**Research in Education, Technology, and Multiculture**

ISSN: 3025-6763 | Journal's DOI: 10.61436/rietm

Homepage: https://imrecsjournal.com/journals/index.php/rietm

# CLV-Enhanced RFM Framework for Customer Segmentation in Indonesian SMEs Using K-Means Clustering

**Dio Rizkyandita[1], Heri Wijayanto[1], I Gde Putu Wirarama Wedashwara Wirawan[1], Mosiur Rahaman[2]**

[1]Master of Information Technology, University of Mataram, Mataram, Indonesia

[2]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand

**ARTICLE INFO**

***Article History:***

Received: 21 April 2026

Accepted: 23 May 2026

Published: 30 May 2026

***Keywords:***

K-Means;
Clustering;
RFM;
Customer Lifetime Value;
Customer Segmentation;
MSMEs.

***Corresponding Author:***

drizkyandita@gmail.com

**ABSTRACT**

Micro, Small, and Medium Enterprises (MSMEs) contribute more than 61% of Indonesia's Gross Domestic Product, yet most of them still face limitations in leveraging transactional data for customer retention strategies. Prior studies have extensively combined Recency, Frequency, and Monetary (RFM) analysis with the K-Means algorithm for customer segmentation, but the majority treat Customer Lifetime Value (CLV) only as a post-hoc label assigned to clusters after the clustering process is finalized, rather than as a feature that shapes the segment structure from the beginning. This study addresses three research questions: how CLV can be effectively integrated as a clustering input, what segmentation structure emerges from this approach, and what concrete retention strategies can be derived for MSMEs with limited analytical capabilities. The proposed framework incorporates CLV, calculated as the discounted historical net sales using a 10% annual discount rate, as the fourth feature in the K-Means feature space with K-Means++ initialization, alongside RFM variables standardized using Z-scores. The framework is applied to a real dataset from a coffee shop MSME in Indonesia, comprising 19,126 transactions, of which 4,310 are member transactions from 472 unique registered customers, recorded throughout January–December 2023. The optimal number of clusters is determined through the convergence of the Elbow Method and the Silhouette Coefficient, both indicating four clusters as the best solution with a silhouette score of 0.5105. The segmentation divided customers into four tiers: Platinum, Gold, Silver, and Bronze. A key finding was a highly concentrated value distribution, with just 1.48% of customers (n = 7) contributing 21.66% of total revenue and CLV. This pattern is significantly more skewed than the traditional 80/20 Pareto rule. This concentration is interpreted through the lenses of habit formation, small base amplification, and comparative empirical evidence. The four-tier framework translates into differentiated retention strategies: VIP retention, value uplift, repeat-purchase incentives, and win-back campaigns, with monetary thresholds calibrated to each segment's median value.

***To cite this article:***

Rizkyandita et al. (2026). CLV-Enhanced RFM Framework for Customer Segmentation in Indonesian SMEs Using K-Means Clustering. *Research in Education, Technology, and Multiculture*, *5*(1), 01-14. doi: https://doi.org/10.61436/rietm/v5i1.pp01-14



## ■ INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) serve as the backbone of the Indonesian economy. According to data from the Ministry of Cooperatives and Small and Medium Enterprises in 2024, there are approximately 65.5 million MSME units in Indonesia, contributing more than 61% to the national Gross Domestic Product (GDP), equivalent to IDR 9,580 trillion, while also absorbing around 97% of the national workforce. At the same time, the rapid growth of the Southeast Asian digital economy, which reached USD 90 billion in GMV in 2024, opens wider opportunities for Indonesian MSMEs to expand their customer base through digital channels. However, these opportunities come with challenges that are difficult to overcome. Today's consumers have greater access to information and more personalized expectations than ever before, making a "one-size-fits-all" marketing approach ineffective. This situation is further complicated by limited analytical resources and low data literacy

among MSMEs, where only around 33.6% of Indonesian MSMEs have undergone substantive digital transformation (Haddadi & Hamidi, 2025). The result is that, even though Southeast Asia's digital economy is an opportunity, most MSMEs have not been able to seize it to gain an advantage by turning data into measurable business decisions.

When MSMEs do manage to leverage their data, prioritizing customer retention over acquisition often yields the most significant business impact. One classic approach in marketing states that acquiring a new customer can cost five to twenty times more than retaining an existing one, and that a 5% increase in retention can lift profit by 25% to 95% (Reichheld & Sasser, 1990). This finding has become an important foundation in Customer Relationship Management (CRM), particularly for MSMEs operating under tight marketing budgets (Kumar & Reinartz, 2018). To manage retention effectively, MSMEs need to identify which customers hold high economic value, which are at risk of churning, and which fall in the mid-value range and could be upgraded to a higher tier. In this regard, Customer Lifetime Value (CLV) plays an important role as a metric measuring the long-term economic contribution of a customer to the business and has been widely used to allocate retention resources more efficiently (Gupta et al., 2004; Haddadi & Hamidi, 2025). Prior studies have also shown that CLV can serve as a strong strategic metric for informing business decisions related to customer retention and acquisition (Iqbal et al., 2024; Laksono et al., 2023).

In the customer segmentation literature, the K-Means algorithm has become one of the most popular methods due to its computational efficiency and ease of interpretation (Tabianan et al., 2022). Several studies have combined K-Means with RFM analysis to identify customer groups with similar behavioral patterns (ASLANTAŞ et al., 2023; Djun et al., 2024; Jamunadevi et al., 2021). However, conventional K-Means has a well-known limitation: its sensitivity to initial centroid placement, where random initialization can trap the algorithm in local minima and produce suboptimal clusters, particularly on datasets with skewed distributions such as retail transaction data (Sinaga & Yang, 2020). To address this issue, introduced the K-Means++ initialization scheme, which probabilistically selects initial centroids that are well spread across the feature space (Arthur & Vassilvitskii, 2007). Recent empirical studies have confirmed that K-Means++ yields more stable clusters and faster convergence in customer segmentation contexts, especially when the monetary variable follows a heavy-tailed distribution (Hicham & Karim, 2022; Tabianan et al., 2022; Wu et al., 2020). For these reasons, the K-Means++ initialization is adopted as the standard approach in this study.

Although the combination of Recency, Frequency, and Monetary (RFM) analysis with the K-Means algorithm has been extensively discussed in the context of customer segmentation (ASLANTAŞ et al., 2023; Djun et al., 2024; Ikotun et al., 2023; Jamunadevi et al., 2021) several limitations remain unresolved, especially in the Indonesian MSME context. First, most prior studies treat CLV as a post hoc label assigned to clusters after clustering is complete, rather than as a feature included from the outset, so that the economic value dimension does not directly shape the structure of the resulting segments. Second, the majority of customer segmentation studies in Indonesia are conducted on relatively small datasets, typically fewer than 1,000 customers, or rely on public datasets such as UCI Online Retail from the United Kingdom, which do not adequately represent the characteristics of Indonesian MSMEs. Third, only a few studies have conducted a sensitivity analysis of the discount rate used in CLV calculations, even though Gupta et al. (2004) showed that small changes in this parameter can substantially affect customer value estimation. Finally, the segmentation results from many studies have not been translated into an actionable framework that MSMEs lacking internal analytical capabilities can implement directly.

Based on these gaps, this study proposes an integrated customer segmentation framework that combines RFM analysis with CLV calculated as the discounted historical net sales used as the fourth feature in the K-Means feature space, with K-Means++ initialization. With this approach, the resulting segments reflect not only transactional patterns (recency, frequency, monetary value) but also the realized economic value each customer has contributed to the business. The framework is applied to a real dataset from MSME X, consisting of 19,126 transactions, of which 4,310 are member transactions belonging to 472 unique registered customers, recorded over a full one-year period (January–December 2023). The optimal number of clusters is determined by combining the Elbow Method and the Silhouette Coefficient. At the same time, Z-score standardization is applied to prevent variables with large ranges, such as Monetary and CLV, from dominating the clustering process. The segmentation results are then interpreted into four customer tiers

(Platinum, Gold, Silver, Bronze), each paired with a specific retention strategy. Conceptually, this approach builds on the theoretical foundations of RFM (Alves Gomes & Meisen, 2023), CLV (Gupta et al., 2004), and data mining for customer segmentation (Han et al., 2012; Khajvand & Tarokh, 2011), and extends the practice reported in prior Indonesian studies (Marisa et al., 2019; Matz & Hermawan, 2020).

This study is directed to answer the following three research questions:

RQ1. How can Customer Lifetime Value be effectively integrated as an input feature in the K-Means clustering process so that the resulting segments simultaneously reflect transactional behavior and the realized economic value of customers?

RQ2. What segmentation structure comes up when the CLV-enhanced RFM framework is applied to a real Indonesian MSME dataset, and what behavioral and economic characteristics distinguish each of the resulting segments?

RQ3. What concrete retention and loyalty strategies can be derived from the resulting segmentation that are feasible for MSMEs operating with limited analytical capabilities?

In this paper, much focus is placed on three main contributions. First, from a methodological standpoint, this study extends the classical RFM framework by integrating CLV as an input feature into K-Means, so that the dimension of realized economic value contributes to the formation of the segment structure from the very beginning of the clustering process. Second, from an empirical standpoint, the framework is validated on a representative Indonesian MSME dataset (4,310 member transactions from 472 unique customers over a full year), which is larger than most similar studies conducted in Indonesia, thereby providing stronger empirical evidence regarding MSME customer behavior. Third, from a practical standpoint, this study produces a four-tier framework (Platinum–Gold–Silver–Bronze) that MSMEs can directly adopt without internal analytical capability, while also revealing an extreme revenue concentration where only 1.48% of top customers (n = 7) contribute 21.66% of the revenue as a basis for designing more focused loyalty programs and retention strategies.

## ■ METHOD

In simple terms, Customer Relationship Management (CRM) is a business strategy aimed at maximizing profit, revenue, and customer satisfaction by organizing customer segments, sustaining actions that enhance satisfaction, and implementing customer-centric processes. CRM strategies are commonly developed around three key aspects: customer profitability, customer acquisition, and customer retention—based on the rationale that retaining existing customers is generally less costly than acquiring new ones. In this study, the CRM strategy focuses on customer retention to strengthen customer loyalty. Compared with traditional one-size-fits-all CRM policies, RFM-driven K-Means segmentation allows firms to align retention and acquisition strategies with empirically derived customer groups, thereby improving the effectiveness of CRM initiatives (Ling et al., 2024). The methodology is organized into four sub-sections following the standard reporting structure for empirical research: (1) Study Area / Materials describing the research object and data; (2) Research Design and Procedures laying out the sequential research stages; (3) Equipment and Parameters detailing the software and technical parameters used; and (4) Data Analysis explaining the data processing techniques and mathematical formulations applied.

### Study Area and Materials

This research object, MSME X, is a retail coffee shop business operating in Mataram City, West Nusa Tenggara Province, in the F&B retail product category. The selection of MSME X as a case study is based on three considerations: the Availability of data transaction recording through a Point of Sales (POS) system, the existence of a customer membership program that allows tracking of individual purchasing behavior, and lastly, the management's willingness to share data for academic research purposes while preserving confidentiality.

The primary materials used in this study consist of MSME X's sales transaction data covering the period from 1 January 2023 to 31 December 2023. The initial dataset comprises 19.126 transaction rows. After filtering to retail-only transactions identified under member customers, this yielded 4.319 member transactions originating from 472 unique customers. The available transaction attributes include date, time, gross sales, discount, net sales, total collect, and total amount.

### Research Design and Procedure

Based on the previous explanation, this research was designed as a retrospective observational study on historical transaction data. The retrospective design was appropriate because the dataset covered a full year (January–December 2023), which was already

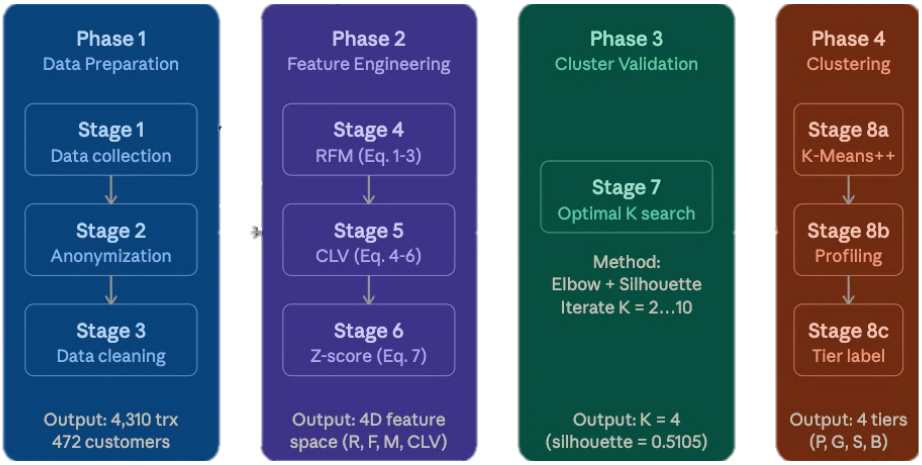


**Figure 1.** Research Procedure

complete by the time of the analysis. Therefore, all customer transactional behavior could be observed in its final form without further experimental intervention or manipulation. The overall research procedure was carried out through several sequential steps, starting from raw data acquisition to the creation of final cluster profiles, as illustrated in Figure 1.

This procedure begins with data collection, in which all transaction records are extracted from the Point-of-Sale (POS) cashier system of MSME X in spreadsheet format (.xlsx) for the entire observation period (January–December 2023). This raw export contains 19,126 rows of transaction data, along with attributes such as transaction date, time, gross sales, discounts, net sales, and total amount, as well as customer name, telephone number, and email if the transaction is a member.

After the raw data was obtained and filtered to only those containing customer identities, the next step was data anonymization. Since the original records, as previously described, contained information that could identify individuals, such as customer names, phone numbers, email addresses, and home addresses, all of these fields were removed from the dataset before any analysis was performed. Each customer was then assigned a pseudonymous identifier, an alphanumeric code, generated by applying a hash function to the original customer ID. The mapping between the original ID and the pseudonymous ID was stored only by the MSME X internal team. It was not accessible to any other party during the modeling process. The analysis was conducted solely at the level of transactional behavior, without revealing the identities of any individual customers, which is consistent with the principles of data privacy that should be applied to research using real consumer data. After the anonymization process, the data still contained inconsistencies introduced during the export process, requiring data cleansing steps. Three procedures were implemented: the decimal separator was converted to a period to ensure consistency in the numeric format, the thousands separator was removed to prevent parsing errors during calculations, and all monetary columns were converted to float data types to support further numeric operations.

With a clean dataset, the analysis moved on to feature construction. Three core variables were derived from transaction history: Recency (the number of days since the last transaction relative to the end of the period), Frequency (the total number of transactions), and Monetary (total spending). In addition to traditional RFM metrics, Customer Lifetime Value (CLV) was integrated as a fourth feature. CLV is calculated by discounting historical net sales at a 10% annual rate. This specific rate accurately reflects the level of barriers to entry for Indonesian MSMEs during the observation period, taking into account Bank Indonesia's benchmark interest rate along with local risk premiums (Gupta et al., 2004). Because these four variables operate at very different scales, a Z-score transformation was applied to standardize them. This scaling ensures that no single feature disproportionately dominates the Euclidean distance calculation during clustering.

Determining the optimal number of clusters involves evaluating standardized data across various k values (from 2 to 10). A combined approach using the Elbow Method and the Silhouette Coefficient is used to find the best configuration. The optimal k value is identified at the point where the within-cluster sum of squares (WCSS) inflection aligns with a robust internal validation metric, ensuring

compact, well-separated segments.

The procedure concludes with clustering and profiling. With the optimal k value determined in the previous step, K-Means is run with K-Means++ initialization to obtain the final cluster assignments. Each cluster is then profiled based on its median R, F, M, and CLV values, the number of members it contains, and the members' contributions to total revenue and total CLV. Finally, the clusters are ranked in descending order by median CLV, and each cluster is labeled with a tier name (Platinum, Gold, Silver, or Bronze) that reflects its relative economic value.

## Equipment and Parameters

All computational procedures were implemented in Python 3.10, using the open-source scikit-learn framework (version 1.3) as the primary engine for K-Means clustering and silhouette coefficient calculations. Supporting libraries include pandas for data manipulation, NumPy for numerical operations, and Matplotlib and Seaborn for visualization (Pedregosa et al., 2012).

The technical parameters applied in this experiment are presented in Table 1. The K-Means parameters used the K-Means++ initialization (sklearn default) to reduce the risk of getting stuck in local optima, with the number of reinitializations (n_init) set to 10 and the maximum number of iterations (max_iter) set to 300. The random_state parameter was set to 42 to ensure reproducibility of the results.

**Table 1.** Parameters

| Parameter | Value |
|---|---|
| Standardization method | Z-score |
| Algorithm | K-Means |
| Initialization (init) | k-means++ |
| Re-initialization (n_init) | 10 |
| Maximum iteration (max_iter) | 300 |
| Reproducibility (random_state) | 42 |
| Cluster range tested | k=2 - k=10 |
| Optimal k selection | Elbow + Silhouette |
| Software | Python 3.10, scikit-learn 1.3 |

The validity and reliability of the measurements in this study were maintained through objective data extraction and standardized evaluation metrics. The Recency, Frequency, and Monetary (RFM) variables were calculated directly from historical transaction records, rather than subjective instruments, resulting in strong construct validity. For the economic dimension, Customer Lifetime Value (CLV) was derived using the standard Net Present Value (NPV) formulation widely accepted in the customer assessment literature. To assess clustering quality, the Silhouette Coefficient was applied as a measure of internal validity. Following a standard interpretation threshold in the range [-1, 1], a score above 0.5 is considered a reasonable cluster structure, while a score above 0.7 indicates a strong cluster structure.

## Data Analysis

In this analysis stage, the data consist of RFM, CLV, standardization, and cluster-validation variables. The following explains the mathematical formulations for each stage (Valentini et al., 2024).

*RFM (Recency, Frequency, Monetary)*

After cleaning the data, RFM analysis was conducted through the following steps:

a. Recency calculation: Measuring the time interval since each customer's most recent transaction.

In this study, Recency ($\mathcal{R}_i$) is measured in days. The reference date $\mathcal{T}_{\mathrm{ref}}$ is set to 31 December 2023 (the end of the observation period), and $T_{i,last}$ is the date of customer i's most recent transaction within the period. For example, a customer whose last transaction occurred on 1 December 2023 has $R_i = 30$ days. Day level granularity was chosen to align with typical retail coffee purchasing cycles in Indonesian MSMEs and to provide sufficient resolution for cluster differentiation.

$$\mathcal{R}_i = \mathcal{T}_{ref} - \mathcal{T}_{i,last}$$

b. Frequency calculation: Counting the number of transactions made by each customer.

$$F_i = \sum_{j=1}^{n} trans_{i,j}$$

Where $trans_{i,j}$ represents the number of unique transactions (transaction IDs) associated with customer i within a given period.

c. Monetary calculation: Summing the total amount of money spent by each customer.

$$M_i = \sum_{j=1}^{n} v_{i,j}$$

Where $v_{i,j}$ denotes the value of the j-th transaction made by customer i. The RFM combination serves as a fundamental basis for measuring the economic value of customers.

*CLV (Customer Lifetime Value)*

CLV is computed as the discounted historical value of each customer's Net Sales within the observation period. The latest transaction date in the dataset is used as the reference date, and the annual discount rate is converted to a monthly rate (Kanchanapoom &

Chongwatpol, 2023). Each transaction is discounted to the reference date based on the difference in months between the transaction date and the reference date, and the customer's CLV is obtained by summing these discounted Net Sales across all transactions. The resulting CLV is treated as a numeric feature and used alongside the RFM variables in subsequent preprocessing and clustering stages (Gupta et al., 2006; Marisa et al., 2019; Valentini et al., 2024).

CLV is calculated using the following formulation:

$$CLV_i = \sum_{j=1}^{n_i} \frac{NS_{ij}}{(1 + r_m)^{mij}}$$

$CLV_i$ = the CLV of customer I (discounted historical value),
$NS_{ij}$ = the Net Sales value of the j transaction made by customer I,
$r_m$ = the monthly discount rate,
$t_0$ = the reference date (set to max{Date}) in the dataset).
*Mij* the difference in months between $t_0$ and the transaction date $t_{ij}$, computed as follows:

$$m_{ij} = 12\left(Y_0 - Y_{ij}\right) + \left(M_0 - M_{ij}\right)$$

where Y denotes the year and M denotes the month component of the date. Monthly discount rate calculation:

$$r_m = (1 + r_a)^{\frac{1}{12}} - 1$$

For each customer i with a set of transactions indexed by j $n_i$, CLV is computed as the sum of the present values (NPV) of historical Net Sales, discounted to the reference date $t_0$ (the most recent date in the dataset). This discount rate calculation uses an annual discount rate ($r_a$) of 10%, representing the typical hurdle rate for MSMEs in Indonesia during the observation period. This figure is based on two main components that can be verified empirically.

To determine the appropriate discount rate for estimating Customer Lifetime Value (CLV), this study adopted a baseline target of 10%. This rate falls within the 8%-15% range (Gupta et al., 2004). for the retail and services context. Achieving this 10% total discount rate involves two components: the risk-free rate and the specific risk premium. The risk-free rate depends on Bank Indonesia's benchmark interest rate (the BI 7-Day Reverse Repo Rate or BI7DRR) during the observation period from January to December 2023. According to official data, the BI7DRR was 5.75% through September, then increased to 6.00% in the last quarter, resulting in an annual average of 5.81%.

**Table 2.** BI 7-Day Reverse Repo Rate 2023

| No | Date | BI-7Day-RR |
|---|---|---|
| 1 | 21 December 2023 | 6.00 % |
| 2 | 23 November 2023 | 6.00 % |
| 3 | 19 October 2023 | 6.00 % |
| 4 | 21 September 2023 | 5.75 % |
| 5 | 24 August 2023 | 5.75 % |
| 6 | 25 July 2023 | 5.75 % |
| 7 | 22 June 2023 | 5.75 % |
| 8 | 25 May 2023 | 5.75 % |
| 9 | 18 April 2023 | 5.75 % |
| 10 | 16 March 2023 | 5.75 % |
| 11 | 16 February 2023 | 5.75 % |
| 12 | 19 January 2023 | 5.75 % |

Since the target discount rate is 10% and the risk-free cost of capital is 5.81%, the required risk premium is 4.19%. The calculation is formulated as follows:

$$r_a = r_{risk-free} + r_{premium} = 5{,}81\% + 4{,}19\% = 10\%$$

Conceptually, this 4.19% premium addition is highly relevant to accommodate the distinct operational constraints of MSMEs. Compared to larger companies, MSMEs inherently face a higher overall cost of capital driven by greater earnings volatility, limited access to formal financial markets, and higher exposure to local economic shocks.

*Standardization Z-Score*

Z-score standardization is a data transformation technique that scales each numeric feature to a comparable range by centering values around the mean and adjusting their spread using the standard deviation. The Z-Score Standardization method chosen in this study is based on the characteristics of the K-Means algorithm, which relies on Euclidean distance calculations. Unlike the Min-Max Scaler, which maps all data to a fixed range [0, 1], the Z-Score transform preserves the relative distribution using the mean and standard deviation.

In RFM analysis, extreme values (outliers) in Monetary variables often represent high-value customers, which are crucial for calculating Customer Lifetime Value (CLV). Using the Min-Max Scaler on data containing extreme outliers will cause most of the data to be concentrated in a very narrow range, thereby reducing the K-Means algorithm's ability to distinguish between clusters. In contrast, the Z-Score mitigates the impact of outliers without

eliminating them, allowing for the formation of clusters that are more representative of diverse customer profiles (Han et al., 2012).

Mathematically, Z-score standardization for a feature x is defined as:

$$x_i = \frac{x_i - \mu_x}{\sigma_x}$$

where $x_i$ is the standardized value, $\mu_x$ is the mean of feature x, and $\sigma_x$ is the standard deviation of feature x. This transformation produces features with a mean close to 0 and a standard deviation close to 1, allowing each variable to contribute more evenly to the clustering process.

*Elbow Method*

One of the main challenges in applying K-Means is determining the optimal number of clusters (k) (Marisa et al., 2019). If the number of clusters (k) is set too low, each cluster becomes highly heterogeneous, reducing detail and making the resulting segments harder to interpret. Conversely, if k is too large, the segments can become overly specific and difficult to manage in practice. Therefore, in this study, as in many prior works, the Elbow Method is used as a primary approach for selecting k. The Elbow Method works as follows:

a. Definition of SSE

The Elbow Method is based on the Sum of Squared Errors (SSE), defined as the sum of the squared distances from each data point to the centroid of its assigned cluster.

$$SSE = \sum_{j=1}^{k} \sum_{x_i \epsilon C_j} \|x_i - C_j\| 2$$

A lower SSE indicates that data points within each cluster are more homogeneous. Plotting SSE against k: In practice, the K-Means algorithm is executed repeatedly for different values of k (e.g., k=2 to k=9 or k=10).Each value of k produces a different SSE value. As a result, a curve is obtained by plotting k on the x-axis and SSE on the y-axis (Marisa et al., 2019). The next step is to identify the "elbow point," where the rate of SSE reduction begins to slow down noticeably. The point at which SSE drops most sharply, after which the curve becomes relatively flatter, is referred to as the "elbow." (Widhyastuti et al., 2022; Perdana et al., 2022). The value of k at this point is typically selected as the optimal number of clusters because it offers the best balance between within-cluster heterogeneity and over-segmentation (Perdana et al., 2022).

a. K-Means Clustering

The K-Means algorithm is one of the most widely used clustering methods in data analysis(Li et al., 2021; Sinaga & Yang, 2020). It partitions data into clusters based on the proximity (distance) between observations (Perdana et al., 2022). The main steps of the K-Means algorithm begin by determining the optimal number of clusters (k) using the Elbow Method and the Silhouette Coefficient. Next,

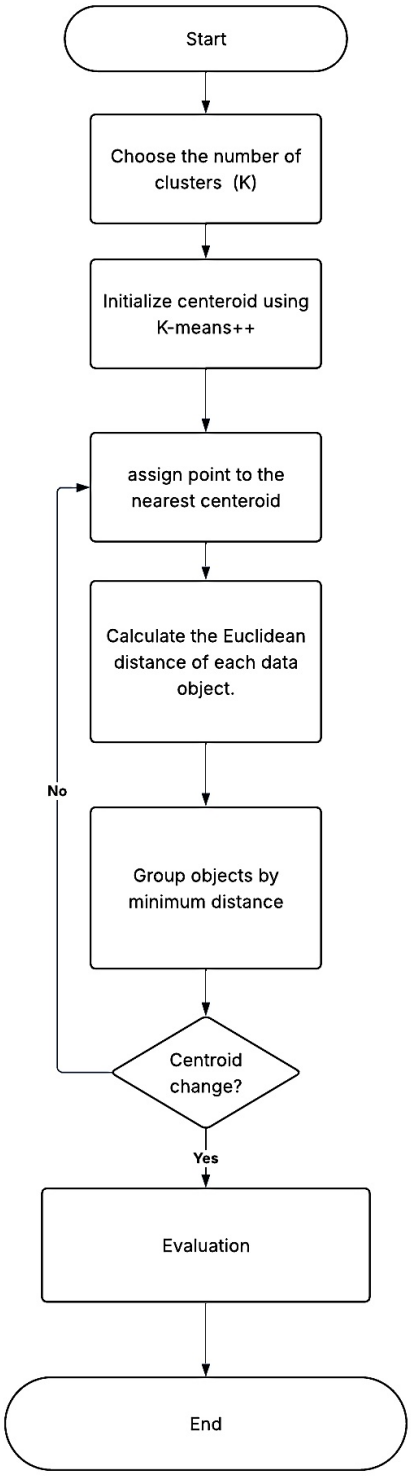


Figure 2. Flowchart K-Means++

the centroids are initialized using the K-Means++ approach. (Arthur & Vassilvitskii, 2007). Each data point is then assigned to the nearest cluster based on Euclidean distance, and the centroid is recalculated repeatedly until convergence is reached (Jamunadevi et al., 2021). To select the optimal number of clusters, the Elbow method is used to evaluate the Sum of Squared Errors (SSE) across different values of k (Gustriansyah et al., 2019; Nugroho & Adhinata, 2022).

At the centroid initiation stage, K-Means++ is used because one of the long-known limitations of conventional K-Means is its sensitivity to the choice of initial centroids. Random initialization can select centroids from adjacent regions, increasing the risk of getting stuck in a local optimum. To overcome this problem, K-Means++ was introduced by Arthur & Vassilvitskii (2007) as a probabilistic initialization scheme. The first centroid is selected randomly, then each subsequent centroid is selected with a probability proportional to the square of its distance from the previously selected centroid. This approach produces initial centroids that tend to be far apart, resulting in faster convergence and more stable clustering results, with a theoretical guarantee of O(log(K)) for an optimal solution(Fader & Hardie, 2009).

Recent empirical studies confirm the advantage of K-Means++ in customer segmentation contexts. (Tabianan et al., 2022) demonstrated that K-Means++ combined with RFM analysis on e-commerce data produces more stable clusters across replications with different random seeds. Similar findings are reported by Hicham & Karim (2022) and Wu et al. (2020) on customer value clustering with heavy-tailed monetary distributions.

a. Silhouette Coefficient (SC)

The Silhouette Coefficient (SC) is an internal cluster validity measure used to evaluate how well an object is assigned to its own cluster compared with other clusters. SC is employed because it simultaneously captures cohesion (how close an observation is to other members within the same cluster) and separation (how far it is from the nearest neighboring cluster). (ASLANTAŞ et al., 2023; Matz & Hermawan, 2020). The general formula of the Silhouette Coefficient for a data point i is defined as follows:

$$s(i) = \frac{[b(i)-a(i)]}{max\{a(i),b(i)\}}$$

a(i) = the average distance between observation i and all other observations within its own cluster. b(i) = the average distance between observation i and all observations in the nearest cluster to which i does not belong.

The Silhouette value s(i) lies in the range [-1, 1]: s(i) close to +1 indicates that the observation is well matched to its own cluster and far from other clusters. s(i) close to 0 suggests that the observation lies near the boundary between two clusters. s(i) close to −1 indicates that the observation may have been assigned to the wrong cluster, as it is closer to another cluster than to its own.

In this study, the clustering quality of K-Means segmentation based on RFM is assessed using the Silhouette Coefficient. This internal validity index simultaneously measures intra-cluster cohesion and inter-cluster separation on a scale from −1 to 1, where higher values indicate more clearly defined clusters (Rhomadhona et al., 2025).

## ■ RESULT AND DISCUSSION

### Data Preparation and Feature Construction

The data used in this study were obtained from UMKM X and consist of 4.310 member transactions belonging to 472 unique registered customers, recorded throughout January–December 2023. The dataset includes multiple transaction attributes, such as transaction date, transaction amount, and customer ID. Table 3 shows a sample of the cleaned transaction records.

The integration of CLV as an input feature, rather than as a post hoc label assigned after clustering, is a key methodological choice that distinguishes this study from prior work. By placing CLV alongside the RFM features in the same feature space, the K-Means algorithm forms clusters that are simultaneously similar in transactional behavior (R, F, M) and economically coherent (CLV). This directly addresses the post hoc weakness, where two customers with similar RFM profiles but very different CLV values could end up in the same segment when CLV is computed only after clustering is finalized.

### Z-Score Standardization

Because the four features have very different scales, recency ranges from 0 to over 350 days, frequency from 1 to 301 transactions, while Monetary and CLV span millions of rupiah. Direct application of K-Means with Euclidean distance would allow large-magnitude features to dominate the distance calculation. Z-score standardization transforms all features to have $\mu = 0$ and $\sigma = 1$, allowing each feature to contribute proportionally to the distance computation. After standardization, all four features have a mean ≈ of 0 and a standard deviation ≈ of 1, satisfying the assumptions of

**Table 3.** Sample of RFM-CLV computation per customer

| Customer | Recency | Frequency | Monetary | CLV |
|---|---|---|---|---|
| CUST_001 | 7 | 5 | 232200 | 221978,82 |
| CUST_002 | 120 | 11 | 566100 | 544913,04 |
| CUST_003 | 83 | 11 | 973600 | 934191,25 |
| CUST_004 | 122 | 2 | 180900 | 170528,11 |
| CUST_005 | 48 | 14 | 1783200 | 1668944,57 |
| CUST_006 | 16 | 89 | 3641900 | 3461998,58 |
| CUST_007 | 185 | 4 | 899700 | 857830,29 |
| CUST_008 | 132 | 3 | 315000 | 303281,09 |
| CUST_009 | 191 | 1 | 55080 | 52516,72 |
| CUST_010 | 84 | 13 | 491400 | 460821,46 |

distance-based clustering. The Z-score is preferred over Min-Max scaling in this context because it preserves the heavy-tailed signals in Monetary and CLV values that represent the highest-contributing customers, rather than compressing noise.

**Optimal Cluster Number Determination**

To determine the optimal number of clusters k, K-Means with K-Means++ initialization was executed for k=2 through k=10, and both the Sum of Squared Errors (SSE) and the Silhouette Coefficient were computed for each run. Figure 3 shows the Elbow plot.

Two observations emerge from these plots. First, the Elbow curve shows a clear inflection point at k=4, after which the rate of SSE reduction slows substantially, indicating that adding more clusters yields diminishing returns in within-cluster cohesion.

The Silhouette score of 0.5105 indicates that the cluster structure is reasonably well formed, yet some degree of overlap remains. Three computational factors explain this overlap. First, the boundary between Silver (Cluster 0) and Bronze (Cluster 1) is naturally fuzzy because both segments share low Frequency values (1–4 transactions), with the main differentiator being Recency. Customers near the threshold, for example, those with Recency around 100–150 days and Frequency = 2–3, sit close to the decision boundary and contribute to lower silhouette values for these clusters. Second, the heavy-tailed distribution of Monetary and CLV values means that a small number of customers are far from any cluster centroid, thereby lowering the average silhouette score. Third, K-Means assumes spherical and equally sized clusters, which does not perfectly fit retail transaction data, where some clusters (Silver, Bronze) are large and dense, while others (Platinum, Gold) are small and dispersed. Despite this overlap, the score of 0.5105 is consistent with similar studies in retail clustering: Wu et al. (2020) reported 0.47, and Tabianan et al. (2022) reported scores in the 0.48–0.52 range, suggesting the result is within the acceptable range for real-world transactional data with skewed distributions.

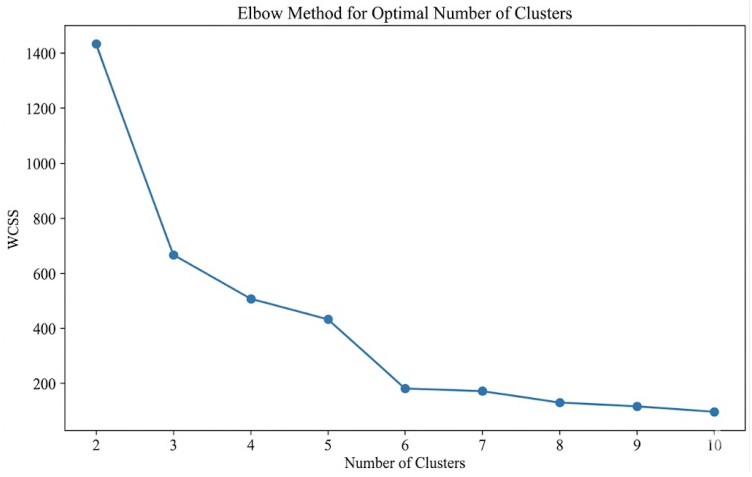


**Figure 3.** Elbow Method Result

**Table 4.** Silhouette Coefficient Result

| Silhouette Score per K: | |
|---|---|
| **Number of K** | **Silhouette Score** |
| K=2: | 0.4045 |
| K=3: | 0.4851 |
| K=4: | 0.5105 |
| K=5: | 0.4281 |
| K=6: | 0.4763 |
| K=7: | 0.4724 |
| K=8: | 0.4761 |
| K=9: | 0.4764 |
| K=10: | 0.4283 |

## Customer Segmentation Result and Cluster Profiling

After running K-Means++ on the standardized feature space, the algorithm produced four well-separated clusters. Each cluster was then profiled using its median R, F, M, CLV, member count, and contributions to total revenue and CLV. Clusters were ranked by median CLV in descending order and assigned the labels Platinum, Gold, Silver, and Bronze. The complete profile is presented in Table 5.

Three patterns shown from the cluster profiling. First, the relationship between Recency and Frequency is sharply nonlinear: moving from Bronze (R=231.5, F=1) to Platinum (R=14, F=95) represents a 16x reduction in recency paired with a 95x increase in frequency. This indicates that customer activation is not gradual but discrete; once a customer crosses a threshold of repeat purchasing, their behavior shifts qualitatively. Second, the Monetary-to-Frequency ratio increases across tiers (Bronze: Rp 136k/ transaction, Silver: Rp 81.6k/transaction, Gold: Rp 77.1k/transaction, Platinum: Rp 80.5k/ transaction), suggesting that high-tier customers do not necessarily spend more per transaction but rather transact more frequently. Third, the CLV-to-Monetary ratio remains stable at around 96-97% across all clusters, indicating that the discounted historical value is dominated by recent transactions across all customer types, a reasonable outcome given the 12-month observation window.

The data revealed that a small percentage of customers, just 1.48% (n=7), accounted for 21.66% of total revenue. While this finding reflects the well-known Pareto principle (80/20 rule) often observed in the retail sector, the distribution in this study is much more skewed. Instead of the traditional 80/20 ratio, the revenue concentration here appears sharper, closer to a 22/0.16 pattern. Three main reasons could explain this highly concentrated purchasing behavior. First, in MSME retail with a small registered customer base (472 unique members), reseller behavior may inflate the Platinum tier. These are likely repeat bulk buyers who serve as informal distribution channels, a pattern documented in studies of Indonesian SMEs (Sinaga & Yang, 2020; Wood & Rünger, 2016). Second, the Indonesian MSME consumer market exhibits strong relational loyalty; once trust is established, customers tend to consolidate purchases with a single vendor rather than diversify, thereby amplifying the dominance of top customers. Third, the heavy-tailed distribution is consistent with the broader economic literature on customer value, including the Pareto/NBD model (Reinartz et al., 2000; Schmittlein et al., 1987), where a small subset of customers consistently drives disproportionate value. The practical implication is significant: losing even one Platinum customer translates to losing ~3% of total annual revenue, making retention investment in this tier yield an extraordinarily high ROI compared to mass-market acquisition campaigns.

The Bronze segment, with a median Recency of 231.5 days, requires interpretation in light of industry-specific repurchase patterns. For coffee shop MSMEs, the typical repurchase

**Table 5.** Profile Cluster

| Cluster | Segment | n | R | F | M | CLV | Revenue share | CLV share |
|---|---|---|---|---|---|---|---|---|
| 0 | Silver | 266 | 53.0 | 4.0 | Rp326.600 | Rp315.423 | 39.43% | 39.72% |
| 1 | Bronze | 172 | 231.5 | 1.0 | Rp136.350 | Rp126.552 | 12.13% | 11.90% |
| 2 | Gold | 27 | 26.0 | 39.0 | Rp3.006.150 | Rp2.844.900 | 26.77% | 26.72% |
| 3 | Platinum | 7 | 14.0 | 95.0 | Rp7.654.140 | Rp7.361.103 | 21.66% | 21.66% |

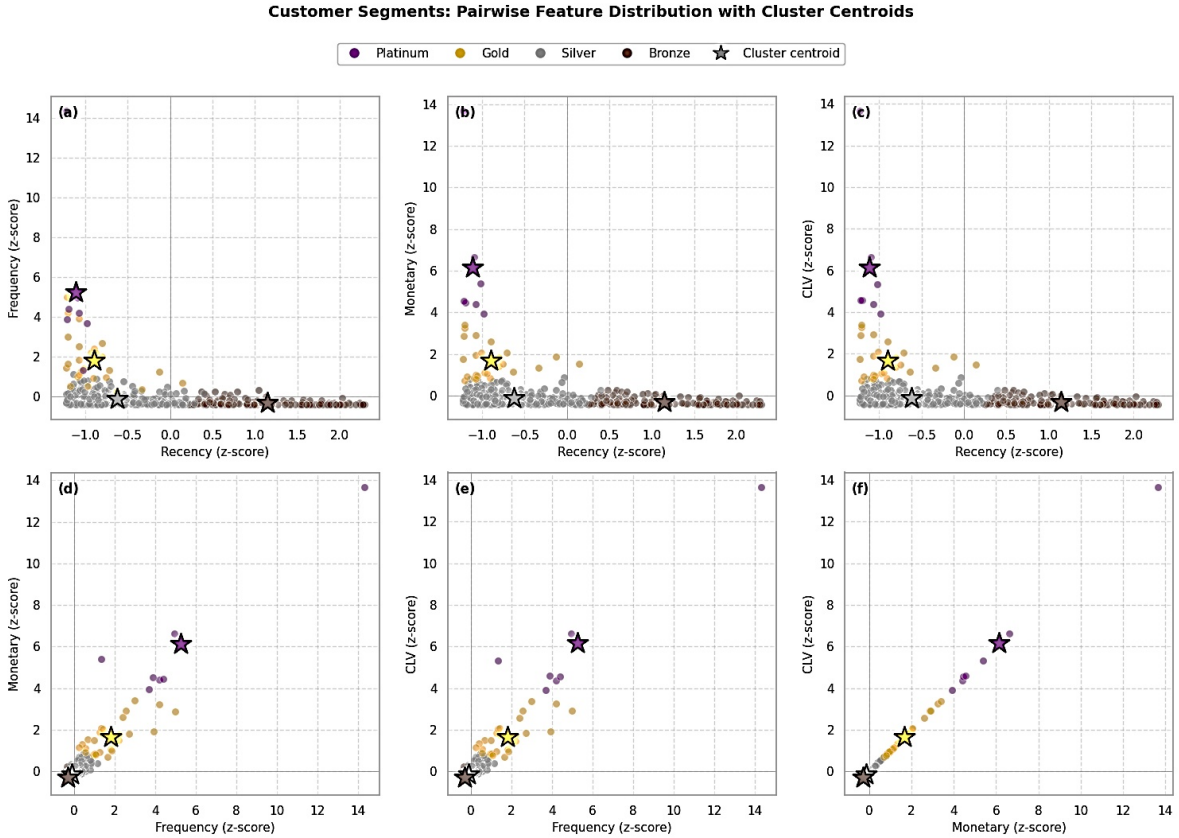


**Figure 4.** Feature Distribution Customer Segment

cycle for active customers ranges from 7 to 30 days. A Recency value exceeding 200 days substantially exceeds this normal range, providing a defensible basis for classifying these customers as at risk of churn. The classification threshold is calibrated to the coffee shop context; for product categories with longer repurchase cycles, such as electronics, furniture, or high-end apparel, a Recency value of 200+ days does not necessarily indicate churn. In this dataset, the gap between Bronze (Recency 231.5 days) and Silver (Recency 53 days) is approximately fourfold, supporting the at-risk classification with reasonable confidence.

### Strategic Implications for MSME Retention

The four-segment structure provides a basis for tier-specific retention strategies. Each strategy below is grounded in the empirical median Monetary value of its corresponding tier, with thresholds and incentives calibrated to actual customer spending behavior.

## ■ CONCLUSION

This study aims to integrate Customer Lifetime Value (CLV) into the traditional RFM segmentation framework for Indonesian MSMEs, with a specific focus on the coffee shop sector in Mataram City. Rather than treating CLV as an additional label after the process is complete, this study demonstrates that CLV can be effectively integrated as a core input feature in the K-Means clustering process. By applying a 10% annual discount rate derived from Bank Indonesia benchmarks, combined with MSME risk premiums, this expanded framework successfully captures both

**Table 6.** Specific Retention Strategies

| Cluster | Monetery | Action | Concrete |
|---|---|---|---|
| Platinum | Rp7.654.140 | Personal account manager program. Target: Maintain retention >95%. | Exclusive invitations to annual events, premium product bundles with a minimum spend of Rp. 8,000,000 (at least 5% above the mean), early access 30 days before the product's public launch. |
| Gold | Rp3.006.150 | Value uplift via cross-selling. Target: Drive 20% of Gold members to Platinum within 12 months. | Tier up rewards for spending exceeding IDR 4,500,000 (50% above the mean); tiered cashback for transactions exceeding IDR 750,000. |

| | | | |
|---|---|---|---|
| Silver | Rp326.600 | Repeat purchase incentive. Target: Increase Average transactions are from 4 to 8 per year. | 10% discount program with a minimum spend of Rp 400,000 (22% above the mean) according to The reviewer's suggestion: "buy 3 get 1" coupon valid for 30 days. |
| Bronze | Rp136.350 | Win back campaign. Target: Reactivate 15% of Bronze within 90 days. | Reactivation voucher worth Rp 50,000 with a minimum spend of Rp 200,000 (47% above the mean), valid for 14 days only to create urgency. |

transactional behavior and realized economic value simultaneously during segment formation.

When validated on a real-world dataset of over 4,000 transactions from nearly 500 registered customers over a full year, this approach yields a robust four-tier segmentation structure: Platinum, Gold, Silver, and Bronze. The optimal formation of these four clusters is confirmed through the convergence of the Elbow Method and the Silhouette Coefficient. A striking empirical finding from this distribution is the extreme concentration of value in the MSME context. Specifically, only 1.48% of the customer base accounts for nearly 22% of total revenue and CLV, depicting a Pareto-style pattern that is much more skewed than traditional large-scale retail observations.

By translating these insights into practical applications, the resulting framework provides highly actionable, differentiated retention strategies tailored to the empirical median monetary value of each tier. MSMEs can immediately implement VIP retention initiatives for Platinum members, value-enhancing and cross-selling tactics for Gold members, repeat-purchase incentives for Silver members, and targeted customer-return campaigns for Bronze members. As digital adoption continues to increase among Indonesian MSMEs, the availability of POS data and digital payments will further facilitate data-driven Customer Relationship Management. Ultimately, by combining an easy-to-understand methodology with accountable parameters and concrete thresholds, this proposed model equips MSMEs lacking dedicated analytics teams to translate raw transaction data into measurable retention strategies, laying the groundwork for immediate business applications and future academic refinements.

## ■ DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

During the drafting of this manuscript, the author(s) utilized [Claude by Anthropic, ChatGPT by OpenAI, and Gemini by Google] for the purpose to improve readability and correcting English grammar. Following the use of this tool, the author(s)reviewed and revised the content as necessary and accept full responsibilityfor the final content of the article.

## ■ REFERENCES


Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and E-Business Management*, *21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07) (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Aslantaş, G., Gençgül, M., Rumelli, M., Özsaraç, M., & Bakırlı, G. (2023). Customer segmentation using K-means clustering algorithm and RFM model. *Deu Muhendislik Fakultesi Fen ve Mühendislik, 25*(74), 491–503. https://doi.org/10.21205/deufmd.2023257418

Bank Indonesia. (2024). BI 7-day reverse repo rate (BI7DRR). Bank Indonesia. https://www.bi.go.id/en/statistik/indikator/bi-7day-rr.aspx

Djun, S. F., Gunadi, I. G. A., & Sariyasa, S. (2024). *Analisis Segmentasi Pelanggan pada Bisnis dengan Menggunakan Metode K-Means Clustering pada Model Data RFM* [Customer Segmentation Analysis in Business Using the K-Means Clustering Method on the RFM Data Model]. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *5*(4), 354–364. https://doi.org/10.35746/jtim.v5i4.434

Fader, P. S., & Hardie, B. G. S. (2009). Probability Models for Customer-Base Analysis. *Journal of Interactive Marketing*, *23*(1), 61–69. https://doi.org/10.1016/j.intmar.2008.11.003

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling Customer Lifetime Value. *Journal of Service Research*, *9*(2), 139–155. https://doi.org/10.1177/1094670506293810

Gupta, S., Lehmann, D. R., & Stuart, J. A.

(2004). Valuing Customers. *Journal of Marketing Research*, *41*(1), 7–18. https://doi.org/10.1509/jmkr.41.1.7.25084

Gustriansyah, R., Suhandi, N., & Antony, F. (2019). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Haddadi, A. M., & Hamidi, H. (2025). A hybrid model for improving customer lifetime value prediction using stacking ensemble learning algorithm. Computers in Human Behavior Reports, 18, 100616. https://doi.org/10.1016/j.chbr.2025.100616

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Hicham, N., & Karim, S. (2022). Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering. International Journal of Advanced Computer Science and Applications, 13(10), 514–523. https://doi.org/10.14569/IJACSA.2022.0131016

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences, 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Iqbal, N. M., Iskandar, Y. A., & Zulvia, F. E. (2024). Customer Segmentation Using the K-means Clustering Algorithm and Recency Frequency Monetary Model at Pharmaceutical Product Wholesaler. *International Journal of Research in Vocational Studies*, *4*(2), 53–60. https://doi.org/10.53893/ijrvocas.v4i2.293

Jamunadevi, C., Tamil Selvan, S., Govindarajan, M., Saravanan, C., & Janaki Raman, B. R. (2021). LRFM model for customer purchase behaviour using K-Means algorithm. *IOP Conference Series: Materials Science and Engineering*, *1055*(1), 012111. https://doi.org/10.1088/1757-899x/1055/1/012111

Kanchanapoom, K., & Chongwatpol, J. (2023). Integrated customer lifetime value (CLV) and customer migration model to improve customer segmentation. *Journal of Marketing Analytics*, *11*(2), 172–185. https://doi.org/10.1057/s41270-022-00158-7

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science*, *3*, 1327–1332. https://doi.org/10.1016/j.procs.2011.01.011

Kumar, V., & Reinartz, W. (2018). Customer relationship management: Concept, strategy, and tools (3rd ed.). Springer. https://doi.org/10.1007/978-3-662-55381-7

Laksono, F. A., Rachmat, B., & Sutarso, Y. (2023). B2B customer segmentation based on customer lifetime value concept and RFM modeling. International Journal of Economics Development Research, 4(3), 1185–1199. https://doi.org/10.37385/ijedr.v4i3.2952

Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. Applied Soft Computing, 113, 107924. https://doi.org/10.1016/j.asoc.2021.107924

Ling, S. S., Too, C. W., Wong, W. Y., & Hoo, M. H. (2024). Customer Relationship Management System for Retail Stores Using Unsupervised Clustering Algorithms with RFM Modeling for Customer Segmentation. *2024 IEEE 14th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 1–6. https://doi.org/10.1109/ISCAIE61308.2024.10576353

Widhyastuti, L. P. W., Sukajaya, I. N., & Aryanto, K. Y. E. (2022). *Customer profiling berdasarkan model RFM dengan metode K-Means pada institusi pendidikan untuk menunjang strategi bisnis di masa pandemi Covid-19* [Customer profiling based on the RFM model with the K-Means method in educational institutions to support business strategies during the Covid-19 pandemic]. JTIM: Jurnal Teknologi Informasi dan Multimedia, 4(2), 94–108. https://doi.org/10.35746/jtim.v4i2.232

Marisa, F., Ahmad, S. S. S., Yusof, Z. I. M., Fachrudin, & Aziz, T. M. A. (2019). Segmentation model of customer lifetime value in small and medium enterprise (SMEs) using K-Means clustering and LRFM model. *International Journal of Integrated Engineering*, *11*(3), 169–180. https://doi.org/10.30880/ijie.2019.11.03.018

Matz, A., & Hermawan, A. T. (2020). Customer

loyalty clustering model using K-Means algorithm with LRIFMQ parameters. Jurnal INFORM, 5(2), 54–61. https://doi.org/10.25139/inform.v0i1.2691

Nugroho, N., & Adhinata, F. (2022). Penggunaan Metode K-Means dan K-Means++ Sebagai Clustering Data Covid-19 di Pulau Jawa. *Teknika*, *11*, 170–179. https://doi.org/10.34148/teknika.v11i3.502

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Perdana, S. A., Florentin, S. F., & Santoso, A. (2022). *Analisis segmentasi pelanggan menggunakan K-Means clustering studi kasus aplikasi Alfagift* [Customer segmentation analysis using K-Means clustering case study of Alfagift application]. *Sebatik*, *26*(2), 420–427. https://doi.org/10.46984/sebatik.v26i2.2134

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: quality comes to services. *Harvard Business Review*, *68*(5), 105—111. http://europepmc.org/abstract/MED/10107082

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. Journal of Marketing, 64(4), 17–35. https://doi.org/10.1509/jmkg.64.4.17.18077

Rhomadhona, H., Kusrini, W., Aprianti, W., & Permadi, J. (2025). Implementation of K-Means Clustering for Social Assistance Recipients with Silhouette Score Evaluation. *Brilliance: Research of Artificial Intelligence*, *5*(1), 136–143. https://doi.org/10.47709/brilliance.v5i1.5900

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? Management Science, 33(1), 1–24. https://doi.org/10.1287/mnsc.33.1.1

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243

Valentini, T., Roederer, C., & Castéran, H. (2024). From redesign to revenue: Measuring the effects of servicescape remodeling on customer lifetime value. Journal of Retailing and Consumer Services, 77, 103681. https://doi.org/10.1016/j.jretconser.2023.103681

Wood, W., & Rünger, D. (2016). Psychology of habit. Annual Review of Psychology, 67, 289–314. https://doi.org/10.1146/annurev-psych-122414-033417

Wu, J., Shi, L., Lin, W.-P., Tsai, B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering*, *2020*, 1–7. https://doi.org/10.1155/2020/8884227