Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 2 (2025) 218 - 224 e-ISSN: 2580-0760

Application of Formal Concept Analysis and Clustering Algorithms to Analyze Customer Segments

I Gede Bintang Arya Budaya[1*], I Komang Dharmendra[2], Evi Triandini[3]
[1]Information Technology Department, Institute of Technology and Business STIKOM Bali, Denpasar, Indonesia
[2]Information System Departement, Institute of Technology and Business STIKOM Bali, Denpasar, Indonesia
[3]Postgraduate Information System Department, Institute of Technology and Business STIKOM Bali, Denpasar, Indonesia
[1]bintang@stikom-bali.ac.id, [2]dharmendra@stikom-bali.ac.id [3]evi@stikom-bali.ac.id

Abstract

Business development cannot be separated from relationships with customers. Understanding customer characteristics is important both for maintaining sales and even for targeting new customers with appropriate strategies. The complexity of customer data makes manual analysis of the customer segments difficult, so applying machine learning to segment the customer can be the solution. This research implements K-Means and GMM algorithms for performing clustering based on the Transaction data transformed to the Recency, Frequency, and Monetary (RFM) data model, then implements Formal Concept Analysis (FCA) as an approach to analyzing the customer segment after the class labeling. Both K-Means and GMM algorithms recommended the optimal number of clusters as the customer segment is four. The FCA implementation in this study further analyzes customer segment characteristics by constructing a concept lattice that categorizes segments using combinations of High and Low values across the RFM attributes based on the median values, which are High Recency (HR), Low Recency (LR), High Frequency (HF), Low Frequency (LF), High Monetary (HM), and Low Monetary (LM). This characteristic can determine the customer category; for example, a customer that has HM and HR can be considered a loyal customer and can be the target for a specific marketing program. Overall, this study demonstrates that using the RFM data model, combined with clustering algorithms and FCA, is a potential approach for understanding MSME customer segment behavior. However, special consideration is necessary when determining the FCA concept lattice, as it forms the foundation of the core analytical insights.

Keywords: Customer Hierarchical Relationships; Data-Driven Marketing; Gaussian Mixture Model; K-Means Clustering; RFM Analysis

How to Cite: I. G. B. A. Budaya, I. K. Dharmendra, and E. Triandini, “Application of Formal Concept Analysis and Clustering Algorithms to Analyze Customer Segments”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 2, pp. 218 - 224, Mar. 2025.
DOI: https://doi.org/10.29207/resti.v9i2.6184

1. Introduction

Marketing strategies are generally related to Segmentation, Targeting, and Positioning, which are essential in understanding value for customers. Which then can help the development of the business products and services [1], [2]. First, it is important to focus on identifying the characteristics that differentiate each customer segment in the market, the results can be used to plan market targeting and product positioning in today's competitive market. The business needs a method to understand customers well to formulate appropriate and effective marketing strategies. However, understanding a large and diverse customer base data is challenging [3], [4]. Because of that, before implementing any marketing strategy, companies will try to find a way to segment customers (grouping) based on certain similarities. This approach aims to keep the relationship between the company and customers strong and increase profitability. In addition, customer segmentation serves as one of the main tools in customer understanding, which is a fundamental step in customer relationship management (CRM) [5]. It has the potential to contribute to customer retention and loyalty and aids in recognizing the value that each customer brings [6]-[8].

In business, there are segments called Micro, Small, and Medium Enterprises (MSMEs) where in countries like Indonesia, it has become one of the national concerns to support as its positive role in impacting the country economy [9], [10]. The support is about the potential to

Received: 21-11-2024 | Accepted: 18-02-2025 | Published Online: 15-03-2025

increase the business capability of MSME, in which the priority topic will be related to financial recording, digital marketing, and digitalization, especially those in the micro and small categories. Digitalizing each data in the MSME business means that there is a new path for MSMEs to run their business based on data-based strategies. However, the implementation of data-based strategies typically includes complex analysis, where not all MSME has the capacity of human resources to implement it. In the era of artificial intelligence, the implementation of machine learning can be one of the best ways to automatically give insightful analysis. The challenge for MSMEs to understand their customer in customer relationship management also becomes one of the data-based strategies that can be solved using a machine learning approach.

Understanding potential customers through customer relationship management (CRM) is a key component of business intelligence. MSMEs can improve their decision-making process by utilizing standard transaction data, which often contains variables that can be transformed into RFM (Recency, Frequency, Monetary) metrics. This allows the application of machine learning algorithms, such as clustering, to segment customers effectively. RFM variables are closely tied to customer characteristics, making them important for interpreting customer segments [11]-[13]. While some studies have relied on extensive datasets—including behavioral, psychographic, geographic, and demographic data to gain a deeper understanding of customers [14], these approaches often introduce high feature complexity, making data analysis more challenging. Moreover, some research focuses only on the optimal clustering outcomes to support business strategies [15]-[17], which limits the ability to comprehensively understand customer characteristics across different segments.

To address this limitation, further analysis is needed to gain a better understanding of customer segmentation. Study [18] introduced FCA with a hierarchical approach to build a knowledge structure for customer segmentation, effectively identifying groups of customers with common properties. The study highlighted the advantages of FCA in discovering both explicit knowledge from the co-occurrence of attributes and implicit knowledge from the implication rules within the FCA lattice. Additionally, the research compared FCA with traditional clustering methods, such as K-Means and Agglomerative Clustering, demonstrating how FCA provides more interpretable and actionable insights into customer behaviors. This methodology enables the creation of a hierarchical structure that facilitates a deeper understanding of customer segmentation.

Building on this, this research proposes the use of FCA to enhance the understanding of customer characteristics based on the RFM data model, which is particularly useful for MSMEs. Unlike Study [18] which applied FCA to hierarchical clustering, this study applies FCA after clustering with K-Means [4], [7], [19] and GMM [20], [21]. The application of FCA to the RFM dataset allows for a deeper exploration of these patterns, especially after the data is labeled by the clustering results. FCA in this study identifies the structural patterns and binary relationships [18], [22], where High (1) and Low (0) RFM values are determined based on the median values. The clustering results are used to label the data, and FCA is subsequently employed to explore customer characteristics. This approach makes the analysis more interpretable and directly applicable to MSMEs, making it easier for businesses to implement data-based strategies.

The paper is organized as follows: The first section introduces the research, providing background and the research objectives. The second section outlines the research methods used in the study. The third section presents the results and discusses the findings in relation to the research questions. Finally, the conclusion summarizes the key findings and implications of the study.

2. Research Methods

2.1 Data Collection and Preprocessing

In this research, the dataset comes from the transaction data from one MSME that has recorded 1,253 rows of transactions in one year of 2022. Before transforming data into the RFM variables, data preprocessing was conducted to identify missing values and the consistency of the data. The original variables that were used to transform the dataset become RFM variables dataset are shown in Table 1.

Table 1. Dataset description

No	Column	Description	Included in RFM Dataset (Yes/No)
1	Date	The date of the transaction occurred.	Yes
2	Name	The name of the customer making the transaction.	No
3	Email	The customer’s email.	Yes
4	Address	The customer’s residential or delivery address.	
5	Phone	The customer’s contact phone number.	No
6	Item Name	The name or description of purchased items.	No
7	Price	The price of an item purchased.	No
8	Number of Purchase	The quantity of items purchased in a transaction.	No
9	Total Price	The total cost of all items before discounts, calculated by the number of items purchased multiplied by each price.	No
10	Discount	The amount of percentage of discount applied to the total price.	No
11	Final Price	The total price after deduction by discount.	Yes

Specifically, the columns Date, Email, and Final Price are used to calculate the RFM variables. An example of transactional data used for RFM analysis is shown in Table 2, providing a brief overview of the key information used for customer segmentation.

Table 2 Sample of data used for RFM analysis

No	Date	Email	Final Price (Rp)
1	05/01/2022	x@mail.com	467.500
2	05/01/2022	y@mail.com	2.550.000
….	….	….	….
1253	24/12/2022	z@mail.com	1.275.000

2.2 RFM Modelling

This study uses the RFM model to categorize customer characteristics based on three key metrics: Recency (R), Frequency (F), and Monetary (M) Value. Recency reflects the number of days since the customer's last transaction, indicating recent engagement with the business. Meanwhile, Frequency measures the total number of transactions within a specific period, representing customer activity over time. Monetary Value refers to the total amount spent by each customer, providing insights into their overall contribution to the business revenue. These metrics are calculated for each customer, and the Relationship is calculated based on the last transaction date in the dataset, which is December 24, 2022, with the Relationship calculated starting from one day after that date.

After calculating these metrics, they are normalized to ensure consistency and accurate comparison, allowing for proper grouping and segmentation. Initially, the dataset contained 1,253 transaction records, but after filtering and summing the Monetary value for customers with multiple transactions, the dataset was reduced to 929 unique customer records. The visualization of RFM data points before the implementation of the clustering algorithm is shown in Figure 1 by using a pair plot graph. Table 3 shows the RFM dataset before normalization, which describes how the scales of R, F, and M are significantly different, and it might influence the analysis. Table 4 shows the normalized RFM dataset as the approach to make sure each variable contributes equally.

Table 3. Sample of the RFM dataset

No	Email	Recency	Frequency	Monetary
1	x@mail.com	354	1	467.500
2	y@mail.com	354	1	2.550.000
….	….	….	….	….
359	w@mail.com	109	4	1.593.750
….	….	….	….	….
929	z@mail.com	1	1	1.275.000

Table 4. Sample of the normalized RFM dataset

No	Email	Recency	Frequency	Monetary
1	x@mail.com	-0.761976	-0.367500	-0.426625
2	y@mail.com	-0.761976	-0.367500	1.079616
….	….	….	….	….
359	w@mail.com	-0.030038	2.925819	0.491241
….	….	….	….	….
929	z@mail.com	-0.690815	-0.367500	-0.144204

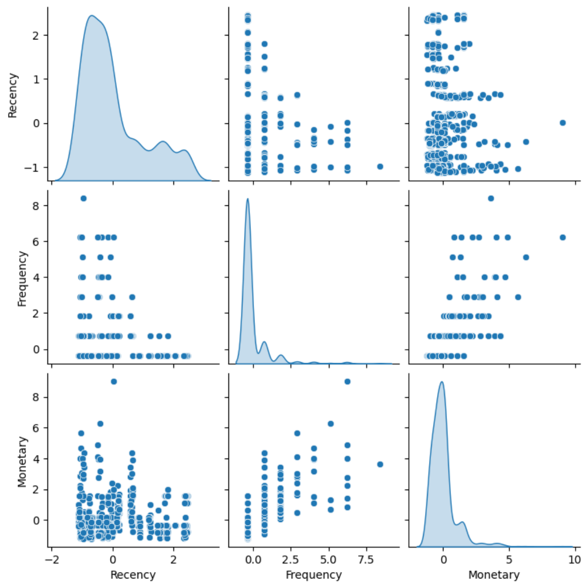

Figure 1. Pair plot visualization of the RFM dataset prior to clustering

2.3. *Clustering Analysis and Evaluation*

K-Means and GMM are the two algorithms that are implemented to perform the clustering process. K-Means itself is an algorithm that is categorized as a hard clustering algorithm which means it gives fixed information of cluster. GMM, contrary to K-Means is categorized as a soft clustering algorithm because of its probabilistic nature about the cluster information. The optimum number of clusters can be interpreted as the best customer segment number. Hence, each number of cluster scenarios must be evaluated. In this research, silhouette scores and the Davies-Bouldin Index are used as the metrics. The higher the values of the silhouette score, the better, and for the Davies-Bouldin Index, the lower the value, the better, which both direct the decision to the final number of clusters.

2.4 Formal Concept Analysis

The implementation of FCA as a data analysis method is trying to find a deeper understanding of the RFM dataset pattern after data labelling by the result of K-Means and GMM. The FCA shows the structural pattern and binary relation, in this case, the binary concept built by the category of High (1) and Low (0) RFM values based on its median values. The concept lattice then visualizes the High attribute to identify which variables between High R, High F, and High M influence the clusters as shown in Formulas 1-3.

$$\mathrm{HR} = \{\begin{smallmatrix} 1 & \text{if Recency} \leq \text{Median(Recency)} \\ 0 & \text{if Recency} > \text{Median(Recency)} \end{smallmatrix}\} \quad (1)$$

$$\mathrm{HF} = \{\begin{smallmatrix} 1 & \text{if Frequency} > 1 \\ 0 & \text{if Frequency} \leq 1 \end{smallmatrix}\} \quad (2)$$

$$\mathrm{HM} = \{\begin{smallmatrix} 1 & \text{if Monetary} \geq \text{Median(Monetary)} \\ 0 & \text{if Monetary} < \text{Median(Monetary)} \end{smallmatrix}\} \quad (3)$$

The concept lattice was analyzed by using its visualization between the clusters and the customer characteristics. The visualization shows the node that is connected by arrows between the attribute and the related cluster. In this research, the highlight point was the cluster that has High attributes either individually by R, F, and M or a combination of two from R, F, and M.

2.5 Strategic Business Intelligence

The practice implication related to the use of FCA for understanding K-Means and GMM clustering results as an approach to finding customer segments. The visualization between both algorithms shows the practical usage of FCA in RFM analysis. The strategic implication related to the potential insight gained after the analyses can become the potential information to make data-based decisions, for example, targeting the high-value segment with special campaigns and loyalty programs.

3. Results and Discussions

3.1 Results

Figure 2 shows the evaluation of clustering performance between K-Means and GMM. For K-Means clustering, the silhouette score and The Davies-Bouldin Index indicate the four clusters are the optimum cluster by comparing each value for each cluster scenario. The value for Silhouette scores is 0.55 and the value of the Davies-Bouldin Index is 0.82, which both balance the result in four clusters, as shown in Figures 2(a) and 2(b). When it comes to GMM clustering, Figure 2 (c) shows that four clusters are considered ideal by the Davies-Bouldin Index, which has a score of 0.75. Given that this score is somewhat higher than K-Means, GMM is able to handle more complex data distributions. The best four-cluster solution, on the other hand, produces a score of 0.50, which is marginally lower than K-Means but still indicates reasonable clustering quality, according to Figure 2(d), which displays the Silhouette Score for GMM. Figure 3 presents the visualization of pair plots from the RFM dataset after grouping for K-Means and GMM. Figure 3(a) illustrates the K-Means cluster assignment, where clusters are well separated based on the dimensions of recency, frequency, and monetary value. Clusters can be visually distinguished, especially along the Monetary vs. Frequency and Recency vs. Frequency axes, where different customer groups are clearly differentiated. The distribution of Recency and Frequency indicates certain high-value customer segments that frequently interact or have significant monetary involvement. Figure 3(b) visualizes the GMM cluster assignments, showing a similar distribution but with slightly more overlap compared to K-Means. The probabilistic nature of GMM clustering explains this, as customers can belong to multiple clusters with varying degrees of membership. However, the clustering separation along the Recency and Monetary axes still highlights different customer segments.

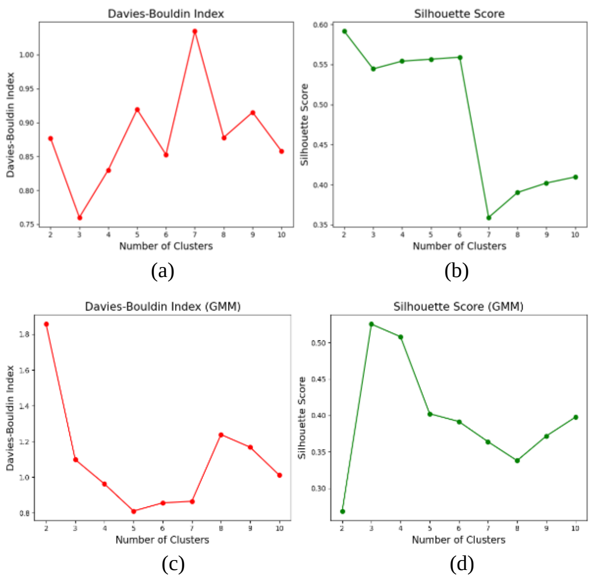

Figure 2. Clustering performance evaluation for (a) Davies-Bouldin Index (K-Means); (b) Silhouette Score (K-Means); (c) Davies-Bouldin Index (GMM); (d) Silhouette Score (GMM)

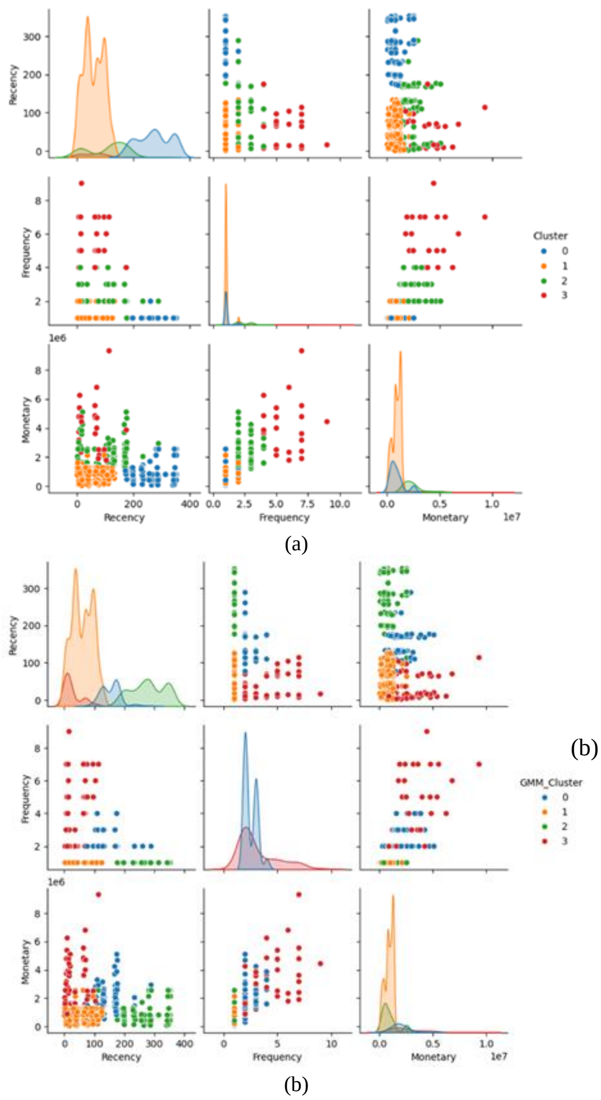

Figure 3. Pair plot visualization of the RFM dataset after clustering: (a) K-MEANS labeling; (b) GMM labeling.

For the FCA concept lattice, Figure 4 presents the concept lattice for K-Means, which illustrates the relationship between customer attributes and cluster membership (C0, C1, C2, C3). Each node in the lattice represents a unique combination of high RFM attributes and cluster membership. The relationships between nodes describe the hierarchical structure of customer segments. Starting from the *HighRecency* node, this group represents customers who made a recent purchase. This node is connected to *HighRecency, K-Means_Cluster_3*, indicating that customers in Cluster 3 exhibit high recency. Similarly, the *HighFrequency* node identifies customers with frequent transactions. This node is linked to *HighFrequency,* K-Means_Cluster_2, representing customers in Cluster 2 who demonstrate frequent interactions. Both of these nodes, *HighRecency* and *HighFrequency,* reflect general customer characteristics before transitioning to more specific high-attribute combinations.

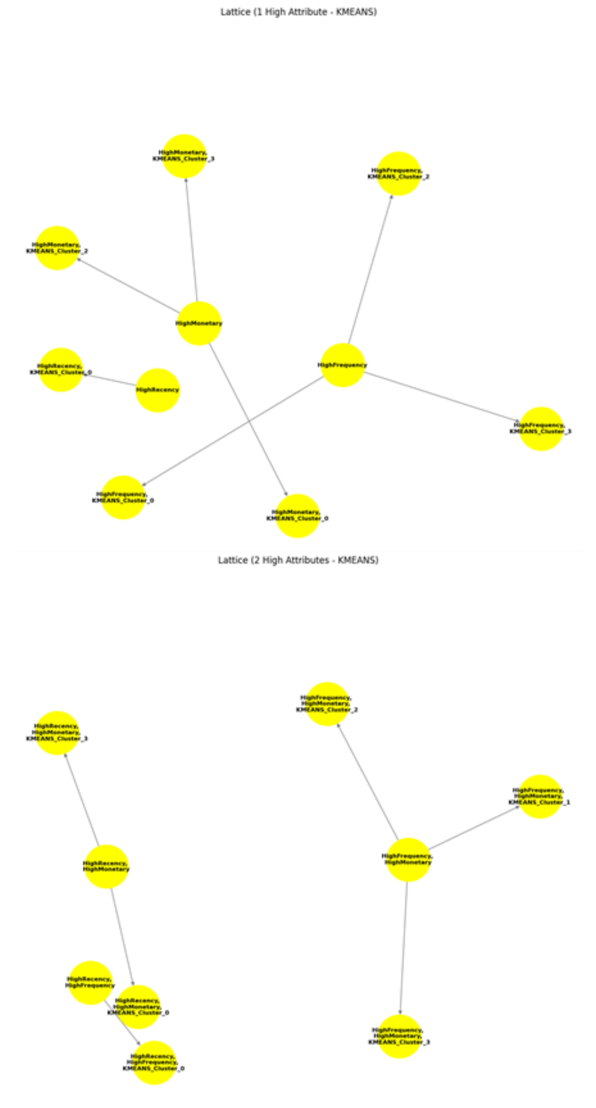

Figure 4. Lattice Visualizations of Clustering Based on High Attributes for K-MEANS

Moving up in the hierarchy, the *HighMonetary* nodes indicate customers who spend a significant amount and are connected to several specific clusters, such as *HighMonetary, K-Means_Cluster_1,* and *HighMonetary, K-Means_Cluster_3*, suggesting that these clusters contain high-value customers based on their expenditures. The nodes *HighRecency, HighMonetary*, and *K-Means_Cluster_3* highlight that customers in Cluster 3 not only spend more but are also frequently engaged. The *HighFrequency* node connected to *HighFrequency, K-Means_Cluster_1,* indicates that frequent customers are concentrated in Cluster 1 as well. Furthermore, the *HighMonetary, HighFrequency, and K-Means_Cluster_2* contain two attributes that indicate customer segments with high spending (monetary) and high frequency within Cluster 2. The K-Means grid illustrates how common attributes, such as high RFM value, converge within specific clusters, assisting businesses in identifying crucial customer segments that direct targeted marketing efforts. Figure 5 illustrates the GMM FCA visualization, which can be analyzed using the same interpretation process.

Figure 5. Lattice Visualizations of Clustering Based on High Attributes for GMM

3.2 Discussions

The approach of using FCA as a data analyst method for customer segmentation after the clustering process using K-Means and GMM is applicable. There are several important steps to consider in this approach. The first is the dataset modeling, in this research the data modeling is based on the RFM data model, which has three variables. These three variables become the base information in forming the FCA concept lattice. The next step is to determine the binary relation for the concept lattice, where in this research the binary values state as High and Low in each R, F, and M variables. To determine the High and Low, the median value can be used because the median value in transaction data can be more representative of customer behavior, for example in the monetary that has a high-range scale. For the clustering result, after the optimum cluster was determined, the next step was to assign each class. The K-Means optimum cluster results are assigned accordingly, but for GMM, the assigned process to the data point is based on the highest probability. So, when the concept lattice was formed both algorithms were already assigned fixed clusters to each data point.

The visualization of the FCA concept lattice proved to be helpful. The node that contains the cluster attribute and cluster label, and its relation with the R, F, and M attributes can show the potential grouping of customer characteristics. The individual high attribute of each R, F, and M can become a guideline for specific CRM strategies, such as the customer segment with HR attribute which means those have the most recent transactions. They tend to have a good memory about their experience in a recent transaction which means the after-sales program needs to be maximalised. The customer segment who become part of HF can be categorized as loyal customers so marketing strategies like membership discounts can be the solution. Lastly, for the customer segment that is part of HM, which is the biggest spender customer, special treatment like VIP treatment might interest them. It can be highlighted that the combination of two High attributes can become another potential insight that shows the characteristics of the customer. Finally, this structured approach can become the guideline for MSMEs to interact with their potential customer based on customer characteristics.

4. Conclusions

This research objective is to implement FCA to analyze the clustering result of K-Means and GMM to understand MSMEs' customer segment behavior based on the RFM data model. The result indicating FCA indeed can become a practice approach to do deeper analysis for the RFM data model, especially for clustering problems. However, there is a need for special attention to determine the FCA concept lattice as it will become the core analysis knowledge. The performance of clustering algorithms K-Means and GMM both resulted in four optimum clusters with different data groups. This research suggests the potential implementation of this approach for MSMEs in the case of data-based decision-making. The limitation of this research is the small dataset sample. Future research could explore the potential of implementing different clustering algorithms such as hierarchical clustering and using transaction datasets from different industries.

Acknowledgements

This work was supported by the Directorate General of Higher Education, Research, and Technology (*Ditjen Diktiristek*) through the Directorate of Research, Technology, and Community Service (DRTPM) under the 2024 Beginner Lecturer Research (*Penelitian Dosen Pemula/*PDP) Program grant, Grant Number 110/E5/PG.02.00.PL/2024, dated June 11, 2024.

References

[1] S. Gupta, T. Justy, S. Kamboj, A. Kumar, and E. Kristoffersen, “Big data and firm marketing performance: Findings from knowledge-based view,” *Technol Forecast Soc Change*, vol. 171, p. 120986, 2021.

[2] S. Munusamy and P. Murugesan, “Modified dynamic fuzzy c-means clustering algorithm–Application in dynamic customer segmentation,” *Applied Intelligence*, vol. 50, no. 6, pp. 1922–1942, 2020.

[3] H. Chen, L. Zhang, X. Chu, and B. Yan, “Smartphone customer segmentation based on the usage pattern,” *Advanced Engineering Informatics*, vol. 42, p. 101000, 2019.

[4] Y. Deng and Q. Gao, “A study on e-commerce customer segmentation management based on improved K-means algorithm (Retraction of Vol 18, Pg 497, 2020),” 2022, *SPRINGER HEIDELBERG TIERGARTENSTRASSE 17, D-69121 HEIDELBERG, GERMANY*.

[5] F. Leisch, S. Dolnicar, and B. Grün, *Market segmentation analysis: Understanding it, doing it, and making it useful*. 2018.

[6] M. T. Ballestar, P. Grau-Carles, and J. Sainz, “Customer segmentation in e-commerce: Applications to the cashback business model,” *J Bus Res*, vol. 88, pp. 407–414, 2018.

[7] E. Y. L. Nandapala and K. P. N. Jayasena, “The practical approach in Customers segmentation by using the K-Means Algorithm,” in *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, IEEE, 2020, pp. 344–349.

[8] J. Wu *et al.*, “An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm,” *Math Probl Eng*, vol. 2020, 2020.

[9] R. M. Saputra and H. Darmawan, “EFFECTS OF URBANIZATION AND THE GROWTH OF MICRO, SMALL, AND MEDIUM ENTERPRISES (MSMES) AS SUPPORTS FOR THE ECONOMY OF INDONESIA IN AN URBAN CONTEXT,” *Journal of Social Political Sciences*, vol. 4, no. 2, pp. 201–214, 2023.

[10] F. Rinaldi, S. Maarif, S. Thamrin, and A. A. Supriyadi, “Role of Micro, Small, and Medium Enterprises (MSMEs) in Supporting National Defense from Economic Perspective,” *Journal of Positive School Psychology*, pp. 8914–8920, 2022.

[11] T. BELHADJ, “Customer Value Analysis Using Weighted RFM model: Empirical Case Study.,” *Al Bashaer Economic Journal*, vol. 7, no. 3, 2021.

[12] M. Kovacs, D. Buryakov, and V. Kryssanov, “An Unsupervised Approach for Customer Need Assessment in E-commerce: A Case Study of Japanese Customer Reviews,” in *Proceedings of the 2021 6th International Conference on Cloud Computing and Internet of Things*, 2021, pp. 41–48.

[13] D. Gerritsen and R. van Olderen, “Events and customer relationship marketing.,” 2020.

[14] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, no. 3, pp. 527–570, 2023.

[15] I. G. Harsemadi, D. P. Agustino, and I. G. B. A. Budaya, "Klasterisasi Pelanggan Tenant Inkubator Bisnis STIKOM Bali Untuk Strategi Manajemen Relasi Dengan Menggunakan Fuzzy C-Means," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 4, pp. 232–243, 2023.

[16] D. P. Agustino, I. G. Harsemadi, and I. G. B. A. Budaya, "Edutech Digital Start-Up Customer Profiling Based on RFM Data Model Using K-Means Clustering," *Journal of Information Systems and Informatics*, vol. 4, no. 3, pp. 724–736, 2022.

[17] D. P. Agustino, I. G. B. A. Budaya, I. G. Harsemadi, I. K. Dharmendra, and I. M. S. A. Pande, "Comparison of the DBSCAN Algorithm and Affinity Propagation on Business Incubator Tenant Customer Segmentation," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 2, pp. 315–321, 2023.

[18] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J. Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA," *Expert Syst Appl*, vol. 237, p. 121449, 2024.

[19] S. F. Djun, I. G. A. Gunadi, and S. Sariyasa, "Analisis Segmentasi Pelanggan pada Bisnis dengan Menggunakan Metode K-Means Clustering pada Model Data RFM," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 5, no. 4, pp. 354–364, 2024.

[20] E. Patel and D. S. Kushwaha, "Clustering cloud workloads: K-means vs gaussian mixture model," *Procedia Comput Sci*, vol. 171, pp. 158–167, 2020.

[21] Y. Fu, X. Liu, S. Sarkar, and T. Wu, "Gaussian mixture model with feature selection: An embedded approach," *Comput Ind Eng*, vol. 152, p. 107000, 2021.

[22] A. Braud, A. Buzmakov, T. Hanika, and F. Le Ber, *Formal Concept Analysis*. Springer, 2021.