ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Centroid Optimization of K-Means Using Ant Colony Optimization for Culinary MSME Clustering

Muhammad Fharahbi Fachri[1], **Lisna Zahrotun**[2]

[1, 2]Informatics Department, Ahmad Dahlan University, Yogyakarta, Indonesia

Received:
December 12, 2025
Revised:
January 23, 2026
Accepted:
February 15, 2026
Published:
March 2, 2026

Corresponding Author:
Author Name*:
Muhammad Fharahbi Fachri
Email*:
2200018132@webmail.uad.ac.id

DOI:
10.63158/journalisi.v8i1.1443

Abstract. Micro, Small, and Medium Enterprises (MSMEs) are economic activities conducted by individuals or groups, particularly in the culinary sector. The rapid expansion of culinary MSMEs, especially in tourism-oriented regions such as the Special Region of Yogyakarta, necessitates effective data clustering to systematically analyze their characteristics. High-quality clustering plays a crucial role in supporting informed decision-making, including business development planning, MSME assistance programs, and the formulation of well-targeted policies. This study applies the K-Means algorithm to cluster culinary MSME data; however, its performance is sensitive to centroid initialization, which may result in suboptimal clustering outcomes. To address this limitation, Ant Colony Optimization (ACO) is employed as a centroid optimization approach. ACO is a metaheuristic algorithm inspired by the foraging behavior of ant colonies, where pheromone trails guide the search toward optimal solutions. The results indicate that the integration of ACO enhances clustering performance compared to K-Means. The silhouette scores obtained are 0.88 and 0.89 for two clusters, 0.80 and 0.86 for three clusters, and 0.80 and 0.92 for four clusters for K-Means and ACO-optimized K-Means, respectively. These findings demonstrate that ACO effectively improves centroid initialization, with four clusters identified as the optimal configuration.

Keywords: Data Mining, Ant Colony Optimization, Clustering Optimization, K-Means, MSME

1. INTRODUCTION

Micro, Small, and Medium Enterprises are economic entities operated by individuals or groups and play a vital role in supporting regional and national economic growth. MSMEs are classified based on the amount of capital and annual sales turnover and are widely recognized for their adaptability to change and innovation, particularly in technology adoption. Due to their relatively small operational scale, MSMEs tend to maintain closer working relationships and demonstrate greater flexibility among business actors. MSMEs operate across various sectors including services, fashion, handicrafts, laundry, automotive, beauty, health, and culinary. Among these sectors, the culinary sector is the most prevalent and widely distributed across regions, offering strong appeal to both business owners and the public [2]. Tourism development is one of the key factors contributing to the growth of culinary MSMEs. The Special Region of Yogyakarta is a prominent example of a tourism-oriented area with significant culinary MSME potential. According to data from the Central Statistics Agency (BPS), in 2024 the number of domestic tourists visiting Yogyakarta reached approximately 38 million, while foreign tourist arrivals exceeded 7 thousand visitors [3]. This condition highlights the urgency of MSME development, as the sector has proven to absorb a substantial portion of the workforce, strengthen local economic resilience, and serve as a backbone of national economic stability.

In the context of digital transformation and increasingly dynamic market competition, effective MSME data management is essential to support accurate and targeted decision-making. One important approach is clustering MSME data to extract meaningful patterns and characteristics. Through clustering, various attributes related to education level, business activities, marketing practices, ownership structure, access to electronic media, government assistance, capital loans, turnover, health insurance ownership, workforce composition, and business age can be analyzed in a more structured manner [1]. Previous studies have utilized variables related to business activities, marketing objectives, business credit loans, workforce composition, and business age; however, the resulting clustering performance remained unsatisfactory, with the best silhouette score reported at only 0.60 [1].

In this study, the analysis is limited to seven key variables, namely annual turnover, highest education level, land or building ownership status, health insurance ownership, availability of electronic media facilities, government assistance, and gender. This variable selection is based on prior research [4] that identifies critical success factors for MSMEs, including technology utilization, effective management, government support, and labor or owner competence. These seven variables are considered representative of the most influential factors affecting the clustering of culinary MSMEs.

Clustering is a data mining technique used to group data objects based on similarity in their characteristics. Data mining itself refers to the process of extracting valuable information from large datasets to support critical decision-making processes [5]. One of the most commonly used clustering algorithms is K-Means, which partitions data into a predefined number of clusters by minimizing the distance between data points and cluster centroids [6]. Despite its simplicity and efficiency, K-Means is highly sensitive to centroid initialization, which can lead to suboptimal clustering results. Previous studies have reported relatively low clustering performance when using K-Means, with silhouette scores as low as 0.56 in certain applications [7]. Since the silhouette score ranges from -1 to 1, higher values indicate better clustering quality[8]. A higher silhouette score indicates better clustering quality. This is where the role of Ant Colony Optimization is very relevant, as it is able to improve the centroid initialization process and help to obtain more accurate and meaningful clustering results [9].

Ant Colony Optimization (ACO) is a metaheuristic algorithm inspired by the foraging behavior of ant colonies, where artificial ants communicate indirectly through pheromone trails to discover optimal paths or solutions in complex search spaces [10]. Several studies have demonstrated the effectiveness of ACO in solving optimization problems and improving solution quality. For instance, ACO has been successfully applied to various real-world datasets, achieving accuracy improvements ranging from 23% to 41% [11]. With this approach, Ant Colony Optimization can avoid inappropriate clustering results that usually occur due to random centroid initialization in K-Means. The end result is a more accurate cluster centroid, so that the clustering of culinary MSME data becomes more accurate and informative [12].

Previous studies have applied K-Means to MSME data, producing three clusters without reporting clustering evaluation results [13]. Furthermore, study [1] compared K-Means and K-Medoids with the two best clusters, but the accuracy obtained was still relatively low (0.60). Meanwhile, study [12] combined the K-Means algorithm with Ant Colony Optimization (ACO) and showed an increase in clustering accuracy, but it was still limited to the simple and controlled Iris dataset. Moreover, the application of ACO in this study is still based on a Traveling Salesman Problem analogy and is not explicitly focused on centroid position optimization. Thus, there are still research gaps related to the application of K-Means–ACO on complex MSME data with objective clustering evaluation and more targeted centroid optimization.

Based on the identified research gap, this study makes several key contributions. First, it proposes an optimization approach for K-Means centroid initialization using Ant Colony Optimization to improve the clustering quality of culinary MSME data. Second, this study conducts an empirical evaluation of clustering performance using silhouette score across multiple cluster configurations. Third, it provides an interpretation of culinary MSME cluster characteristics to support data-driven decision-making and MSME development policies. By integrating K-Means and Ant Colony Optimization, this study aims to produce an optimal clustering model that accurately represents the characteristics of culinary MSMEs.

2. METHODS

This study applies clustering techniques by utilizing the K-Means algorithm and the metaheuristic algorithm Ant Colony Optimization. Then the silhouette score of K-Means and Ant Colony Optimization is compared to see the best performance. The goal is to see how much the accuracy of the results of the data grouping of Culinary MSMEs in Yogyakarta has increased after centroid optimization using Ant Colony Optimization.

2.1 Culinary MSME Data

The dataset used in this study was obtained from the Yogyakarta City Office of Industry, Cooperatives, and MSMEs. The data were collected before and during the COVID-19 pandemic period. Prior to the pandemic, data collection was conducted through direct field verification, while during the pandemic, data were collected remotely via WhatsApp due to mobility restrictions.

The dataset consists of MSMEs from various sectors, with a specific focus on the culinary sector. A total 1,336 culinary MSME records were used in this study. To ensure data privacy, all personal identifiers such as owner names, contact numbers, and exact business addresses were anonymized or excluded from the dataset. The data were used solely for research purposes and analyzed in aggregated form. Details of Culinary MSME data are presented in Table 1.

Table 1. Culinary MSME data

Education	Province	City/District	Subdistrict	...	Age
High School	Special Region of Yogyakarta	Yogyakarta City	Tegalrejo	...	-
High School	Special Region of Yogyakarta	Yogyakarta City	Danurejan	...	35-50 Years
0	Special Region of Yogyakarta	Yogyakarta City	Tegalrejo	...	35-50 Years
High School	Special Region of Yogyakarta	Yogyakarta City	Umbulharjo	...	25-35 Years
Junior High School	Yogyakarta Region	Gunungkidul District	Karangmojo	...	-
...	...	...	...	...	...

2.2 Research Stage

According to [14] The research stage is a guide or structured flow that guides the researcher throughout the research process. The stages of the research used are to use techniques from Knowledge Discovery in Database (KDD) is the process of finding new information from a set of data or database [15]. The picture of the research stages is in Figure 1.

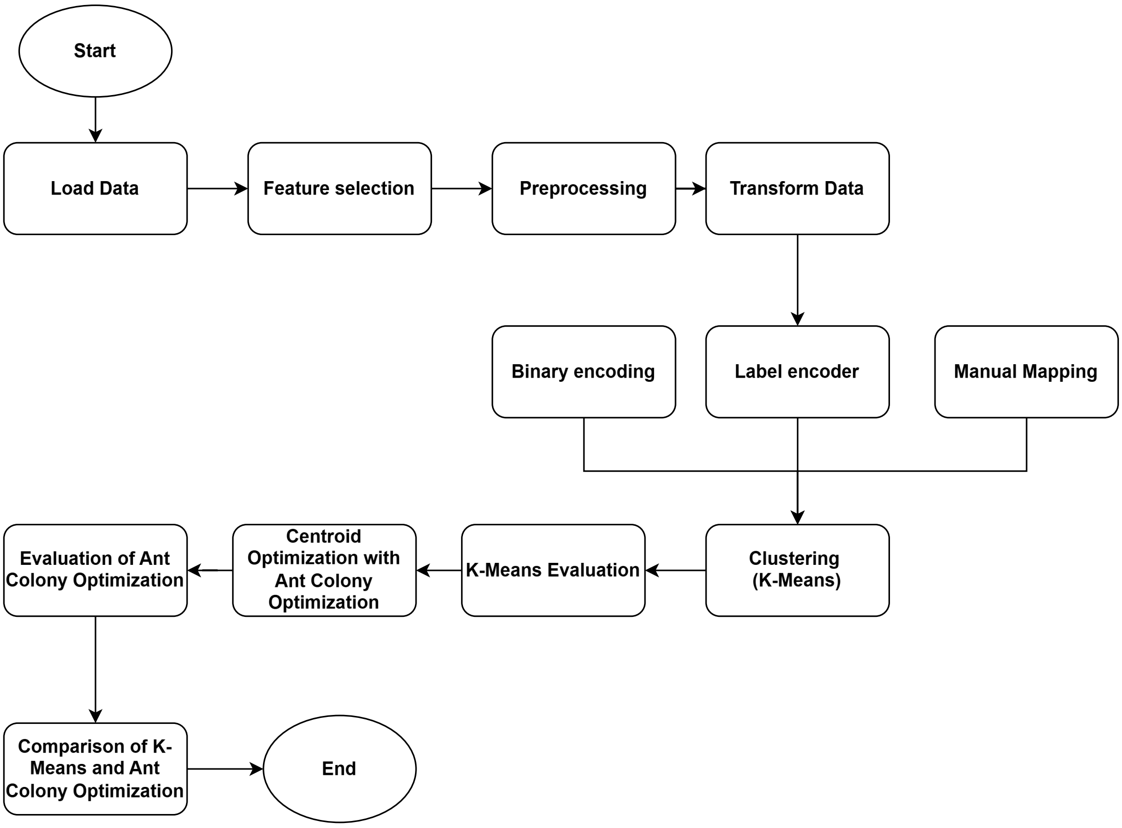


Figure 1. Stages of research

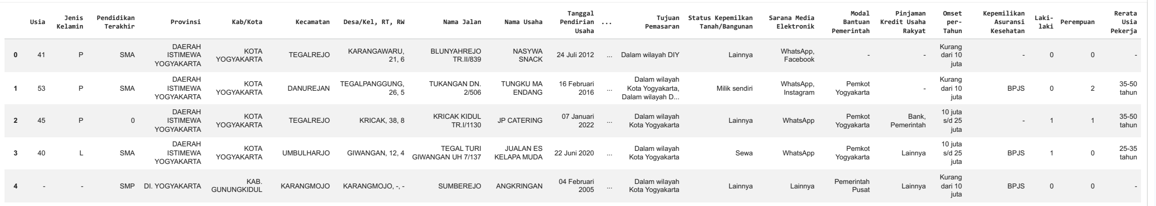
Table 2. Explained stages of research

Stage	Process	Description
1	Load Data	Load Culinary MSME dataset
2	Feature Selection	Select relevant clustering attributes
3	Preprocessing	Handle missing value and duplicated
4	Transform Data	Changing category data using binary encoding, label encoders, and manual mapping
5	Clustering (K-Means)	K-Means Clustering
6	K-Means Evaluation	Silhouette Score K-Means
7	Optimization With Ant Colony Optimization	Centroid Optimization using Ant Colony Optimization
8	Evaluation Ant Colony Optimization	Silhouette Score Ant Colony Optimization
9	Comparison	K-Means vs Ant Colony Optimization

Table 2 presents Explained stages of research process used in this study A more detailed explanation of each stage is provided in the following subsections:

1) Load Data

This stage involves loading the culinary MSME dataset obtained from the Yogyakarta City Office of Industry, Cooperatives, and MSMEs into the analysis environment. In this process, the data is entered into the target system, where each record and attribute is organized in a table format to ensure compatibility with the subsequent preprocessing and grouping stages. In other words, the data is entered into the target system, enabling further processing and analysis[16]. The results of the data load are shown in Figure 2.



	Usia	Jenis Kelamin	Pendidikan Terakhir	Provinsi	Kab/Kota	Kecamatan	Desa/Kel, RT, RW	Nama Jalan	Nama Usaha	Tanggal Pendirian Usaha	...	Tujuan Pemasaran	Status Kepemilkan Tanah/Bangunan	Sarana Media Elektronik	Modal Bantuan Pemerintah	Pinjaman Kredit Usaha Rakyat	Omset per-Tahun	Kepemilikan Asuransi Kesehatan	Laki-laki	Perempuan	Rerata Usia Pekerja
0	41	P	SMA	DAERAH ISTIMEWA YOGYAKARTA	KOTA YOGYAKARTA	TEGALREJO	KARANGAWARU, 21, 6	BLUNYAHREJO TR.II/839	NASYWA SNACK	24 Juli 2012	...	Dalam wilayah DIY	Lainnya	WhatsApp, Facebook	-	-	Kurang dari 10 juta	-	0	0	-
1	53	P	SMA	DAERAH ISTIMEWA YOGYAKARTA	KOTA YOGYAKARTA	DANUREJAN	TEGALPANGGUNG, 26, 5	TUKANGAN DN. 2/506	TUNGKU MA ENDANG	16 Februari 2016	...	Dalam wilayah Kota Yogyakarta, Dalam wilayah D...	Milik sendiri	WhatsApp, Instagram	Pemkot Yogyakarta	-	Kurang dari 10 juta	BPJS	0	2	35-50 tahun
2	45	P	0	DAERAH ISTIMEWA YOGYAKARTA	KOTA YOGYAKARTA	TEGALREJO	KRICAK, 38, 8	KRICAK KIDUL TR.I/1130	JP CATERING	07 Januari 2022	...	Dalam wilayah Kota Yogyakarta	Lainnya	WhatsApp	Pemkot Yogyakarta	Bank, Pemerintah	10 juta s/d 25 juta	-	1	1	35-50 tahun
3	40	L	SMA	DAERAH ISTIMEWA YOGYAKARTA	KOTA YOGYAKARTA	UMBULHARJO	GIWANGAN, 12, 4	TEGAL TURI GIWANGAN UH 7/137	JUALAN ES KELAPA MUDA	22 Juni 2020	...	Dalam wilayah Kota Yogyakarta	Sewa	WhatsApp	Pemkot Yogyakarta	Lainnya	10 juta s/d 25 juta	BPJS	1	0	25-35 tahun
4	-	-	SMP	DI. YOGYAKARTA	KAB. GUNUNGKIDUL	KARANGMOJO	KARANGMOJO, -, -	SUMBEREJO	ANGKRINGAN	04 Februari 2005	...	Dalam wilayah Kota Yogyakarta	Lainnya	Lainnya	Pemerintah Pusat	Lainnya	Kurang dari 10 juta	BPJS	0	0	-

Figure 2. Load data culinary MSME

2) Feature Selection

Features selection is the process of selecting relevant variables from the Culinary MSME dataset to be used in the clustering process. Feature selection aim to reduce or eliminate less relevant features that do not significantly contribute to the clustering objectives[17]. Not all available attributes are included, as some variables are considered irrelevant or redundant. In this study, the selected features represent economic, demographic, and business characteristics of Culinary MSMEs, namely annual turnover, last education, ownership status, insurance ownership, assistance capital, gender and electronic media facilities. These features are expected to better describe the characteristics of Culinary MSMEs and support the formation of optimal clusters.

3) Preprocessing

At this stage, the Culinary MSME dataset is examined to identify missing values and duplicate records in the selected features. Data with incomplete or duplicate information are handled appropriately to prevent bias and inaccuracies in the clustering results. Through proper preprocessing, the data become ready to be analyzed and modeled[18].

4) Transform Data

Transforms is the process of converting data from its initial form to a format that suits the needs of the analysis [19]. In the research, the research carried out was to convert category data into numerical by using the *label encoder*, *binary encoding* and manual mapping. Manual mapping is applied to categorical attributes that have an inherent order or logical meaning, allowing the numerical representation to better reflect the actual relationship between categories. This approach helps prevent misleading interpretations that may occur when arbitrary numerical assignments are used and ensures that the transformed data remain meaningful for the clustering process [20]. However, manual mapping relies on predefined numerical assignments, which may introduce subjectivity and limit flexibility when applied to different datasets.

5) Clustering (K-Means)

Clustering is performed using the K-Means algorithm to group Culinary MSME data that have undergone preprocessing and data transformation. K-Means is widely used in data mining to discover patterns and acquire knowledge from data [21]. In this study, Euclidean distance is applied as the distance metric, as shown in Equation (1), because it is suitable for numerical data and effectively measures similarity between data points by calculating the straight-line distance in multidimensional space. The use of Euclidean distance is appropriate since all categorical attributes have been transformed into numerical form prior to the clustering process as shown in Equation 1 [22].

$$d_{x,y}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

To improve clustering performance, Ant Colony Optimization (ACO) is employed to optimize the initial centroid positions before executing the K-Means algorithm. Although K-Means++ is commonly used for centroid initialization, it is not applied in this study in order to clearly evaluate the contribution of ACO as an alternative centroid optimization approach. Therefore, K-Means initialization is used as the baseline for comparison.

6) Evaluation K-Means

After clustering, the K-Means algorithm was evaluated with using silhouette score as defined in equation (4). The Silhouette Score measures how well each data point is assigned to its corresponding cluster [23], as shown in Equation 2.

$$S(i)=\frac{b_i-a_i}{\max(a_i,b_i)} \tag{2}$$

7) Ant Colony Optimization

Ant Colony Optimization is an algorithm metaheuristic which is affected by the natural behavior of ant colonies in finding the shortest path to a food source. In the real world, ants communicate through pheromone, which is the chemical substances they leave along the path they travel. Pathways with higher pheromone intensity will be more attractive to other ants, so they will be further strengthened and potentially become the best solution Collective [24], as shown in Equation 3.

$$P_{ij}=\frac{\tau_{ij}^{\alpha}\times\eta_{ij}^{\beta}}{\sum_k \tau_{ik}^{\alpha}\times\eta_{ik}^{\beta}} \tag{3}$$

Description:

$\tau_{(ij)}$ = the value of pheromones in element j

$\eta_{(ij)}$ = heuristic value (e.g. inverse distance)

α = weight effect of pheromones

β = heuristic influence weight

$\sum k$ = sum of all possible element choices

$$\tau_{ij}=(1-\rho)\tau_{ij}+\Delta\tau_{ij} \tag{3}$$

Description:

ρ = Evaporation rate

The parameter values used in this study were adopted from the approach reported in [11], [12] and adjusted to suit the characteristics of the dataset used in this research. Through experimental tuning, the best performing parameter values were obtained and applied in the clustering process. Although effective in improving centroid initialization, Ant Colony Optimization introduces additional computational overhead due to its iterative process and repeated pheromone updates. As a result, the execution time of the clustering process becomes longer compared to standard K-Means, with an average computational time of approximately 20-40 minutes per cluster. This increased runtime represents a limitation of the proposed approach particularly when applied to larger datasets.

8) Evaluation Ant Colony Optimization

After clustering using the Ant Colony Optimization based K-Means approach, the resulting clusters evaluated the silhouette score metric as defined in Equation (4). This metric was employed to assess cluster quality in terms of cohesion and separation[23].

9) Comparison

The clustering performance of K-Means and Ant Colony Optimization was compared using the silhouette score. This comparison aims to quantitatively assess the impact of centroid optimization using Ant Colony Optimization on cluster quality in terms of cohesion and separation. By applying the same evaluation metric and dataset, the comparison ensures a fair and consistent performance assessment between the two methods [25].

3. RESULTS AND DISCUSSION

3.1 Selection Data

After an initial inspection of the MSME dataset, several columns were identified as irrelevant to the clustering objective and therefore removed. The features retained were annual turnover, final education, land/building ownership status, insurance ownership, electronic media facilities, government assistance capital, and gender. These attributes were selected based on their relevance to MSME performance in the primary sector, as suggested in previous studies [4]. The complete list of selected features and their categorical descriptions is presented in Table 3, which serves as the foundation for subsequent preprocessing and clustering stages.

Table 3. Features used

Features	Remarks
Turnover per Year	Under 10 million, 10 to 25 million, 40 to 55 million, 55 to 70 million, 70 to 85 million, 85 to 100 million, 100 to 120 million, 120 to 150 million, more than 150 million
Final Education	-, 0, SD, SMP, SMA, SMK, D1, D2, D3, D4, S1, S2, S3
Land or Building Ownership Status	Other, Magersari (custom), Owned, Rent
Insurance Ownership	-, Private Insurance, BPJS, BPJS and Private Insurance
Electronic Media Facilities	-, Facebook, Gojek, Grab, Instagram, Shopee, Tokopedia, Twitter, WhatsApp, more
Government Assistance Capital	-, Yogyakarta Regional Government, Central Government, Yogyakarta City Government
Gender	-, L, P

Table 3 demonstrates that the selected attributes capture both economic capacity (turnover, capital assistance), human capital (education, gender), and technological adoption (electronic media facilities), which are critical dimensions in explaining MSME heterogeneity.

3.2 Preprocessing

Preprocessing was conducted to ensure data quality prior to clustering. According to [26] data cleansing is a crucial step in the Knowledge Discovery in Databases (KDD) process. This stage involved checking for missing values, duplicated records. The results, summarized in Table 4, show that all selected features have zero missing values and no duplicate records.

Table 4. Missing value and duplicate data

Features	Amount of Blank Data	Duplicate	Data Type
Turnover per Year	0	0	Object
Final Education	0	0	Object
Land/Building Ownership Status	0	0	Object
Insurance Ownership	0	0	Object

Features	Amount of Blank Data	Duplicate	Data Type
Electronic Media Facilities	0	0	Object
Government Assistance Capital	0	0	Object
Gender	0	0	Object

Based on Table 4, it can be concluded that the dataset is clean and suitable for transformation and clustering analysis without requiring imputation or record removal, thereby minimizing potential bias in the clustering results.

3.3 Transform Data

Data transformation was performed to convert categorical attribute into numerical representations suitable for distance-based clustering algorithms. Several techniques were applied:

1) Binary encoding was used for the electronic media facilities attribute to represent multiple platform usage in separate binary columns. The results of this process are shown in Table 5, which illustrates how combination of media usage (e.g., WhatsApp, Facebook, Instagram) are encoded into binary vectors.

Table 5. Binary encoding

Electronic Media Facilities	Media Facilities Elektronik_0	...	Media Facilities Elektronik_3	Media Facilities Elektronik_4	Media Facilities Elektronik_5	Media Facilities Elektronik_6
WhatsApp, Facebook	0	...	0	0	1	1
WhatsApp, Instagram	0	...	0	1	0	0
WhatsApp	0	...	0	1	0	1
WhatsApp	0	...	0	1	0	1
Others	0	...	1	0	0	0

2) Label encoding was applied to ordinal and nominal attributes such as final education, land ownership status, insurance ownership, and government assistance capital. The encoding scheme is presented in Table 6.

Table 6. Label encoder

Features	Before	After
Final Education	SMA, S1, SMP, …	13, 9, 15
Land Ownership Status	Owned, rented	2, 3, …
Insurance Ownership	BPJS, -, …	2, 0, …
Government Assistance Capital	-, Central Government, …	0, 2, …

3) Manual mapping was used for annual turnover by converting categorical income range into their average monetary values, as shown in Table 7, to better reflect economic scale differences.

Table 7. Revenue mapping

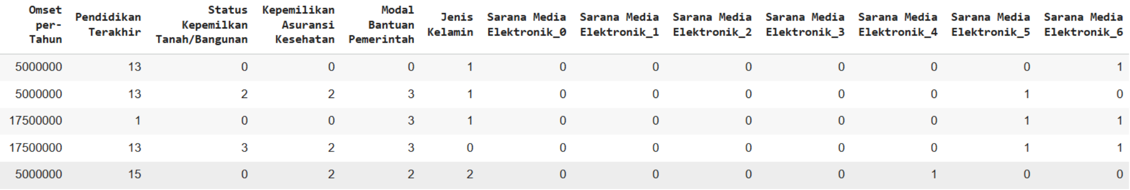
Annual Turnover	Average
Under 10 million	5.000.000
10-25 million	17.500.000
25-40 million	32.500.000
40-55 million	47.500.000
55-70 million	62.500.000
70-85 million	77.500.000
85-100 million	92.500.000
100-120 million	110.000.000
120-150 million	135.000.000
More than 150 million	175.000.000

4) Gender was transformed using manual mapping, as presented in Table 8, to maintain consistency with numerical processing requirements.

Table 8. Gander mapping

Gender	Value
-	2
P	1
L	0

The overall result of the transformation stage is visualized in Figure 3, which illustrates the fully numerical dataset that serves as input for the clustering algorithms. This transformation ensures that all variables contribute quantitatively to distance calculations in K-Means and Ant Colony Optimization.



Omset per-Tahun	Pendidikan Terakhir	Status Kepemilkan Tanah/Bangunan	Kepemilikan Asuransi Kesehatan	Modal Bantuan Pemerintah	Jenis Kelamin	Sarana Media Elektronik_0	Sarana Media Elektronik_1	Sarana Media Elektronik_2	Sarana Media Elektronik_3	Sarana Media Elektronik_4	Sarana Media Elektronik_5	Sarana Media Elektronik_6
5000000	13	0	0	0	1	0	0	0	0	0	0	1
5000000	13	2	2	3	1	0	0	0	0	0	1	0
17500000	1	0	0	3	1	0	0	0	0	0	1	1
17500000	13	3	2	3	0	0	0	0	0	0	1	1
5000000	15	0	2	2	2	0	0	0	0	1	0	0

Figure 3. Result transform data

3.4 K-Means Clustering

K-Means is an example of an algorithm clustering. This algorithm works by determining cluster and divide the data into K groups based on the proximity of the distance between the data and centroid (cluster center point) [27]. The next stage is to use Culinary MSME data to be grouped with K-Means. The process includes:

1) Determine K as 2,3 and 4. This range was intentionally selected to maintain interpretability of the clustering results while enabling a fair comparison of clustering quality. Using a small number of clusters also allows the optimization process with Ant Colony Optimization to focus on improving centroid initialization rather than increasing model complexity.
2) At the initial stage, K-Means randomly initializes centroid positions for each cluster.
3) Each MSME data point is assigned to the nearest centroid based on the distance calculation, as defined in Equation (1). Data points with minimum distance to a centroid are grouped into the same cluster.
4) After data assignment, new centroids are calculated by computing the mean value of all data points within each cluster. This step is repeated iteratively until the centroids no longer change significantly or convergence is achieved.
5) The evaluation results for each value of K are presented in Table 9. Based on the table, K = 2 yields the highest Silhouette Score (90.88) for K-Means; however, this configuration provides less detailed segmentation of MSME characteristics.

Although K-Means demonstrates fast convergence, it is sensitive to initial centroid placement and may converge to local optimal solutions. This limitation motivates the integration of Ant Colony Optimization for centroid optimization [28].

Table 9. Grouping Results K-Means

Number of Clusters (K)	Silhouette Score
2	0.88
3	0.81
4	0.81
Best Silhouette Score K-Means	**0.88**

3.5 Ant Colony Optimization Cluster

In context clustering, Ant Colony Optimization used to optimize position centroid algorithm crash K-Means. The main goal is to fix weaknesses K-Means in the initial centroid initialization that often results in a local solution. With Ant Colony Optimization, position centroid Optimized for quality clustering (e.g. using Silhouette Score), so that the grouping results become more accurate and stable [29]. Application Ant Colony Optimization (ACO) in this study is used to optimize the position centroid algorithm K-Means, so that the grouping of data becomes more accurate and stable. The ACO-based clustering process was carried out through the following stages:

1) Initialize parameters, such as α, β, number of ants. Parameter initialization refers to the research conducted ρ[11]. However, in this study, the parameter values taken were α = 1 and 2, β = 2 and 3, and 0.5 and 0.7, the number of ants was taken as many as 50 and 100. The complete parameter configurations are shown in Table 10.
2) Initialize the initial pheromone value for each possible centroid position.
3) The virtual ants move in search of a solution, which is to determine the position of the centroid.
4) Each ant chose a position based on the pheromone trail, which indicates the quality of the position from the ant's previous experience, and heuristic information, such as the distance between the data and the centroid, so the ant prefers a good position and close to the data.
5) The choice of an ant is determined by probability, which is how likely it is that the ant chooses a position. These probabilities are calculated based on the strength of pheromones and heuristic values, with weights of each using Equation (2).
6) Update the value of pheromones based on the best cluster results with the press Equation (3).

Table 10. Parameters used Ant Colony Optimization

Cluster	a	b	ρ	N_ant	Silhouette Score
2	1	2	0.5	50	**0,89**
2	1	2	0.5	100	0,89
...	...	...	...	...	...
2	2	3	0.7	100	**0.89**
3	1	2	0.5	50	0.86
...	...	...	...	...	...
3	2	3	0.7	100	0.86
4	1	2	0.5	50	0.88
4	1	2	0.7	100	0.88
...	...	...	...	...	...
4	1	3	0.5	50	**0.92**
4	1	3	0.5	100	0.92
...	...	...	...	...	...
4	2	3	0.7	100	0.88

7) Evaluate with silhouette score.

The best results for each cluster size summarized in Table 11. The highest overall Silhouette Score of 0.92 was achieved with K = 4, indicating that Ant Colony Optimization successfully improves clustering quality and stability.

Table 11. Best Silhouette Score Aco

Cluster	a	b	ρ	N_ant	Best Silhouette Score
2	1	2	0.5	50	0,89
3	1	2	0.5	50	0,86
4	1	3	0.5	50	0.92
		Best Silhouette Score ACO			**0.92**

3.6 Comparison of K-Means and Ant Colony Optimization

A direct comparison between standard K-Means and Ant Colony Optimization is presented in Table 12 and visualized in Figure 4. The comparison shows that Ant Colony Optimization consistently improves clustering quality across all tested values of K. For K

= 4, Ant Colony Optimization improves the Silhouette Score from 0.81 to 0.92, representing the most significant gain among all configurations.

Table 12. Comparison K-Means & ACO

Cluster	Silhouette Score K-Means	Silhouette Score Ant Colony Optimization
2	0,88	0,89
3	0,81	0,86
4	0,81	0,92
Best Cluster **4**		

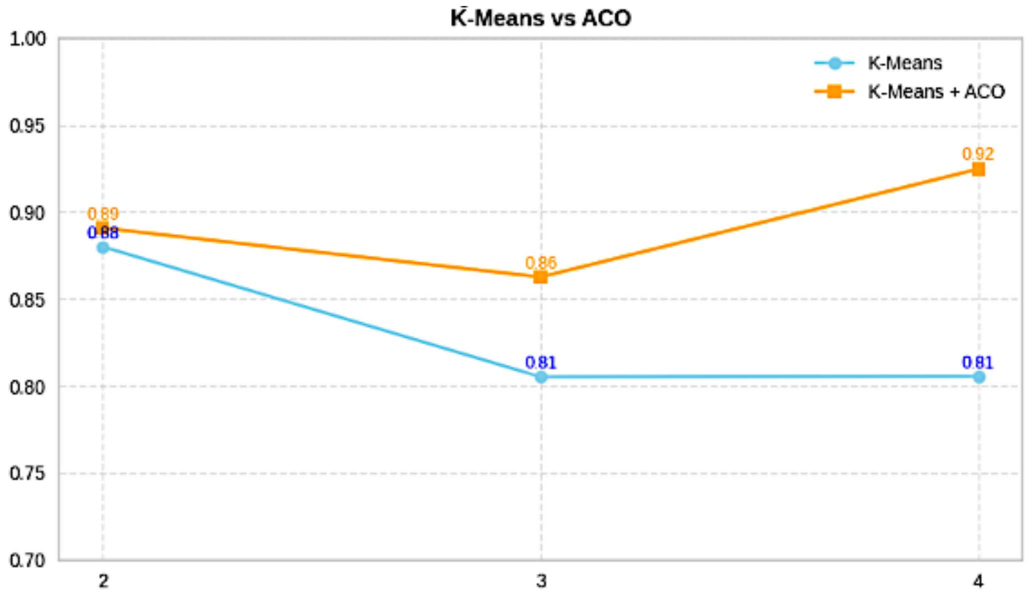


Figure 4. Comparison K-Means & ACO

Figure 5 illustrates the cluster visualization produced by K-Means, while Figure 6 shows the more compact and well-separated clusters obtained after Ant Colony Optimization.

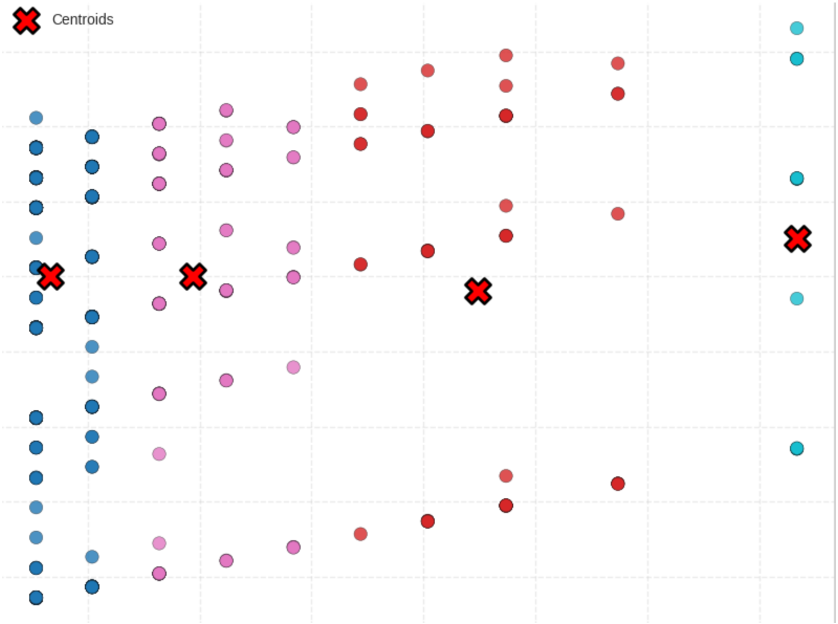


Figure 5. The result of grouping with K-Means

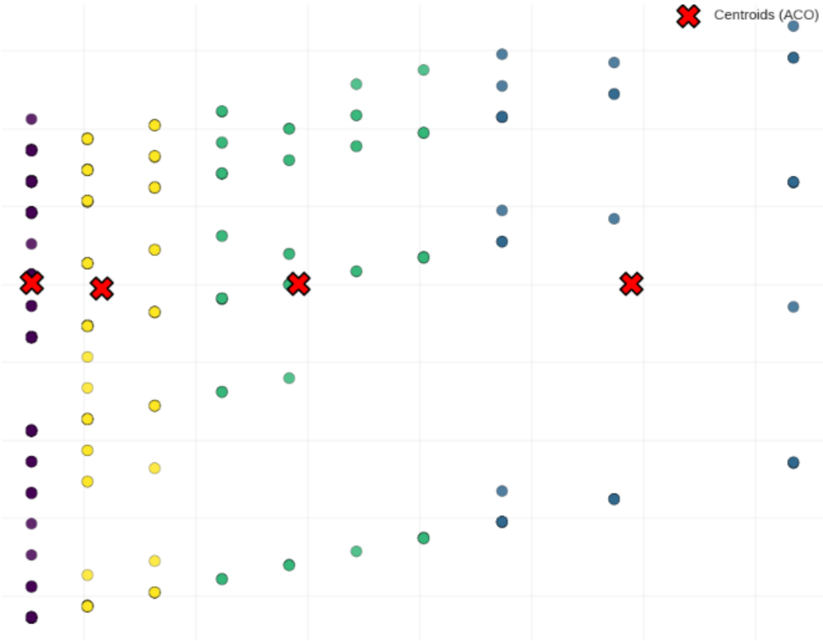


Figure 6. Grouping Results ACO

3.7 Interpretation

The characteristics of each group member obtained are that cluster 1 has an income of 85 to more than 150 million, the last education is dominated by s1 by 25.9%, 60.3% owns own land, the majority has BPJS as much as 62.1%, is very active in using WhatsApp and social media such as Facebook or Instagram, some even use Shopee, Gojek and Grab. Receiving government assistance by 19%, with gender dominated by women at 60%. Cluster 2 has an income of 10 to 25 million, the last education is dominated by high school 28%, vocational school 17.7%, 49.1% of the status of own land ownership, with 70.6% of members having BPJS. Almost all of the media facilities used use WhatsApp, a small part uses Gojek, Grab, and Instagram. Receiving 34.5% of assistance from the Yogyakarta city government and the central government, dominated by 67% women and 29% by men. Cluster 3 with an income of 25 to 85 million, dominated by the last education of high school 23.2%, S1 21.6%, with 50% of land or buildings owned by themselves, 80.8% have BPJS as health insurance, electronic media facilities use WhatsApp, Instagram, Shopee, Gojek, and Grab, receive government assistance of 36.8%. Women dominate as much as 65.6% and men 32%. Cluster 4 has an income of less than 10 million, with the last education being high school 29.4%, vocational school 20.8%. The status of land and buildings is 52% owned, with 67% owning BPJS for health insurance. The use of technology is not good where generally WhatsApp and only a few uses other media. It received 31.3% of government assistance, especially the central government and 70% of its members were women. The characteristics of each cluster resulting from the best ACO configuration (K = 4) are summarized in Table 13.

Table 13. Summary of Cluster Characteristics and Implications

Cluster	Characteristics	Managerial / Policy Implications
1	High income (85-150+ million), dominant S1 education, high land ownership, active multiplatform digital usage	Suitable for advanced digital marketing support, export facilitation, and innovation-driven programs
2	Low-middle income (10-25 million), high school/vocational education, moderate land ownership, WhatsApp dominant	Requires basic financial access, digital literacy training, and local government assistance
3	Middle income (25-85 million), mixed education (SMA-S1), strong BPJS coverage, active e-commerce usage	Potential for scaling through marketplace integration, logistics

Cluster	Characteristics	Managerial / Policy Implications
		support, and business mentoring
4	Very low income (<10 million), low technology adoption, limited education, high dependency on assistance	Priority target for empowerment programs, micro-financing, and foundational capacity building

3.8 Discussion

The four-cluster solution demonstrates the most robust structure, achieving the highest Silhouette Score (0.92), which reflects both excellent intra-cluster cohesion and inter-cluster separation. This result highlights the superior clustering quality provided by the four-cluster configuration compared to K = 2 and K = 3. While K = 2 may oversimplify the MSME landscape by grouping distinct business profiles into broad categories, K = 4 offers a balanced trade-off between granularity and interpretability. This clustering configuration allows for a more nuanced understanding of MSME characteristics, including variations in scale, technology adoption, and support dependency. Consequently, the four-cluster solution enables policymakers and stakeholders to design more targeted interventions while avoiding unnecessary complexity.

From a methodological perspective, the Silhouette Score, a key indicator of clustering quality, reveals that the four-cluster model outperforms both the two- and three-cluster models in terms of data differentiation and accuracy. With a higher Silhouette Score, the four-cluster solution better captures the heterogeneity of MSMEs, facilitating the identification of distinct groups with differing needs and capabilities. This increased resolution provides a clearer view of the unique characteristics of MSMEs in the culinary sector, which is essential for informed policy-making.

In practical terms, fewer clusters (e.g., K = 2) risk overlooking crucial variations within the MSME data, potentially leading to generalized, one-size-fits-all strategies. The four-cluster model, however, offers sufficient detail without becoming overly complex, making it a more effective tool for decision-makers seeking to support MSMEs with tailored interventions. This model's clear differentiation of MSME characteristics—such as income levels, educational backgrounds, and technology adoption—aligns with the need for more precise, data-driven strategies to strengthen MSME development and competitiveness, particularly in a dynamic and competitive environment like the culinary sector in Yogyakarta.

By integrating clustering techniques with Ant Colony Optimization for centroid initialization, this study advances the potential for more accurate and reliable clustering, especially for complex datasets like MSME data, where traditional K-Means clustering may fall short. The results of this study suggest that the optimal configuration (K = 4) not only improves clustering accuracy but also provides valuable insights for designing interventions that are both effective and feasible.

4. CONCLUSION

Based on the results of the research conducted, Ant Colony Optimization was able to optimize and improve the performance of the K-Means algorithm in the data clustering process of Culinary MSMEs in the Special Region of Yogyakarta. The use of Ant Colony Optimization has been proven to be able to optimize the centroid initialization process so that the grouping results are stable and accurate. The results of the comparison showed that the silhouette score value produced by the K-Means method increased after optimization with Ant Colony Optimization, namely from 0.88 to 0.89 for two clusters, 0.81 to 0.86 for three clusters, and from 0.81 to 0.92 for four clusters. From these findings, it can be concluded that the combination of the K-Means and Ant Colony Optimization methods not only improves the accuracy of the grouping results, but also provides a clearer picture of the profile and characteristics of Culinary MSME actors.

This study has several limitations. First, the clustering analysis was conducted using a single dataset limited to Culinary MSMEs in the Special Region of Yogyakarta, which may restrict the generalizability of the results to other regions or business sectors. Second, the evaluation of clustering quality relied solely on the Silhouette Score. Although this metric effectively measures cluster cohesion and separation, using a single evaluation index may not fully capture all aspects of clustering performance, such as cluster density or distribution balance.

Future research may expand this study by applying other metaheuristic optimization algorithms, such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA), or Firefly Algorithm, to further compare and enhance centroid optimization performance. Additionally, future studies could use MSME datasets from different regions or sectors to improve the generalizability of the findings. Another promising direction in the development of a real-time or decision-support clustering system that can dynamically update MSME groupings as new data become available, thereby supporting more adaptive and data-driven policy interventions.

REFERENCES

[1] U. Linarti, A. H. Soleliza Jones, L. Zahrotun, and A. Rahmawati, “Penerapan Metode K-Medoids Guna Pengelompokan Data Usaha Mikro, Kecil dan Menengah (UMKM) Bidang Kuliner Di Kota Yogyakarta,” *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 37–45, 2024, doi: 10.55338/jikomsi.v7i1.2194.

[2] C. Andriani, “Pemberdayaan Umkm Dengan Pendaftaran Nomor Induk Berusaha Melalui Oss Di Kelurahan Krembangan Selatan Surabaya,” *PATIKALA J. Pengabdi. Kpd. Masy.*, vol. 2, no. 1, pp. 406–413, 2022, doi: 10.51574/patikala.v2i1.487.

[3] Badan Pusat Statistik Provinsi Daerah Istimewa Yogyakarta, “Perkembangan Pariwisata Daerah Istimewa Yogyakarta Desember 2024,” *Berita Resmi Statistik*, no. 11, pp. 1–12, Feb. 2025.

[4] N. A. Sudibyo, M. Najb, M. S. Andrianto, E. Alimovich, M. Marhadi, and A. Boros, “The Rise of ASEAN SMEs: How to Successfully Enter the Global Market,” *Preprints*, Jul. 2023.

[5] C. Zai, “Implementasi Data Mining sebagai Pengolahan Data,” *J. Portal Data*, vol. 2, no. 3, pp. 1–12, 2022.

[6] E. Muningsih, I. Maryani, and V. R. Handayani, “Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Davies–Bouldin Index untuk Clustering Provinsi Berdasarkan Potensi Desa,” *J. Sains dan Manajemen*, vol. 9, no. 1, pp. 96–104, 2021.

[7] S. Andriani, "Distribution Analysis Active Small and Medium Industries Bogor City Using K-means Clustering," *Komputasi J. Ilm. Ilmu Komput. dan Mat.*, vol. 20, no. 1, pp. 56–70, 2022, doi: 10.33751/komputasi.v20i1.6559.

[8] M. Hernita, E. Pramesty, Y. U. Aprilia, R. Bryan, J. Purba, and M. Athoillah, "Klasterisasi Kabupaten dan Kota Menggunakan Algoritma K-Means dengan Metode Elbow dan Silhouette Score," in *Seminar Nasional Hasil Riset dan Pengabdian*, Indonesia, 2023, pp. 359–368.

[9] I. Yaputera, R. Hanafi, and M. Rusman, "Optimasi Rute Kunjungan Cluster Sales Officer (CSO) Menggunakan Ant Colony Optimization (ACO)(Studi Kasus: Indosat Ooredoo Hutchison Micro Cluster Mamuju)," *J. Penelit. Enj.*, vol. 26, no. 2, pp. 82–90, 2022, doi: 10.25042/jpe.112022.04.

[10] E. Febianti, Y. Muharni, D. Falti, L. Herlina, and K. Kulsum, "Usulan Penjadwalan Mesin Paralel Menggunakan Metode Ant Colony Optimization Algorithm dan Longest Processing Time," *J. Integr. Syst.*, vol. 6, no. 1, pp. 42–52, 2023, doi: 10.28932/jis.v6i1.5610.

[11] R. Y. C. Sianturi, B. Rahayudi, and A. W. Widodo, "Implementasi Algoritma Ant Colony Optimization untuk Optimasi Rute Distribusi Produk Kebutuhan Pokok dari Toko Sasana Bonafide Mojoroto," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 7, pp. 3190–3197, 2021.

[12] D. J. Ratnaningsih, "Clustering with K-Means Hybridization Ant Colony Optimization (K-ACO)," *Int. J. Math. Model. Comput.*, vol. 12, no. 2, pp. 143–152, 2022.

[13] I. Sari, Y. Maulita, L. Arliana, and N. Kadim, "Pengelompokan UMKM Kota Binjai Menggunakan Metode Clustering K-Means untuk Identifikasi Pola Perkembangan Bisnis," *Jurnal Sistem Informasi*, vol. 3, no. 1, pp. 45–52, 2024.

[14] M. Tonggiroh, S. Nurhayati, Yakub, and Jusmawati, "Sistem Pendukung Keputusan Pemilihan Wireless Router Menggunakan Pendekatan Rank Reciprocal dan ARAS," *J. Fasilkom*, vol. 14, no. 1, pp. 206–215, 2024, doi: 10.37859/jf.v14i1.6838.

[15] S. Widaningsih, "Penerapan Data Mining untuk Memprediksi Siswa Berprestasi dengan Menggunakan Algoritma K Nearest Neighbor," *JATISI (Jurnal Tek.*

Inform. dan Sist. Informasi), vol. 9, no. 3, pp. 2598–2611, 2022, doi: 10.35957/jatisi.v9i3.859.

[16] D. Andriansyah, "Implementasi Extract-Transform-Load (ETL) Data Warehouse Laporan Harian Pool," *J. Teknik Informatika*, vol. 8, no. 2, pp. 45–49, Jun. 2022, doi: 10.51998/jti.v8i2.486.

[17] S. A. Salasa and W. Maharani, "Personality Detection of Twitter Social Media Users using the Support Vector Machine Method," *J. Sist. Komput. dan Inform.*, vol. 4, no. 2, p. 263, 2022, doi: 10.30865/json.v4i2.5345.

[18] D. Prasetyawan and R. Gatra, "Analisis Cluster untuk Pengelompokan Kemampuan Penguasaan ICT Menggunakan K-Means dan Autoencoder," *JISKA*, vol. 10, no. 2, pp. 145–157, 2025, doi: 10.14421/jiska.2025.10.2.145-157.

[19] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[20] J. C. Quiroz et al., "Extract, Transform, Load Framework for the Conversion of Health Databases to OMOP," *PLoS One*, vol. 17, no. 4, pp. 1–13, Apr. 2022, doi: 10.1371/journal.pone.0266911.

[21] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Deep Learning," in *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA: Morgan Kaufmann, 2017, ch. 10, pp. 389–420.

[22] P. A. Leonardo, M. Arifin, and S. Y. Siswanto, "Penerapan Euclidean Distance untuk analisis driver variable Perubahan Penggunaan Lahan dari Jarak Jalan di Sub-DAS Cikapundung," *Soilrens*, vol. 22, no. 1, pp. 61–66, 2024, doi: 10.24198/soilrens.v22i1.57248.

[23] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan DBSCAN," *J. Inform. dan Teknik Elektro Terapan*, vol. 12, no. 3S1, pp. 60–74, 2024, doi: 10.23960/jitet.v12i3s1.5001.

[24] N. Alfa Husna et al., "Implementasi Algoritma Ant Colony Optimization untuk Penentuan Jalur Terpendek Klinik," in *Seminar Nasional SENTIMAS*, Indonesia, 2023, pp. 112–119.

[25] N. Rohman and A. Wibowo, "Perbandingan Metode K-Medoids dan Metode K-Means Dalam Analisis Segmentasi Pelanggan Mall," *SINTECH (Science Inf. Technol. J.)*, vol. 7, no. 1, pp. 49–58, 2024, doi: 10.31598/sintechjournal.v7i1.1507.

[26] W. Lidysari, H. S. Tambunan, and H. Qurniawan, "Penerapan Data Mining Dalam Menentukan Kelayakan Penerima Bantuan Sosial Pemko Dengan Algoritma C4.5 (Kasus Kantor Kelurahan Martoba)," *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 3, no. 1, pp. 53–61, 2022, doi: 10.30645/kesatria.v3i1.97.

[27] D. N. Yoliadi, "Data Mining dalam Analisis Tingkat Penjualan Barang Elektronik Menggunakan Algoritma K-Means," *Insearch Journal*, vol. 3, no. 1, Feb. 2023.

[28] S. Harris and R. C. De Amorim, "An Extensive Empirical Comparison of k-means Initialization Algorithms," *IEEE Access*, vol. 10, no. 2, pp. 58752–58768, 2022, doi: 10.1109/ACCESS.2022.3179803.

[29] S. Wang, "Ant Colony Optimization for Clustering College Students' Physical Exercise Behavior Patterns," *Inform.*, vol. 48, no. 20, pp. 179–190, 2024, doi: 10.31449/inf.v48i20.6566.

Revision

Reviewer A

No	Reviewer	Revision Locate
A1	Suggested revision: "Centroid Optimization of K-Means Using Ant Colony Optimization for Culinary MSME Clustering"	Title section
A2	Suggestions: 1. Reduce repetition related to K-Means limitations. 2. Add one sentence explaining why improved clustering is important for MSME decision-making. 3. Improve clarity and academic tone through language editing. 4. Keywords: Consider adding "Clustering Optimization" or "Business Data Analytics" for better indexing.	1. Abstract section (sentence 4) 2. Abstract section (sentence 3) 3. Abstract section 4. Keyword section
A3	1. Add a concise research gap paragraph, explicitly stating what previous studies lack. 2. Clearly list research contributions, such as: - Optimization of K-Means centroid initialization using ACO - Empirical evaluation using silhouette score - Interpretation of culinary MSME cluster characteristics	1. Introduction section (paragraph 6) 2. Introduction section (paragraph 7) 3. Introduction section

	3. Improve language consistency.	
A4	Suggestions 1. Figure 2 unclear, transform to table or enhance figure quality 2. Dataset Description: Clarify data provenance and collection period. also Briefly mention data privacy or anonymization. 3. Research Stages: Reduce repetition between “Selection” and “Preprocessing” explanations. also Improve figure resolution and caption clarity. 4. Data Transformation: Briefly justify why manual mapping is suitable. also Mention transformation limitations. 5. K-Means Clustering: Justify distance metric choice. also Mention why K-Means++ was not used as a baseline. 6. Ant Colony Optimization (ACO): Briefly explain how parameter values were selected. also Discuss runtime or computational overhead as a limitation. 7. Every step in Figure 2 must be explained in details, also cite some article in this section	1. Methods section (research stage, figure 2 transform to table) 2. Methods section (Culinary MSME Data) 3. Methods section (research stages, Explained stages of research process no 2 and 3). For the record, Figure 2 became Figure 1 after revision, Figure 1 was previously changed to a table in the Culinary MSME Data section, and Figure 2 after revision is the result of data loading. 4. Methods section (research stages, Explained stages of research process no 4 sentence 3-5) 5. Methods section (research stages, Explained stages of research process no 5 paragraph 3) 6. Methods section (research stages, Explained stages of research process no 7 paragraph 2 and 3) 7. Methods section
A5	Suggestions 1. Explain why 4 clusters	1. Result and discussion (Discussion, paragraph 1 and 2)

	produce the best structure. also Discuss how improved clustering supports MSME policy or business strategies. 2. Summarize cluster characteristics in a table. also Highlight key managerial or policy implications. 3. Every Figure or Table must be mentioned in paragraphs, also must have narrative 4. ADD DISCUSSION SECTION to analysis "WHY?"	2. Result and discussion (Interpretation, Table 13) 3. Result and discussion 4. Result and discussion (Discussion)
A6	Suggestions 1. Add a limitations paragraph (single dataset, silhouette-only evaluation). 2. Suggest future work (other optimization algorithms, different regions, real-time systems).	1. Conclusion (paragraph 2) 2. Conclusion (paragraph 3)
A7	Suggestions 1. Standardize reference formatting. 2. Ensure consistency with journal citation guidelines. 3. Add more citation, minimum 25. also add all article metadata	References

Reviewer B

No	Reviewer	Revision Locate
B1	Capitalization and wording need improvement	Title page
B2	Too descriptive	Abstract section

B3	Research gap not clearly summarized	Introduction (paragraph 6)
B4	Parameter selection unclear	Methods section (research stages, Explained stages of research process no 7 paragraph 2)
B5	Limited discussion of implications, also add discussion section	Result and discussion (Discussion)
B6	No explicit limitations	Conclusion
B7	No explicit limitations	References