize the partition matrix randomly.
- Calculate cluster center with the formula:

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^{w} * X_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \quad (3)$$

- Calculate the distance to the objective function with the following formula:

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}\left(X_{ij} - V_{kj}\right)^2\right] * (\mu_{ik})^{w}\right) \quad (4)$$

- Update the partition matrix:

$$\mu_{ik} = \frac{\left[\sum_{i=1}^{n}(X_{ij} - V_{kj})^2\right]^{\frac{1}{(w-1)}}}{\sum_{k=1}^{n}\left[\sum_{i=1}^{n}(X_{ij} - V_{kj})^2\right]^{\frac{1}{(w-1)}}} \quad (5)$$

- If the difference between the current partition matrix (P_n) and the initial partition matrix (P_0) is less than the epsilon value, or if the number of iterations (t) has reached the maximum number of iterations. The iteration stop.

4) Mini-batch K-means: Mini-batch k-means is like an alternative version of k-means algorithm because the working steps are quite similar but mini-batch k-means is designed for reduce the computational time [26] and use a random mini batch from the dataset to update the clusters [27]. Here are the steps:

- Randomly select a batch of data points from the dataset for the k-means Mini-batch algorithm.
- Assign the nearest centroid to each batch. The formula for finding the shortest distance can be done the same as the K-means algorithm.
- Update the centroids by calculating the average of data points within each cluster until convergence is reached. Convergence means the algorithm has stabilized, typically by comparing changes in centroid positions using the formula,

$$\Delta C = \sqrt{\sum \frac{k}{i=1}\left|\left|C_i - C_i'\right|\right|^2} \quad (6)$$

- Repeat steps 2 to 3 for the next batch in the predetermined number of iterations.

E. Performance Evaluation

The performance evaluation of the four algorithms above will also use and compare several performance metrics, the Silhouette Coefficient and Calinski-Harabasz Index. Here are the formulas for calculating the Silhouette coefficient and Calinski-Harabasz Index:

1) Silhouette coefficient: Silhouette coefficient is one of the performance metrics to calculate the quality of the algorithms. Check whether the algorithm used has reached a goodness point or not in clustering the data. A greater Silhouette coefficient indicates a better cluster [28]. These are the formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}} \quad (7)$$

2) Calinski-Harabasz index: CH Index metrics will calculate the variance ratio within clusters and also between clusters. The higher the CH value compared, the better the performance of the clustering algorithm [29]:

The best silhouette coefficient is a value that is close to 1, indicating that clustering has a good separation between clusters and consistent objects in the cluster. Meanwhile, the best CH value is the value that gives the highest number among all the k clusters.

F. RFM + Discount Proportion Analysis

In this stage, the data distribution of each customer segment will be explained in detail, as well as the benefits of the Discount Proportion for the dataset being used.

IV. Result and Analysis

Our background in conducting this research is that we want to test and compare several clustering algorithms for

customer segmentation using primary data from one of the stores on the e-commerce platform and using the RFM + Discount Proportion model which can produce more deeper result of segmentation. The data that has been collected from MurahJaya888 consists of 1458 rows after being preprocessed and will be processed first because there are many missing values and unused columns which will result in many outliers. These are the steps of our analysis:

A. Integrate Data Files

The dataset are obtained in separate .xlsx files per month. Therefore, we use the Python glob module to group specific file extension names. This approach helps us to combine data from multiple files, making analysis more comprehensive. These are the available columns of the dataset:

```
Index(['OrderID', 'Order Status', 'Cancellation/Return Status',
       'Tracking Number', 'Shipping Options', 'Counter/Pick-up Delivery',
       'Must Ship Before (Avoiding Delays)', 'Scheduled Delivery Time',
       'Order Creation Time', 'Payment Time', 'Parent SKU', 'Product Name',
       'SKU Reference Number', 'Variation Name', 'Original Price',
       'Price After Discount', 'Quantity', 'Total Product Price',
       'Total Discount', 'Discount from Seller', 'Shopee's Discount',
       'Product Weight', 'Quantity Ordered', 'Total Weight',
       'Seller-Funded Voucher', 'Cashback in Coins', 'Shopee-Funded Voucher',
       'Discount Package', 'Package Discount (Shopee's Discount)',
       'Package Discount (Seller's Discount)', 'Shopee Coin Deduction',
       'Credit Card Discount', 'Buyer-Paid Shipping Cost',
       'Estimated Shipping Cost Deduction', 'Return Shipping Cost',
       'Total Payment', 'Estimated Shipping Cost', 'Buyer's Notes', 'Notes',
       'Buyer Username', 'Recipient's Name', 'Phone Number',
       'Shipping Address', 'City/District', 'Province',
       'Order Completion Time'],
      dtype='object')
```

Fig. 2. Available columns.

Fig. 2 shows the columns available from the data we obtained to be further preprocessed including, data selection, cleaning, and the construction of the RFM + DP attributes for use in clustering. To calculate the DP attributes, we divide the total discount price by the total payment and then multiply the result by 100% (https://shorturl.at/afGKP).

B. Preprocessing Data

1) Data selection: From the 46 available columns, our initial step was to select useful features for creating the RFM + DP attributes and drop columns that were irrelevant for analysis. Table III displayed the columns that we selected for analysis after this selection process, providing a clear overview of the chosen attributes. This streamlined approach ensured that only pertinent data were considered for the RFM + DP model development.

TABLE III. COLUMNS AFTER DATA SELECTION

Columns	Description
Order ID	Unique identifier for each order
Product Name	Name of the product
Original Price	The price of the product before any discounts
Order Status	Indicates the current status of the order
Cancellation/Return Status	Indicates whether the order was canceled or returned
Quantity	The number of product purchased in the order
Total Payment	Total amount paid each order
Total Discount	Total discount applied to the order
Last Transaction Time	Timestamp of last transaction order
Username	User account name

We opted to choose a subset of 10 columns from the total 46 columns available. These specific columns are derived from the RFM + DP attribute formula, crucial for the clustering analysis aimed at customer segmentation and insights. It's worth noting that certain details such as customer demographics and shipping information will be excluded to uphold privacy concerns. Additionally, the Last Transaction Time column is included to track the most recent shopping date for each customer.

2) Data cleaning: In the data cleaning stage, we perform several tasks such as renaming columns from Indonesian to English, removing missing values in the Cancellation/Return Status column which indicate cancelled orders. Then, we eliminate duplicate values and ensure that all Order IDs are unique.

3) Feature construction for RFM + DP: In this data transformation stage, firstly, we convert the data type from object to numerical for several columns. Additionally, we perform feature construction for RFM + DP. Fig. 3 is an example dataset of RFM + DP attributes:

Recency	Monetary	Frequency	DiscountProportion
678	246000	3	3.529412
200	895700	1	0.000000
197	1785800	4	5.872757
681	184840	3	12.083333
592	296000	1	6.944444

Fig. 3. RFM + DP attributes before normalization.

Furthermore, we also create distributions to observe the spread of data for each variable on a scale of 1-5, Fig. 4 indeed show data imbalance, particularly in Frequency and Monetary. Fig. 5 to Fig. 7 shows different types of scale. However, we cannot remove such data as it naturally occurs in transactions:

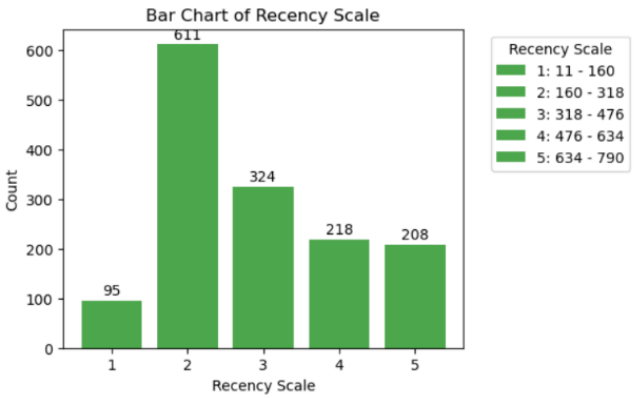

Fig. 4. Recency scale.

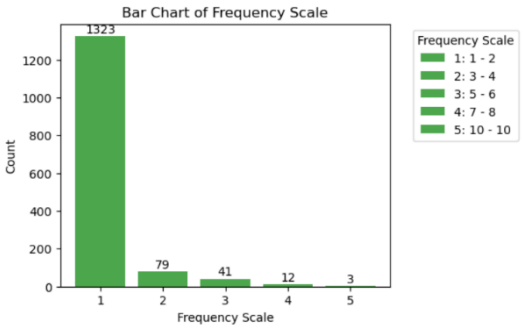

Fig. 5. Frequency scale.

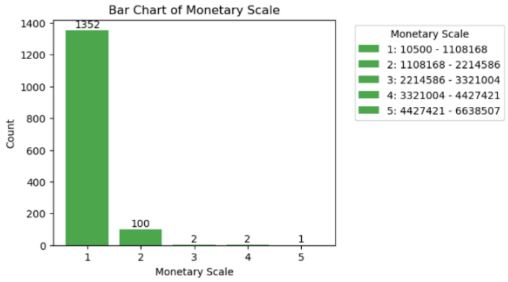

Fig. 6. Monetary scale.

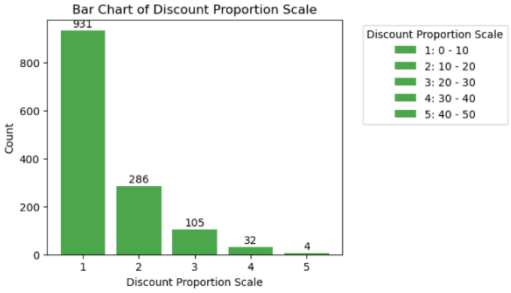

Fig. 7. Discount proportion scale.

C. Min-max Normalization

We changed the scale of the RFM + Discount Proportion attribute to 0-1. The aim is to minimize outliers and the value of attributes are relatively close to each other as shown in Fig. 8. These are the final results of the data with the normalized one:

Recency	Monetary	Frequency	DiscountProportion
0.856226	0.035531	0.222222	0.070588
0.242619	0.133554	0.000000	0.000000
0.238768	0.267848	0.333333	0.117455
0.860077	0.026304	0.222222	0.241667
0.745828	0.043075	0.000000	0.138889

Fig. 8. Min-Max normalization result.

D. Cluster Analysis

Table IV represents the information of customer types in MurahJaya888 store, there are four types of customers such as, Platinum, Gold, Silver, and Bronze after done the segmentation to ensure the right marketing strategy:

1) K-Means: The implementation of clustering algorithms will be done on an apple-to-apple basis, meaning we will use a number of k=4 based on the elbow method, with random state of 42. This approach ensuring reliability and comparability in our analysis across various clusters. Fig. 9 shows that the optimal number of k is 4 because there is a significant change in the points that form an elbow, indicating that adding clusters after k = 4 provides less decreases of the inertia value. This provides a balance between minimizing inertia in the resulting clusters and maintaining a fair interpretation of the existing data. We also use library named KneeLocator to ensure the optimal number and the result shown that is 4.

TABLE IV. CUSTOMER CHARACTERISTICS

Type of Customers	Values
Platinum	High Recency (↑) High Frequency (↑) High Monetary (↑) High Discount Proportion (↑)
Gold	High Recency (↑) Low Frequency (↓) Neutral Monetary (-) Neutral Discount Proportion (-)
Silver	Low Recency (↓) Low Frequency (↓) Low Monetary (↓) High Discount Proportion (↑)
Bronze	Low Recency (↓) Low Frequency (↓) Low Monetary (↓) Low Discount Proportion (↓)

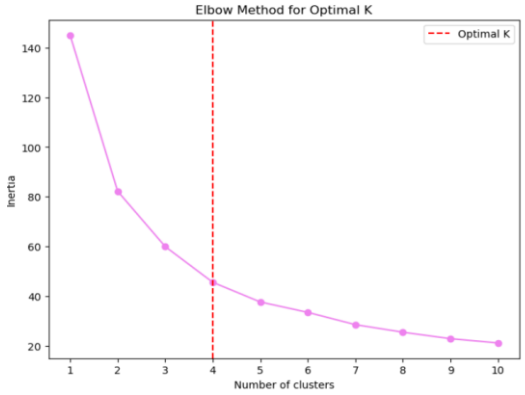

Fig. 9. Elbow Method for Optimal K.

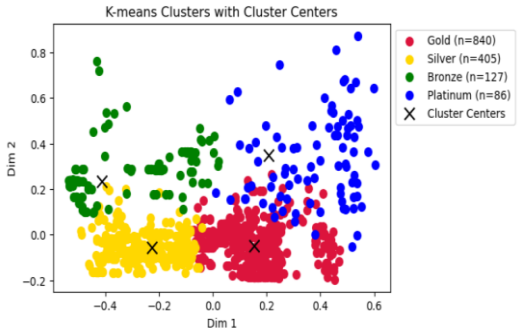

Fig. 10. K-means clustering result.

We used PCA to reduce our components or features to get better visualized in two dimensions. Since we can't plot all four features at once, it would make it difficult for us to understand the data distribution. Thus, PCA serves as a valuable technique for simplifying the data representation while preserving its essential characteristics. Fig. 10 shows the groups of data points in four colors: gold, silver, bronze, and platinum. The number of points in each group is shown by a number (n). The gold group has the most points (840), then silver (407), bronze (125), and platinum (86). There are 'X' marks that show the center or average spot of each group.

2) Mini-batch K-means: Mini-batch K-Means is often applied in larger datasets due to its characteristic nature of dividing the dataset into smaller portions.

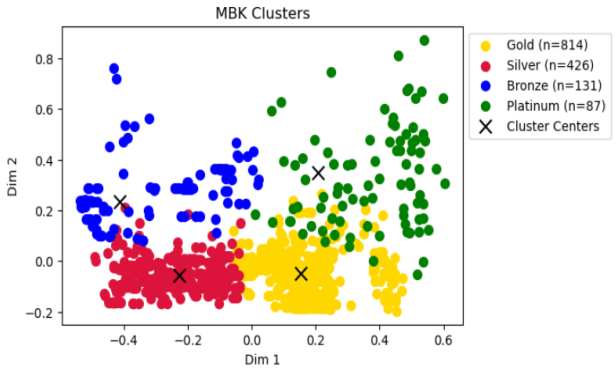

Fig. 11. Mini-batch K-means clustering result.

Fig. 11 demonstrates that with four clusters, its results align closely with those of the k-means algorithm, though there are slight differences in the number of points within some clusters. The distribution of points across the clusters is as follows: the gold cluster contains the highest number of points (814), followed by the silver cluster (426), the bronze cluster (131), and the platinum cluster (87). Surprisingly, the mini-batch k-means algorithm performs well, especially known for its efficiency in processing larger datasets. This efficiency is attributed to its method of dividing the data into smaller batches for each iteration. To achieve optimal clustering results, it's beneficial to experiment with different batch sizes to evaluate their impact on performance and computational time. Table V shows batch size iterations.

TABLE V. BATCH SIZE ITERATIONS

Batch Size	Execution Time	Silhouette Score
50	0.0638 seconds	0.4534
100	0.0439 seconds	0.3476
150	0.0499 seconds	0.4829
200	0.0449 seconds	0.4816
250	0.0439 seconds	**0.5002**
300	0.0519 seconds	0.4928
350	0.0568 seconds	0.4899
400	0.0559 seconds	0.4921

Mini-batch k-means performance was tested across different batch sizes ranging from 50 to 400, revealing a trade-off between execution time and silhouette scores. Smaller batch sizes like 50 and 100 offered faster execution but lower silhouette scores. As batch sizes increased, silhouette scores also improved, peaking at 0.5002 with a batch size of 250 which was identified as the optimal batch size considering both execution speed and performance.

3) K-medoids: The K-Medoids algorithm used here is quite similar to K-Means. However, in this case, the cluster centers are not represented by centroids but medoids. Medoids are points that represent the center of a cluster by assessing the distance between the medoid and all other points within the cluster. Fig. 12 is the results obtained from implementing the K-Medoids algorithm that produced four clusters.

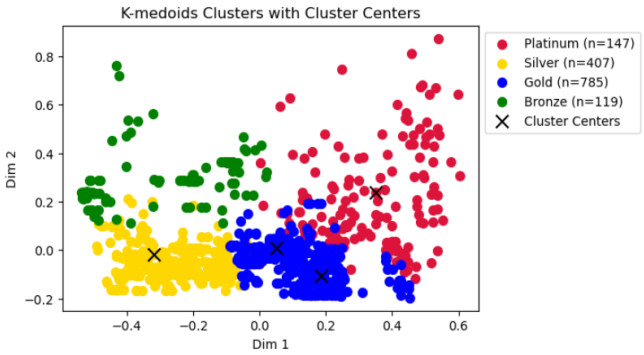

Fig. 12. K-medoids clustering result.

The points are distributed across the clusters as follows: the gold cluster contains the highest number of points (785), followed by the silver cluster (407), the bronze cluster (119), and the platinum cluster (147). K-medoids utilize the concept of medoids as the center point, rather than centroids. A medoid is a real member of the dataset, not an average value. It's possible that for the dataset we employed, K-medoids may not be as suitable, even though one of its advantages is its resilience to outliers.

4) Fuzzy c-means: The last algorithm tested was fuzzy c-means (FCM). This research implements FCM clustering using FCM in the fcmeans library. The FCM algorithm provided quite well-clustered results, similar to K-Means and Mini batch k-means, even though this algorithm is soft clustering, meaning it allows one data point to be included in different clusters. Therefore, there is a fuzziness parameter in the Fuzzy C-means algorithm to regulate the highest probability of each cluster data being most suitable for which cluster number.

We found that the optimal value is 1.1 based on its Silhouette score. Fig. 13 shows optimal fuzziness parameter. By selecting the optimal fuzziness parameter, we achieved improved cluster cohesion and separation in the FCM algorithm's results. The following Fig. 14 is the clustering results of FCM:

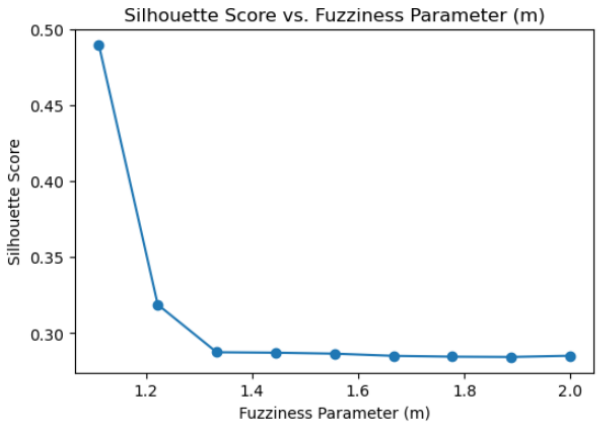

Fig. 13. Optimal fuzziness parameter (m).

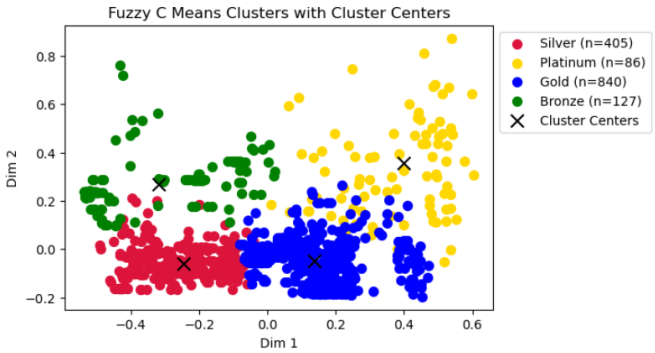

Fig. 14. FCM clustering result.

The FCM algorithm resulted in clusters with distributions almost identical to those produced by K-Means, with only slight differences in the dataset's data points. Among the clusters, the Platinum customers consist of a total of 88 customers considered potential. Next are the Gold customers, totaling 840, indicating that over 50% of customers either made their first purchase or returned to MurahJaya888? However, the Silver and Bronze clusters show fewer potential customers based on the reduced RFM + Discount proportion attributes.

E. Performance Evaluation

In Table VI, we can see that there is very little significant difference between the algorithms K-means, FCM, and Mini batch k-mean in the measurement of the Silhouette score. However, here, there is the highest result obtained by the mini-batch k -mean, namely, 0.5002. Silhouette score values that are > 0.5 are indicated as well clustered, which means that the data points are grouped well enough based on their similar characteristics. However, again that evaluation is not only influenced by the nature of each algorithm but also caused by other factors.

TABLE VI. CLUSTERING PERFORMANCE

Model	Silhouette Score	CH Index
K-Means	0.4921	1056.40
Mini batch K-means	0.5002	1053.18
K-Medoids	0.4714	966.45
FCM	0.4899	1055.39

Other factors include the characteristics of the datasets. In this sales data, we do have outliers like Recency and Monetary values that are unbalanced. But we can't just delete it because the data is important. For example, in this case of sale, it must be in one store to sell a lot of goods at different prices. If there are a few users who just buy expensive stuff. That could have caused the outlier not to go shopping too often. When we measured by the CH Index, the highest result was achieved by K-means with a value of 1056.40.

F. RFM + Discount Proportion Analysis

The use of RFM strategy in analyzing RFM model is quite common to be implemented either by scoring the RFM attributes or by using machine learning. However, our focus here is to add a new attribute, namely Discount Proportion in Percentage. An example illustration of DP is as follows in Fig. 16 and Fig. 15 shows RM +DP model.

Fig. 15. RFM + DP Model.

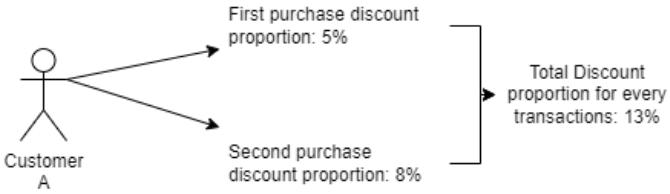

Fig. 16. Discount proportion illustration.

So, suppose client A spends 500,000 (IDR) in one transaction and gets a discount of 25,000 (IDR) so that when calculated the discount obtained from the purchase is 5%. It keeps counting and calculating for all their spending on different dates. These are the advantages of using this RFM + DP model:

- We can see it not just from the side of their shopping behavior. However, also from the internal side that is, the use of discounts. For example, in the case of this small MurahJaya888 dataset. They sell items that are not so varied and their customers can be categorized not so much. However, the feature construction of this Discount Proportion can give them insight into whether a customer can be classified as a loyal customer even though the discount given by the seller is large.
- This model can also be applied to cases of large datasets or retail stores that like to give discounts to their customers. Seeing from this side, they can determine whether giving continuous discounts can attract customers or cause losses.

- RFM usage can vary not only basic RFM, Discount Proportion provides insight on increasing customer loyalty by giving appreciation for their loyalty.

We attempted to analyze based on the cluster results produced by each algorithm and calculated the mean for each cluster. The overview can be observed in Table VII:

TABLE VII. CLUSTERING RESULT FOR RFM + DP ATTRIBUTES

		Mean				
Algorithms	**Customer Types**	**Recency (Days)**	**Monetary (Rupiah)**	**Frequency (Times)**	**Discount Proportion (Rate)**	**Count**
K-Means	Platinum	144	1.245.129	5.24	16.4	86
	Gold	272	407.360	1.2	7.5	840
	Silver	586	364.544	1.19	5.59	405
	Bronze	608	130.274	1.16	24.9	127
Mini-batch k-means	Platinum	137	1.269.237	5.3	14.9	87
	Gold	275	408.150	1.15	7.14	814
	Silver	611	327.639	1.2	7.4	126
	Bronze	495	174.592	1.24	28.6	131
K-Medoids	Platinum	151	1.039.860	4.01	12.8	147
	Gold	282	380.022	1.08	6.9	785
	Silver	591	361.116	1.2	5.7	407
	Bronze	604	121.179	1.15	25.4	119
Fuzzy c-means	Platinum	144	1.245.159	5.25	16.38	86
	Gold	273	407.363	1.13	7.15	840
	Silver	586	364.445	1.2	5.6	405
	Bronze	610	129.290	1.16	25.1	127

The customer segmentation reveals different clusters with unique characteristics. The "Platinum" segment, consisting of around 6-8% of customers, represents the highest tier. They enjoy significant discounts and remain loyal. Implementing exclusive discount programs and referral rewards can enhance their satisfaction. The "Gold" cluster, while stable overall, shows lower frequency values possibly due to its larger size. The "Silver" group, with many customers, requires analysis of their preferences for tailored marketing strategies. On the other hand, "Bronze" segment relies heavily on discounts but makes fewer purchases at MurahJaya888.

V. CONCLUSION

This study compares four partition methods of clustering algorithms, k-means, k-medoids, FCM, and mini-batch k-means, by producing three clusters, namely clusters 0, 1, 2, and 3. Based on the results we obtained and concluded, The Mini-batch K-means algorithm obtained the highest silhouette index value compared to the others, around 0,5 of accuracy. Meanwhile, the k-means algorithm managed to get the highest CH index value, namely 1056, 40. Therefore, the conclusion that can be drawn from this dataset is that the MurahJaya888 store still has not reached a large number of potential customers. Due to the limitations of dataset rows, additional datasets could be utilized to further explore the effectiveness of this model. For future research, other clustering algorithms such as density-based or hierarchical clustering could also be considered.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by Binus University Jakarta for this research, as well as all those involved in the writing process.

REFERENCES

[1] F. Rahmanov, M. Mursalov, and A. Rosokhata, “Consumer behavior in digital era: impact of COVID 19,” Mark. Manag. Innov., vol. 5, no. 2, pp. 243–251, 2021, doi: 10.21272/mmi.2021.2-20.

[2] M. Batool et al., “How COVID-19 has shaken the sharing economy? An analysis using Google trends data,” Econ. Res. Istraz. , vol. 34, no. 1, pp. 2374–2386, 2021, doi: 10.1080/1331677X.2020.1863830.

[3] P. M. Alamdari, N. J. Navimipour, M. Hosseinzadeh, A. A. Safaei, and A. Darwesh, “A Systematic Study on the Recommender Systems in the E-Commerce,” IEEE Access, vol. 8, pp. 115694–115716, 2020, doi: 10.1109/ACCESS.2020.3002803.

[4] S. S. Y. Shim, V. S. Pendyala, M. Sundaram, and J. Z. Gao, “Business-to-business e-commerce frameworks,” Computer (Long. Beach. Calif)., vol. 33, no. 10, pp. 40–47, 2000, doi: 10.1109/2.876291.

[5] M. I. Wanof and A. Gani, “MSME Marketing Trends in the 4.0 Era: Evidence from Indonesia,” Apollo J. Tour. Bus., vol. 1, no. 2, pp. 36–41, 2023, doi: 10.58905/apollo.v1i2.22.

[6] Y. M. Ginting, T. Chandra, I. Miran, and Y. Yusriadi, “Repurchase intention of e-commerce customers in Indonesia: An overview of the effect of e-service quality, e-word of mouth, customer trust, and customer satisfaction mediation,” Int. J. Data Netw. Sci., vol. 7, no. 1, pp. 329–340, 2023, doi: 10.5267/j.ijdns.2022.10.001.

[7] D. L. Aditya and D. Fitrianah, “Comparative Study of Fuzzy C-Means and K-Means Algorithm for Grouping Customer Potential in Brand Limback,” J. Ris. Inform., vol. 3, no. 4, pp. 327–334, 2021, doi: 10.34288/jri.v3i4.241.

[8] H. Xin and S. Zhang, “Construction of Social E - commerce Merchant Segmentation Model Based on Transaction Data,” 2023, doi: 10.4108/eai.28-10-2022.2328461.

[9] K. Banerjee, "AI Driven Customer Segmentation and Recommendation of Product for Super Mall," no. December, 2023, doi: 10.18311/jmmf/2023/34166.

[10] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[11] D. Panji Agustino, I. Gede Harsemadi, and I. Gede Bintang Arya Budaya, "Edutech Digital Start-Up Customer Profiling Based on RFM Data Model Using K-Means Clustering," J. Inf. Syst. Informatics, vol. 4, no. 3, pp. 724–736, 2022, [Online]. Available: http://journal-isi.org/index.php/isi.

[12] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," Analytics, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.

[13] F. Alzami et al., "Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce with Streamlit," Ilk. J. Ilm., vol. 15, no. 1, pp. 32–44, 2023, doi: 10.33096/ilkom.v15i1.1524.32-44.

[14] M. Y. Smaili and H. Hachimi, "New RFM-D classification model for improving customer analysis and response prediction," Ain Shams Eng. J., vol. 14, no. 12, p. 102254, 2023, doi: 10.1016/j.asej.2023.102254.

[15] D. Mensouri, A. Azmani, and M. Azmani, "K-Means Customers Clustering by their RFMT and Score Satisfaction Analysis," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, pp. 469–476, 2022, doi: 10.14569/IJACSA.2022.0130658.

[16] U. Firdaus and D. N. Utama, "Development of bank's customer segmentation model based on rfm+b approach," ICIC Express Lett. Part B Appl., vol. 12, no. 1, pp. 17–26, 2021, doi: 10.24507/icicelb.12.01.17.

[17] A. F. Hardiyanti and D. Fitrianah, "Perbandingan Algoritma C4.5 dan Multilayer Perceptron untuk Klasifikasi Kelas Rumah Sakit di DKI Jakarta," J. Telekomun. dan Komput., vol. 11, no. 3, p. 198, 2021, doi: 10.22441/incomtech.v11i3.10632.

[18] T. C. Chen et al., "Application of Data Mining Methods in Grouping Agricultural Product Customers," Math. Probl. Eng., vol. 2022, 2022, doi: 10.1155/2022/3942374.

[19] V. Dawane, P. Waghodekar, and J. Pagare, "RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention," SSRN Electron. J., no. Icsmdi, 2021, doi: 10.2139/ssrn.3852887.

[20] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir," IEEE Access, vol. 8, pp. 151847–151855, 2020, doi: 10.1109/ACCESS.2020.3014021.

[21] R. K. H. B, A. H. Salim, and B. D. Meilani, Comparison of the Normalization Method of Data in Classifying Brain Tumors with the k-NN Algorithm, vol. 1. Atlantis Press International BV. doi: 10.2991/978-94-6463-174-6.

[22] I. Chatterjee, Machine Learning and Its Application: A Quick Guide for Beginners. Bentham Science Publishers, 2021.

[23] R. Raja, Rohit; Nagwanshi, Kapil, Kumar; Kumar, Sandeep; Laxmi, K., Data Mining and Machine Learning Applications. John Wiley & Sons, 2022.

[24] L. Zahrotun, U. Linarti, B. H. T. Suandi As, H. Kurnia, and L. Y. Sabila, "Comparison of K-Medoids Method and Analytical Hierarchy Clustering on Students' Data Grouping," Int. J. Informatics Vis., vol. 7, no. 2, pp. 446–454, 2023, doi: 10.30630/joiv.7.2.1204.

[25] I. M. D. Pradipta, A. Eka, A. Wahyudi, and S. Aryani, "Fuzzy C-Means Clustering for Customer Segmentation," Int. J. Eng. Emerg. Technol., vol. 3, no. 1, pp. 18–22, 2018, [Online]. Available: https://ojs.unud.ac.id/index.php/ijeet/article/download/41251/25103.

[26] T. Wahyuningrum, S. Khomsah, S. Suyanto, S. Meliana, P. E. Yunanto, and W. F. Al Maki, "Improving Clustering Method Performance Using K-Means, Mini Batch K-Means, BIRCH and Spectral," 2021 4th Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2021, pp. 206–210, 2021, doi: 10.1109/ISRITI54043.2021.9702823.

[27] D. Deepa, A. Sivasangari, R. Vignesh, N. Priyanka, J. Cruz Antony, and V. GowriManohari, "Segmentation of Shopping Mall Customers Using Clustering," pp. 619–629, 2023, doi: 10.1007/978-981-19-6004-8_48.

[28] Y. E. Wella, O. Okfalisa, F. Insani, F. Saeed, and A. R. C. Hussin, "Service quality dealer identification: the optimization of K-Means clustering," Sinergi (Indonesia), vol. 27, no. 3, pp. 433–442, 2023, doi: 10.22441/sinergi.2023.3.014.

[29] T. Juhari and A. Juarna, "Implementation Rfm Analysis Model for Customer Segmentation Using the K-Means Algorithm Case Study Xyz Online Bookstore," Explore, vol. 12, no. 1, p. 107, 2022, doi: 10.35200/explore.v12i1.548.