Eduvest – Journal of Universal Studies
Volume 5 Number 2, February, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

LEVERAGING LRFM ANALYSIS AND SYNTHETIC DATA FOR CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

Muhibuddin, Erna Budhiarti Nababan, Fahmi
Data Science and Artificial Intelligence Program Universitas Sumatera Utara Medan, Indonesia
Faculty of Electrical Engineering Universitas Sumatera Utara Medan, Indonesia
Email: muhibuddinb@gmail.com, ernabrn@usu.ac.id, fahmimn@gmail.com

ABSTRACT

This research explores the use of synthetic data in Length Recency Frequency Monetary (LRFM) analysis and K-Means clustering for customer segmentation. It is challenging to access accurate and comprehensive customer data, this study generates synthetic data using Time-series Generative Adversarial Networks (TimeGAN) to supplement or replace original data. LRFM analysis is used to measure customer characteristics based on the dimensions of Length, Recency, Frequency, and Monetary, which are then applied to clustering using the K-Means algorithm. The quality of clustering is evaluated using the Silhouette Coefficient and Davies-Bouldin Index. The results show that the Silhouette Coefficient for synthetic data is 0.42, slightly higher compared to the original data which has a value of 0.41. Meanwhile, the Davies-Bouldin Index for synthetic data is 0.90, slightly higher than the original data which has a value of 0.89. This indicates that synthetic data can mimic the characteristics of real data without compromising the accuracy and quality of clustering. By combining synthetic data, LRFM analysis, and K-Means clustering, this research provides in-depth insights into customer segmentation. The findings are expected to help companies develop more effective marketing strategies, enhance customer retention, and optimize overall customer experience. This study asserts that synthetic data is a valid alternative to real data in customer analysis.

KEYWORDS *lrfm, timegan, kmeans clustering, segmentation*

INTRODUCTION

Companies need customer data to understand and analyze customer behavior, leading to effective marketing strategies, increased customer retention, and optimized customer experiences (Gul & Rehman, 2023). Customer segmentation, a technique often used, involves grouping customers with similar characteristics (Marisa et al., 2019; Tomašev & Radovanović, 2016). The Length, Recency, Frequency, Monetary (LRFM) method is a widely recognized approach for customer segmentation (mahmoud Taher et al., 2016). However, its application can face

How to cite:	Muhibuddin et all.(2025). Leveraging LRFM Analysis and Synthetic Data for Customer Segmentation Using K-Means Clustering. Journal Eduvest. *Vol 5*(2): 3001-3009
E-ISSN:	2775-3727

challenges such as privacy concerns, particularly when dealing with sensitive customer data, and the availability of historical data (Supangat & Mulyani, 2023). Prior studies indicate that the effectiveness of LRFM analysis relies heavily on comprehensive and high-quality historical datasets (Hasan, 2024; Ibrahim & Tyasnurita, 2022). Limited access to such data may reduce the method's precision in identifying actionable customer insights (Montenegro et al., 2020; Serwah et al., 2023). Synthetic data can help overcome these obstacles by mimicking real data characteristics without exposing sensitive information (Jordon et al., 2018). Synthetic data can also be used to generate realistic synthetic time series, improving model performance (McCrory & Thomas, 2024; Ramponi et al., 2018).

This research aims to combine synthetic data, LRFM analysis, and the K-Means clustering method to perform customer segmentation using synthetic data (Ros et al., 2023; Yoon et al., 2019). The quality of clustering will be evaluated using metrics like the Silhouette Score and Davies-Bouldin Index (Suraya et al., 2023). This approach provides a comprehensive view of customer segmentation and assists companies in developing more effective strategies to enhance customer relationships.

RESEARCH METHODS

The research uses the LRFM approach and K-Means algorithm to identify customer segments based on purchasing behavior, detailing research methods, data collection, preprocessing, implementation, and clustering outcomes as shown on figure 1.

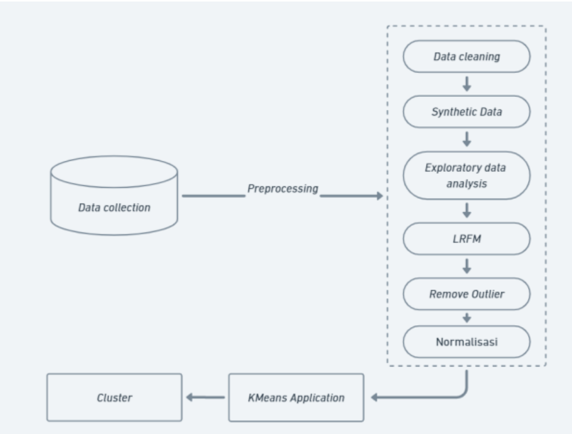

Figure 1. Methodology

This research implements five stages before applying to LRFM and K-means, including:

Data collection

Customer transaction data was obtained from the internal database management system of a Fintech company. This data includes various information such as code, name, data, note, price, status, date_cr, and date_up. The data was collected over a specific period, from October 4, 2023, to May 15, 2024. The data description is as follows:

a. code: The code of the sold product
b. name: The name of the sold product
c. data: Contains the customer number
d. note: Contains transaction notes
e. price: The price of the sold product
f. status: The status of the transaction to determine whether the process was successful or not
g. date_cr: The date the transaction was created
date_up: The date the transaction was successfully processed
h. Total number of entries in the dataset: 6189 transactions Maintaining the Integrity of the Specifications

Data Pre-processing

Based on Figure 2, it can be seen that there are no null data, and the data types for 'date_cr' and 'date_up' have been converted to datetime, and 'price' has been converted to an integer type. After the data cleaning process, the resulting data consists of 5791 rows. The focus is only on the following four columns: name, price, date_up, and anonymous_id, as shown in Table 1.

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	code	5791 non-null	object
1	name	5791 non-null	object
2	data	5791 non-null	object
3	note	5791 non-null	object
4	price	5791 non-null	int64
5	status	5791 non-null	object
6	date_cr	5791 non-null	datetime64[ns]
7	date_up	5791 non-null	datetime64[ns]

Figure 2

name	price	date_up	anonymous_id
Telkomsel 5.000	5342	04/10/2023 15:13	ad2f0ff7baf245cb8c592ede3de359db
Telkomsel Telepon Pas 5.000	5025	13/02/2024 14:50	ad2f0ff7baf245cb8c592ede3de359db
Telkomsel 5.000	5342	17/02/2024 17:21	ad2f0ff7baf245cb8c592ede3de359db
Telkomsel 5.000	5352	09/03/2024 21:48	ad2f0ff7baf245cb8c592ede3de359db
Telkomsel 10.000	10272	05/05/2024 16:48	ad2f0ff7baf245cb8c592ede3de359db
DANA 70.000	70450	04/10/2023 19:39	d1dcc5240c5d477799e7734b232e0a4a
DANA 50.000	50250	05/10/2023 22:39	d1dcc5240c5d477799e7734b232e0a4a
DANA 80.000	80250	09/10/2023 22:45	d1dcc5240c5d477799e7734b232e0a4a
DANA 75.000	75250	11/10/2023 18:08	d1dcc5240c5d477799e7734b232e0a4a

Figure 3

Data Synthetic

At this stage, the data generation process is carried out using the Time-series GAN (TimeGAN) machine learning algorithm. The synthetic data generation process increased the number of unique IDs from 1128 to 10000. The original 5791 rows of data were expanded to 51512 rows. The comparison of value descriptions can be seen in Figures 4. It can be concluded that the data maintains the same structure after the synthetic data process. The purpose of data synthesis is not to alter the data but to enrich its patterns.

	Original	Synthetic
Distinct	884	882
Distinct (%)	15.3%	1.7%
Missing	0	0
Missing (%)	0.0%	0.0%
Infinite	0	0
Infinite (%)	0.0%	0.0%

	Original	Synthetic
Mean	27967.06268	27871.41014
Minimum	1235	1235
Maximum	500235	500235
Zeros	0	0
Zeros (%)	0.0%	0.0%
Memory size	90.5 KiB	402.6 KiB

Figure 4

Feature Engineering

After the LRFM attributes were created, the researcher ensured that these features were properly prepared and suitable for use in customer segmentation analysis. Thus, the LRFM Feature Engineering process became a key step in preparing the data before performing segmentation using K-Means, which is expected to provide deeper insights into customer shopping behavior. The results can be seen in Figure 5.

	anonymous_id	first_transaction	last_transaction	frequency	monetary	length	recency
1	anonymous_id_0	2023-10-04 15:13:00	2024-05-05 16:48:00	5	39593	243	29
2	anonymous_id_1	2023-10-04 19:39:00	2024-05-13 00:10:00	37	2677640	243	21
3	anonymous_id_10	2023-10-06 08:19:00	2023-11-04 21:41:00	3	235750	241	212
4	anonymous_id_100	2023-10-25 09:38:00	2023-10-25 09:38:00	1	9906	222	222
5	anonymous_id_1000	2024-04-04 20:55:00	2024-04-04 20:55:00	1	10488	60	60
6	anonymous_id_1001	2024-04-06 22:25:00	2024-04-06 22:25:00	1	10452	58	58
7	anonymous_id_1002	2024-04-06 22:47:00	2024-04-06 22:47:00	1	11969	58	58
8	anonymous_id_1003	2024-04-07 19:08:00	2024-04-07 19:08:00	1	20145	57	57
9	anonymous_id_1004	2024-04-08 01:42:00	2024-04-08 01:42:00	1	5342	56	56
10	anonymous_id_1005	2024-04-08 18:28:00	2024-05-04 15:50:00	3	50962	56	30

Figure 5

RESULTS AND DISCUSSION

The study analyzed customer segments using the K-Means method and the LRFM approach, providing a descriptive overview of each segment's size, composition, and key characteristics. They also evaluated cluster solutions' quality using metrics and used visualizations like centroid clusters, scatter plot, and silhouette plot to interpret the results.

Determing clustering number with elbow method

Elbow Methods Determining the optimal number of clusters in K-Means modeling can be done using the Elbow method. This method is used to find the appropriate number of clusters in a dataset. Its operation is quite simple and can be easily implemented. In the Elbow method, the optimal number of clusters is determined by identifying the elbow point on the inertia graph, where this point marks a significant change in the decrease of inertia. The chosen number of clusters is where the elbow in the inertia curve is observed.

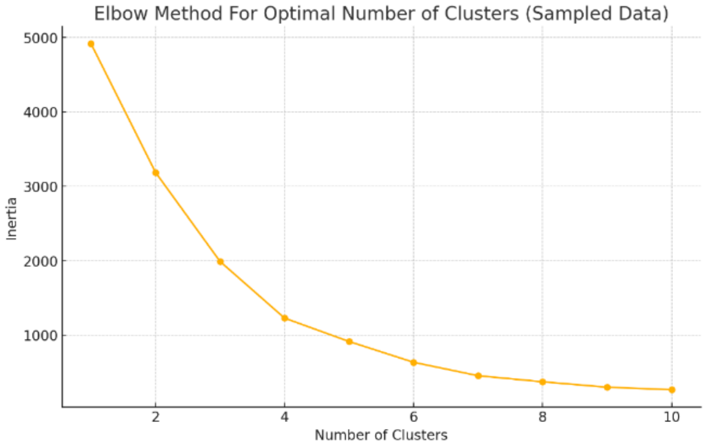

Figure 6

Clustering Dataset using K-Means

In this stage, the K-Means algorithm is applied to the previously prepared dataset by determining the optimal number of clusters. The algorithm works by initializing cluster centers randomly, calculating the distance between each data point and the nearest cluster center, updating the cluster centers based on the average of the data points in each cluster, and repeating these steps until convergence is reached. The result is a number of clusters representing groups of customers with similar characteristics, which will be used for further analysis and decision-making. The results of this process can be seen in Figure 7.

Cluster	Length_mean	Recency_mean	Frequency_mean	Monetary_mean	Count
0	119.748754	62.803589	7.169990	162873.542871	2006
1	6.689162	169.002682	1.490726	37008.728045	4475
2	162.376847	43.012315	22.226601	846204.660099	406
3	15.476670	75.218194	1.909366	44116.353810	2979

Figure 7

Figure 8

Based on Figure 8, the number of customers in each cluster after removing outliers and performing clustering with K=4 is as follows:

a. Cluster 0: 2006 customers
b. Cluster 1: 4475 customers
c. Cluster 2: 406 customers
d. Cluster 3: 2979 customers

Based on Tables 4.5 and 4.6, we can see that the characteristics generated using synthetic data and original data are not significantly different.

Tabel 1. Karakteristik Kluster pada data sintetik

Cluster	***Length***	***Recency***	***Frequency***	***Monetary***	***Jumlah User***
0	119.75	62.8	7.17	162,873.54	2006
1	6.69	169.0	1.49	37,008.73	4475
2	162.38	43.01	22.23	846,204.66	406
3	15.48	75.22	1.91	44,116.35	2979

Tabel 2. Karakteristik Kluster pada data asli

Cluster	*Length*	*Recency*	*Frequency*	*Monetary*	Jumlah *User*
0	14.60	104.92	1.89	43,813.85	363
1	6.60	199.31	1.48	36,460.31	498
2	118.92	93.64	7.05	163,514.15	222
3	157.69	74.31	21.13	821,149.76	45

Cluster Analysis

After removing outliers and performing clustering with K=4, we can analyze the characteristics of each cluster more deeply and provide tailored strategies for each cluster. The characteristics of each cluster are as follows:

Cluster 0:

Customers in this cluster have a long relationship with the company, transact fairly frequently, and spend a significant amount of money. This indicates that they are high-value and loyal customers.

Cluster 1:

This cluster consists of new or less active customers. Their last transaction was quite a while ago, and they have a low transaction frequency. They also spend less money compared to other clusters.

Cluster 2:

Customers in this cluster have very high value. They have a very long relationship with the company, transact very frequently, and spend a substantial amount of money. They are extremely valuable customers.

Cluster 3:

This cluster consists of customers who have a relatively short relationship with the company, made a recent transaction, but have a low frequency of transactions. They spend a moderate amount of money.

Based on these characteristics, the company can develop tailored strategies for each cluster to enhance customer satisfaction and increase business performance.

Evaluating the clustering

The evaluation results of the clusters in this study can be seen in Table 3, using the silhouette score calculation method with a value of 0.42 for synthetic data and 0.41 for original data. The higher Silhouette Coefficient for synthetic data indicates that objects within the cluster have slightly better cohesion compared to the original data. Based on the Davies-Bouldin Index, which is lower for the original data, it shows that the separation between clusters is better in the original data compared to the synthetic data. Overall, these two metrics indicate that both synthetic and original data have almost equivalent clustering quality, with synthetic data being slightly better in cohesion (Silhouette Coefficient), and original data being slightly better in separation (Davies-Bouldin Index).

Tabel 3. Silhouette Coefficient Values

Data	*Silhoette Coefficient*	*Davies-Bouldin Index*
Data Sintetik	0.42	0.90
Data Original	0.41	0.89

CONCLUSION

Based on the discussion outlined above, the following conclusions can be drawn from the clustering results on the LRFM dataset, which divides users into four main clusters. Clusters 0 and 2 consist of high-value customers with long-term relationships, high transaction frequency, and significant monetary value, and the strategy for these clusters is to focus on retention and loyalty by providing special offers, loyalty programs, and excellent customer service. Cluster 1 consists of inactive or new customers with short-term relationships, low transaction frequency, and low monetary value, and the strategy for this group is to run reactivation campaigns and promotions to increase transaction frequency and customer retention. Finally, Cluster 3 consists of customers with short-term relationships who have made recent transactions with moderate frequency and monetary value, and the strategy for this group is to encourage more interactions and purchases through personalized offers and purchase incentives.

DAFTAR PUSTAKA

Gul, M., & Rehman, M. A. (2023). Big data: an optimized approach for cluster initialization. Journal of Big Data, 10(1), 120.

Hasan, Y. (2024). Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan Dbscan. KAKIFIKOM (Kumpulan Artikel Karya Ilmiah Fakultas Ilmu Komputer), 60–74.

Ibrahim, M. R. K., & Tyasnurita, R. (2022). LRFM model analysis for customer segmentation using K-means clustering. 2022 International Conference on Electrical and Information Technology (IEIT), 383–391.

Jordon, J., Yoon, J., & Van Der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. International Conference on Learning Representations.

mahmoud Taher, N., Elzanfaly, D., & Salama, S. (2016). Investigation in customer value segmentation quality under different preprocessing types of RFM attributes. International Journal of Recent Contributions from Engineering, Science & IT (IJES), 4(4), 5–10.

Marisa, F., Ahmad, S. S. S., Yusof, Z. I. M., Hunaini, F., & 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).

McCrory, M., & Thomas, S. A. (2024). Cluster Metric Sensitivity to Irrelevant Features. ArXiv Preprint ArXiv:2402.12008.

Montenegro, M., Meiguins, A., Meiguins, B., & Morais, J. (2020). Improving the Clustering Algorithms Automatic Generation Process with Cluster Quality Indexes. International Conference on Computational Science and Its Applications, 1017–1031.

Ramponi, G., Protopapas, P., Brambilla, M., & Janssen, R. (2018). T-cgan: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling. ArXiv Preprint ArXiv:1811.08295.

Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A partitioning Davies-Bouldin index for clustering evaluation. Neurocomputing, 528, 178–199.

Serwah, A. M. A., KHAW, K. W. A. H., Yeng, C. S. P., & Alnoor, A. (2023). Customer analytics for online retailers using weighted k-means and RFM analysis. Data Analytics and Applied Mathematics (DAAM), 1–6.

Supangat, S., & Mulyani, Y. (2023). Customer Loyalty Analysis Using Recency, Frequency, Monetary (RFM) and K-means Cluster for Labuan Bajo Souvenirs in Online Store. Journal of Information Systems and Informatics, 5(1), 285–299.

Suraya, S., Sholeh, M., & Lestari, U. (2023). Evaluation of Data Clustering Accuracy using K-Means Algorithm. International Journal of Multidisciplinary Approach Research and Science, 2(01), 385–396.

Tomašev, N., & Radovanović, M. (2016). Clustering evaluation in high-dimensional data. In Unsupervised learning algorithms (pp. 71–107). Springer.

Yoon, J., Jarrett, D., & Van der Schaar, M. (2019). Time-series generative adversarial networks. Advances in Neural Information Processing Systems, 32.