Random perturbation bee colony-optimized k-means approach for optimized MSME data clustering

Lisna Zahrotun [a,1,*], Ika Arfiani [a,2], Dwi Normawati [a,3]

[a] Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1] lisna.zahrotun@tif.uad.ac.id; [2] ika.arfiani@tif.uad.ac.id; [3] dwi.normawati@tif.uad.ac.id
* corresponding author

ARTICLE INFO

Article history
Received October 21, 2024
Revised May 22, 2025
Accepted March 1, 2026
Available online August 31, 2026

Keywords
MSMES
Clustering
K-Means
RPBCO
Market expansion

ABSTRACT

Micro, Small, and Medium Enterprises (MSMEs) play an important role in the economy and need strategic support to grow. The K-Means method is often used for cluster analysis, but it has a weakness in determining the initial cluster Centre. This research proposes the K-Means Random Perturbation Bee Colony Optimization (RPBCO) method to overcome the problem. The test results show an increase in cluster accuracy, with the Silhouette Coefficient score increasing by 40.9% (from 0.171 to 0.245). Wilcoxon testing also showed a z-value of -1.342, confirming that RPBCO is superior to standard K-Means. This method proved effective in optimizing clusters on a heterogeneous MSMEs dataset. The analysis revealed that creative MSMEs thrive on Instagram, while retail MSMEs perform best on Shopee. To expand the market and reach MSMEs beyond Java and internationally, marketing strategies tailored to the characteristics of the region and target market are required. Creative MSMEs can utilize Instagram Ads to target ASEAN countries, while retail MSMEs can focus on global platforms such as Amazon or Alibaba. Digital literacy training, platform algorithm workshops, and collaborative marketing campaigns can strengthen this strategy. These measures are expected to increase turnover and support the sustainable growth of MSMEs, positively impacting more than one million businesses.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

Micro, Small and Medium Enterprises (MSMEs) are activities managed by individuals or institutions that have a vital role in the country's economic growth. When the economic crisis hit Indonesia, many large companies were forced out of business, but the MSMEs sector was able to survive and become a bastion of economic resilience in the face of the crisis [1]. MSMEs play a crucial role as a buffer for the economy in the face of various crises [2]. The Department of Industry, Cooperatives, and Small and Medium Enterprises is the agency responsible for managing the MSMEs sector [3]. With such a large number of MSMEs, mapping their condition becomes indispensable. MSMEs data analysis has various benefits, including formulating MSMEs development strategies [4], categorizing data based on their characteristics [5], assessing the competence of resources in each MSME [6], and understanding the level of competition among MSMEs [4].

Financial management is one of the main keys to the success of MSMEs [5]. In the post-COVID-19 pandemic era, the use of information technology and social media is increasingly crucial to encourage the development of MSMEs [6]. The government also continues to strive to strengthen the role and

competitiveness of local MSMEs amid global economic challenges [7]. Seeing the importance of data analysis to understand the potential and challenges of MSMEs, this research focuses on clustering MSME data. This approach is expected to facilitate the formulation of more effective strategies, especially in improving marketing, driving revenue growth, and maximizing the use of marketing media.

K-Means is one of the oldest and most widely used clustering techniques. This method works through a series of iterations that will continue until no cluster member moves to another cluster [8]. The cluster center point, or centroid, serves as the starting point to begin clustering in the K-Means algorithm. The data clustering process is done by calculating the closest distance from the data to the initial cluster center point, which acts as the midpoint in group formation [9]. The result of the clustering process with the K-Means method is highly dependent on the determination of the initial cluster center point [10]. Using various methods to determine the initial center point can improve the accuracy of the output results of this clustering process. One of the approaches used is the Genetic Algorithm (GA) [11], which is proven to provide better accuracy compared to random center point determination. In addition, the use of structural features in determining the initial center point also shows more accurate results with the application of Minkowski Distance for calculating the distance of data to the centroid [12]. Another study using Sum of Squared Error to determine the cluster center point in the K-Means algorithm recorded clustering quality with a Davies-Bouldin Index value of 0.0378, or 3.78% better than the conventional K-Means method [12]. Bee Colony Optimization (BCO) has also been applied in cluster center point selection, where the results of BCO K-Means show higher accuracy compared to random initial center point determination [13]. The use of the BCO method has previously been tested and proven to be able to improve the performance of the type-2 fuzzy method with a significant increase of 21.7% [14].

Based on the advantages of the Bee Colony Optimization (BCO) method that have been proven in previous studies, this research uses the Random Perturbation Bee Colony Optimization (RPBCO) approach to determine the starting point of the cluster center. In the BCO method, there are various initialization techniques that can be used. One of them is initialization using a Gaussian distribution, which has been tested previously and shows more competitive results [15]. Therefore, this study adopts a random initialization approach through the random perturbation method to improve the performance and flexibility of the algorithm.

The data obtained from the process of determining the cluster centroid using the BCO method will be used as a reference to start the clustering process with the K-Means algorithm. To evaluate the difference in clustering results produced by these methods, tests were conducted using the Silhouette Coefficient.

2. Method

This research aims to optimize the determination of the cluster center point in the K-Means method using RPBCO in grouping MSMEs data.

2.1. Dataset

This research analyses 1,339 data points from the MSMEs dataset in the culinary sector managed by the Department of Industry, Cooperatives, Small and Medium Enterprises, Yogyakarta. The dataset includes 12 attributes and has been published on the Hugging Face platform to support further research, and it is attached in the Data Availability Section.

2.2. Research Stages

This paper is organized to provide a clear and logical flow. Section 1 introduces the research background, objectives, and challenges in MSME clustering. Section 2 elaborates on the dataset and preprocessing steps, ensuring data quality and representativeness. Section 3 describes the RPBCO K-Means methodology in detail, including its unique advantages over other initialization methods. Section 4 presents results and a thorough discussion, including real-world implications for MSME development. Finally, Section 5 provides conclusions and actionable recommendations for policymakers and

practitioners, emphasizing strategies for optimizing MSMEs' market reach and turnover. The overall flow of these stages can be seen more clearly in Fig. 1.

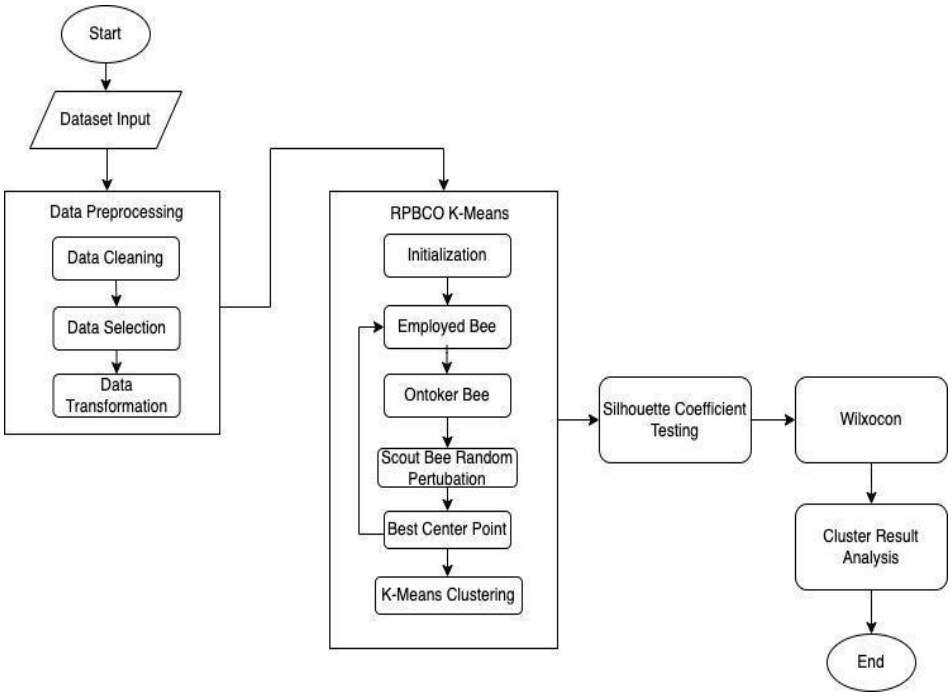


Fig. 1. Research stage using RPBCO K-Means

2.2.1. Dataset Input

This stage involves inputting the research dataset, which is the MSMEs data that is the object of research.

2.2.2. Data Preprocessing

At this stage, the data goes through several processing steps before clustering. The first step is data cleaning, which removes noise and checks for missing data. Next, data selection is carried out, which selects important variables from the MSMEs data that will be grouped. The last step is data transformation, where the data is converted into a numeric format to make it suitable for the clustering process.

2.2.3. RPBCO K-Means Process

In this stage, MSMEs data is clustered using the K-Means method, with the determination of the initial centroid optimized using the RPBCO method. The RPBCO K-Means process consists of six steps, namely Initialization, Employed Bee, Onlooker Bee, Scout Bee, Selection of the best center, and ends with clustering using K-Means.

Bee Colony Optimization is an optimization algorithm inspired by the way bees search for food. Bees are social insects that have excellent organization [16]. In this approach, bee colonies work together to find the optimal solution to the problem at hand [13]. In this research, the Random Perturbation method is used in the Bee Colony Optimization process. Here are the steps in RPBCO K-Means [17]:

1) Initialization Phase. In this phase, initial food sources are randomly initialized or randomly perturbed, with each food source representing one centroid. Each food source will be equipped with one Employee bee to start the search.

2) Employee Bee Phase. Each Employee bee is tasked with searching for new food sources that have more nectar around the current location. Using the K-Means algorithm, they perform clustering and calculate fitness values. Next, they apply greedy selection to evaluate the fitness value of the new food source and compare it with the previous one. The better food source will be accepted by the onlooker bee.

3) Onlooker Bee Phase. In this phase, the onlooker bee waits for information about the food source from the Employee bee. They then perform the second step with the aim of combining the information into the food source with the highest probability for each cluster.

4) Scout Bee Phase. In this phase, the Scout bee generates new food sources as long as the set of food sources has not reached the parameter limit. This phase serves as a bridge between the first iteration and the next iteration, where the Random Perturbation method is used to create new food sources.

5) Selection of the Best Food Source. The three phases-Employee bee, Onlooker bee, and Scout bee-will be repeated continuously until they reach the iteration limit. From all iterations, the best food source will be selected as the final solution, and its fitness value will be calculated. The best fitness value will be the optimal solution determined by the bee.

6) Clustering Using the K-Means Method. This research applies the RPBCO algorithm to optimize the K-Means process in clustering, with the aim of improving the quality and efficiency in determining the initial centroid. The performance of K-Means is greatly affected by the improper selection of the initial centroid, which may result in inaccurate clustering results or convergence to a local solution. To overcome these problems, RPBCO utilizes bee-based exploration strategies (employed bee, onlooker bee, and scout bee) to find better initial solutions. The results show that the initial centroid generated by the BCO algorithm has a more optimal quality compared to the random selection method in conventional K-Means.

After the completion of the solution search process using the RPBCO algorithm, the optimal initial centroid is selected based on the highest fitness value of the resulting solution population. This centroid replaces the random initialization method in the K-Means algorithm, so that the K-Means iteration process starts with the optimized centroid. To evaluate the performance of RPBCO, several metrics are used, including the Sum of Squared Errors (SSE), which calculates the squared error between a data point and its centroid, and the Silhouette Coefficient, which assesses the cluster quality based on the closeness between points in the cluster and the distance to points in the nearest cluster.

The implementation of the RPBCO algorithm in K-Means begins with the selection of the best centroid based on a fitness function that refers to the Euclidean distance or other error metrics. The centroid optimization process is carried out iteratively through the mechanism of worker bees, onlooker bees, and scout bees to update the solution until it reaches the optimal condition. The optimal centroid obtained is then used as the initial input for the K-Means algorithm, replacing the random centroid initialization method. Furthermore, the K-Means iteration is continued until it reaches convergence, which is a condition in which changes in the centroid position become insignificant.

In general, K-Means is an algorithm used for data clustering. The success in forming clusters is highly dependent on the initialization of the initial center point value, or centroid [18]. The steps performed in the K-Means algorithm [19], begin by determining the optimal number of clusters (K) using the RPBCO algorithm, followed by initializing the cluster centroids as the starting points for clustering. Subsequently, the Euclidean distance between each data point and the cluster centroids is calculated to measure their similarity. Based on these distances, each data point is assigned to the cluster with the nearest centroid. Once all data points have been allocated to their respective clusters, the centroids are updated by calculating the mean of the data points within each cluster using the following equation:

$$\bar{v}_{ij} = \frac{1}{N_i}\sum_{k=0}^{n} x_{kj} \tag{1}$$

In Equation (1), C_{ij} represents the centroid of the *i-th* cluster for the *j-th* variable, while n_i denotes the total number of data points assigned to the *i-th* cluster. The indices i, j, and k represent the cluster, variable, and data point within the cluster, respectively, where k ranges from 1 to n_i. Furthermore, x_{ikj} denotes the value of the *j-th* variable for the *k-th* data point belonging to the *i-th* cluster. The centroid is calculated as the arithmetic mean of the corresponding variable values across all data points within each cluster, serving as the representative center for subsequent clustering iterations.

2.2.4. Testing

This stage measures the accuracy of the clustering results using the Silhouette Coefficient, which serves to assess how well the data has been grouped. Silhouette Coefficient is a measure used to assess the quality of clustering by assigning a value between -1 and 1. A value close to 1 indicates that the data is in the right cluster, while a value close to -1 indicates that the data may have been misclustered. This method calculates the distance between the data point and the centroid of the corresponding cluster, as well as the distance to the nearest cluster. By analyzing the ratio between the two, the Silhouette Coefficient gives a clear picture of how well a clustering works. The use of Silhouette Score is helpful in choosing the optimal number of clusters and evaluating clustering results objectively [20].

This method aims to measure the level of closeness between data or objects in a cluster. The following are the steps in the silhouette coefficient calculation process [21]:

- Calculate the average distance between a particular document, e.g., document i, and all other documents belonging to the same cluster [22].

$$a(i) = 1\,|A| \; - 1 \sum j \in A, j \neq i \; d(i,j) \tag{2}$$

- Calculate the average distance from document i to all documents in different clusters, then take the smallest value of the result [22].

$$d(i,C) \; = \; 1\,|A| \sum j \epsilon \; C \; d(i,j) \tag{3}$$

with $d(i,C)$ is the average distance of document i to all objects in another cluster C where A≠C [26]

$$b\,(i) \; = \; minC \; \neq \; A \; d(i,C) \tag{4}$$

- Calculate the Silhouette Coefficient value to evaluate the quality of the clustering performed

$$s(i) = b(i) - a(i)\; max(a(i), b(i)) \tag{5}$$

where s(i) represents the silhouette value of the *i-th* data point, while a(i) denotes the average distance between the *i-th* data point and all other data points within the same cluster (cluster A). Meanwhile, b(i) represents the minimum average distance between the *i-th* data point and all data points in any other cluster, indicating the nearest neighboring cluster. The Silhouette Coefficient measures clustering quality by evaluating both intra-cluster cohesion and inter-cluster separation, with values ranging from −1 to 1, where values closer to 1 indicate better clustering performance.

2.2.5. Wilcoxon

The Wilcoxon method is one of the techniques in nonparametric statistics designed to compare two groups of paired data or test hypotheses related to population medians. Developed by Frank Wilcoxon in 1945, this method has become a reliable analytical tool in situations where the assumption of normal distribution is not met. As a nonparametric approach, the Wilcoxon test provides flexibility in analyzing paired data in a way that is robust to different forms of data distribution [23].

The principles and hypotheses in the Wilcoxon Test focus on comparing the median values of two paired samples. In the context of clustering, the samples being compared are usually the values of clustering result evaluation metrics, such as silhouette score, inertia, or Davies-Bouldin index, obtained from applying two different algorithms on the same dataset. The hypotheses tested in this test include.

- Null hypothesis (H_0): There is no significant difference between the clustering results of method A and method B, or the median difference between the two samples is equal to zero.
- Alternative hypothesis (H_1): There is a significant difference between the clustering results of method A and method B, or the median difference between the two samples is not equal to zero

2.2.6. Cluster Result Analysis

This is the stage of analyzing the clustering results performed using the RPBCO K-Means method. This analysis aims to understand the pattern of MSMEs data groups, which can then be used as a basis for formulating strategies and developing MSMEs.

3. Results and Discussion

The dataset used in this study consists of 1,339 MSME entries across diverse industries and regions within Yogyakarta, including retail, manufacturing, and services sectors. This comprehensive dataset captures demographic, economic, and operational characteristics, ensuring representativeness for broader MSME populations. Stratified sampling was used to ensure proportional representation of each business type and geographic location. Furthermore, sensitivity analysis was performed to validate the robustness of clustering results, confirming the generalizability of findings to MSMEs beyond the study area. After the data is obtained, the next step is to preprocess the data, which includes data cleaning, selection of relevant variables, and data transformation.

3.1. Preprocessing Data

At this stage, the data preparation process is carried out to make it ready for clustering, which includes several preprocessing steps. First, the data cleaning process aims to clean the dataset of data that contains noise or is not suitable for use in the analysis process. In the culinary dataset consisting of 1339 data points, there was a reduction to 1336 data points after the cleaning process. This reduction is caused by missing values or empty data found in the initial dataset. Second, the data selection process was carried out to select relevant variables from all variables contained in the dataset. From a total of 40 variables, 11 variables were selected for use. The results of variable selection include: Last Education, Business Activities, Marketing Objectives, Land/Building Ownership Status, Electronic Media Facilities, Government Aid Capital, People's Business Credit Loans, Turnover per Year, Health Insurance Ownership, Labor, and Female Labor. Third, the Data Transformation Process is a transformation from text format to numeric format in order to be processed further. The variable column lists the type of variable that was transformed, while the transformation result shows how the variable was converted into a numeric value.

3.2. Application of RPBCO K-Means on MSMEs Data Clustering

The use of the K-Means method in clustering MSMEs data has been recognized as effective, but this method has a weakness in the initialization of the centroid, which affects the final result. To overcome this limitation, Random Perturbation Bee Colony Optimization (RPBCO) is introduced as a better optimization technique for determining cluster centroids [24], [25]. In this process, Random Perturbation is an important step in the exploitation phase, where mutations are performed randomly to combine centroid values. This approach not only increases the chance of finding the optimal solution, but also makes the clustering process more dynamic and accurate compared to the conventional K-Means method. The stages of applying RPBCO K-Means in this study are as follows:

1) Initialization Phase

In this phase, the initial initialization of the centroid position is done randomly. This centroid position represents the food source to be explored by the bee colony. Each centroid is taken randomly in the dataset feature space consisting of MSMEs. data. In Table 1, the initial centroid results are shown.

Table 1. Initial Centroid Determination Result Data

Initialization Phase	Centroid Initial Position
0.3033772	0.36715396
0.27038582	0.05920311
0.17923313	0.82152884
0.57535673	0.58374948
0.64780902	0.50116372
0.76679993	0.33897519

This result becomes the foundation for worker bees to continue exploration in the next phase. If the centroid positions are chosen too close together, this may result in suboptimal clustering results, where some clusters may not form perfectly. Based on the initial initialization, the 6 most optimal centroids were obtained, as shown in Fig. 2.

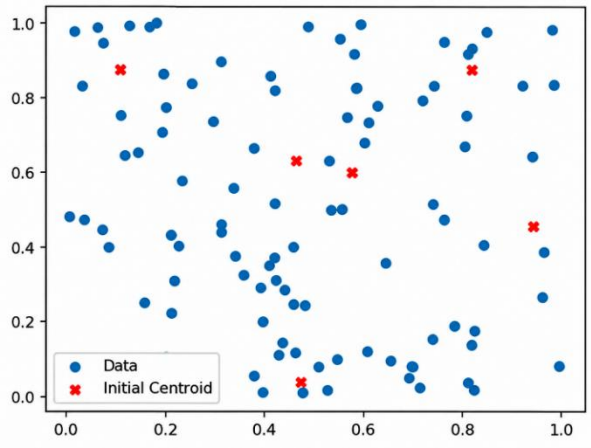


Fig. 2. RPBCO Initial Centroid Distribution Data

2) Employee Bee Phase

In this phase, the employee bee starts to explore around the centroid position that has been initialized before. Each bee will modify the centroid position by adding random perturbations. After the perturbation process is complete, the centroid position will change. Next, the solution quality (fitness) is calculated to evaluate how effective the centroid is in clustering the data, or measured based on the distance between the data and the centroid. The results of the centroid position changes after random perturbation are shown in Table 2.

Table 2. Centroid Change Result Data

Worker Bee	Index Centroid	Centroid X	Centroid Y
0	0	0.3051714336763072	0.3451751958668189
0	1	0.25348620884838796	0.10037808692056131
0	2	0.20872921137552225	0.8522764332232384
0	3	0.5360314987282339	0.5900209734723527
0	4	0.6260235822996165	0.4592594246598012
0	5	0.7202268171692124	0.3412740739692583
1	0	0.26252652794997755	0.3230180034459267
1	1	0.2206808682250842	0.1016934397436383
1	2	0.22175551806380495	0.7985762819834122
1	3	0.6047837290217606	0.5461610976970764
1	4	0.6681335229081145	0.5106034748167592
1	5	0.788790733368038	0.31034185310373913
2	0	0.32873420164681494	0.3550716701648544
2	1	0.27206031986804957	0.08607927049759527
2	2	0.18384991624215685	0.7906987689607101
2	3	0.5964788564634879	0.6135636534378619
2	4	0.6111541783584373	0.5198863506690236
2	5	0.724769311959331	0.3168536710115456
3	0	0.3144352438014049	0.32251180523926526
3	1	0.26404390779128417	0.0955915824789417
3	2	0.2232483686338121	0.808510604545862
3	3	0.537980239160472	0.5636373326707508
3	4	0.6836871270925824	0.5375038711345809
3	5	0.7403708267704789	0.32631656823848665
4	0	0.3475407972676314	0.3372849839051296
4	1	0.3089032670163636	0.10077871356404254

Worker Bee	Index Centroid	Centroid X	Centroid Y
4	2	0.13051753918559744	0.797877457705237
4	3	0.6039557875929508	0.5876638978244557
4	4	0.6517005954596286	0.5194006585953257
4	5	0.7590207345846829	0.28964584709593316
5	0	0.3417785250082269	0.37242633292709776
5	1	0.3071200780120662	0.06843072778829418
5	2	0.15941020391620112	0.8173812894691469
5	3	0.6114838715845834	0.5618087283990651
5	4	0.6299908728209759	0.454389521719836
5	5	0.7590471601098212	0.35016142440360665
6	0	0.34806106716768526	0.35719851085086524
6	1	0.24350998074746338	0.08904607294367979
6	2	0.18859626809133595	0.8703920207055642
6	3	0.5897244513981894	0.5837507607305911
6	4	0.6194792616436416	0.4609243563449872
6	5	0.7941116772217272	0.29717839161286336
7	0	0.29422183726888124	0.38951666995220136
7	1	0.2608418319632879	0.0888150808456701
7	2	0.18140245234554775	0.7762220106992265
7	3	0.599499146446112	0.5572323610096681
7	4	0.6221606971833127	0.5386137722979895
7	5	0.727935580751376	0.29138209236639145
8	0	0.33820797176172485	0.3269092873677554
8	1	0.29084734223803366	0.03382101988170566
8	2	0.17877768043038245	0.8604575718149026
8	3	0.555879201964675	0.6268361007924577
8	4	0.6062297732759256	0.5482943135578087
8	5	0.7623061476209826	0.3879354967944665
9	0	0.3370743220317507	0.39744123420828054
9	1	0.28848799567283717	0.09570152222709459
9	2	0.19077783810156995	0.8114197782792351
9	3	0.6129590580951911	0.5526181474140389
9	4	0.6547022313282861	0.49049047123641304
9	5	0.7939279133928641	0.32035052305425876

Perturbation provides an opportunity to explore solutions around the initial centroid position, which is very useful to avoid the local minimum trap, which is a common problem often encountered in traditional K-Means methods. By using the fitness value to evaluate the quality of the solution, we can determine how good or bad the resulting solution is. A graph showing the centroid changes after modification using BCO can be seen in Fig. 3.

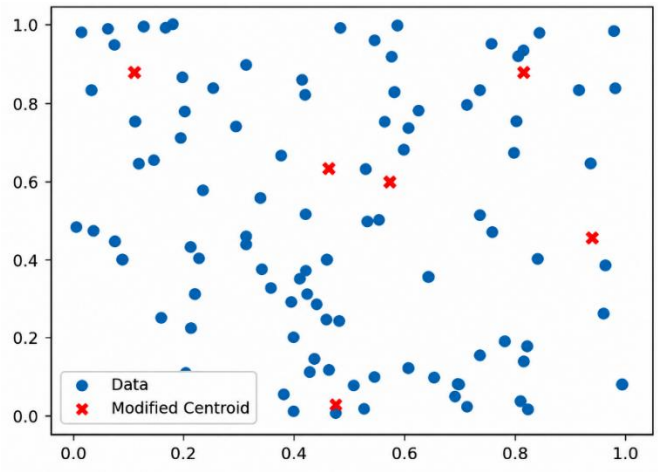


Fig. 3. Modified centroid change data

3) Onlooker Bee Phase

In this phase, the onlooker bee selects the best solution from the results explored by the worker bees. This selection process is done probabilistically, prioritizing the fitness value of each centroid. This approach ensures that superior solutions are prioritized, but less optimal solutions also have a chance to be selected, creating a balance between exploitation (improving existing solutions) and exploration (finding new alternatives). The higher the fitness value of a centroid, the more likely the observer bees are to select it. From this phase, the selected centroid is [0.94730311; 0.45309252], which will be the candidate for the next stage, as shown in Fig. 4.

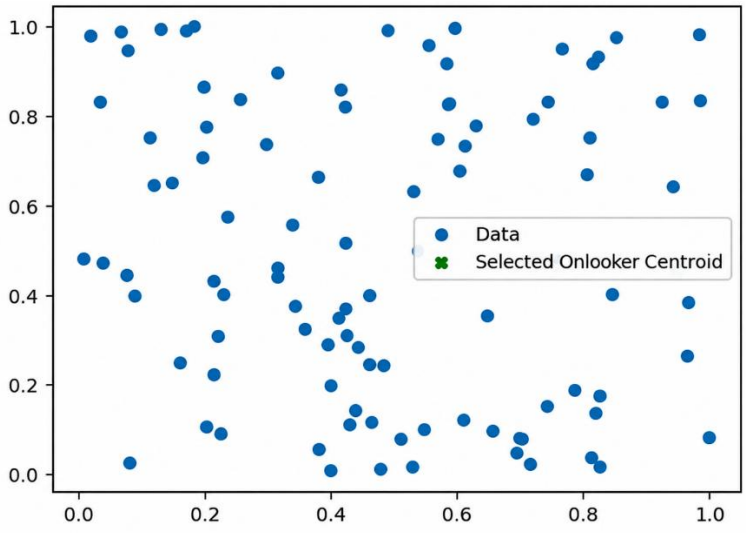


Fig. 4. Selected Best Centroid Data

4) Scout Bee Phase

In the previous phase, there were centroids that did not provide significant improvement, or the existing solution was too stagnant and did not improve in several iterations. Therefore, in this phase, the scout bee will replace the centroid with a new solution randomly. Some centroids that perform poorly will be replaced with new centroids generated randomly. The new centroid value is generated as shown in Fig. 5.

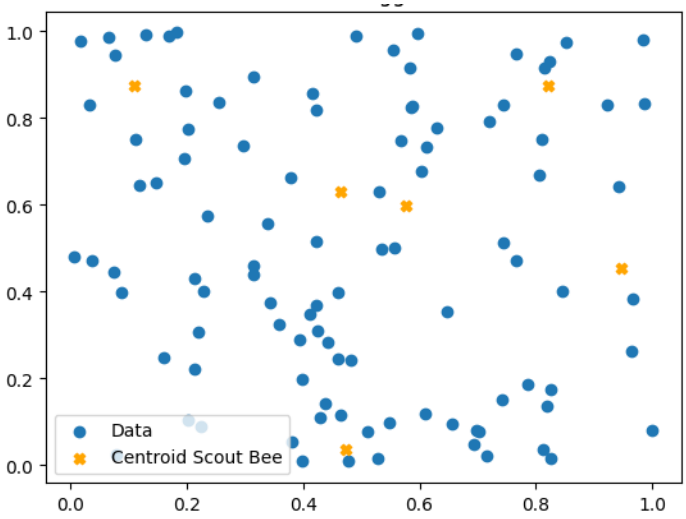


Fig. 5. New Improved Centroid Data

This phase is crucial to avoid getting stuck in a sub-optimal solution, otherwise known as stagnation. Scout bees play an important role in ensuring that the search for solutions keeps moving, not getting stuck in unfavorable local solution spaces, thus providing opportunities to find better solution clusters.

5) Selection of the Best Food Source

In this phase, the best food source is selected, which is the centroid with the highest fitness value from the previous process. This centroid, which is [0.46393086; 0.63166386] as shown in Fig. 6, will be the starting point for the K-Means method in the next stage. By starting from a more optimal centroid, instead of using random initialization, it can increase the chance of K-Means achieving better clustering results in less time.

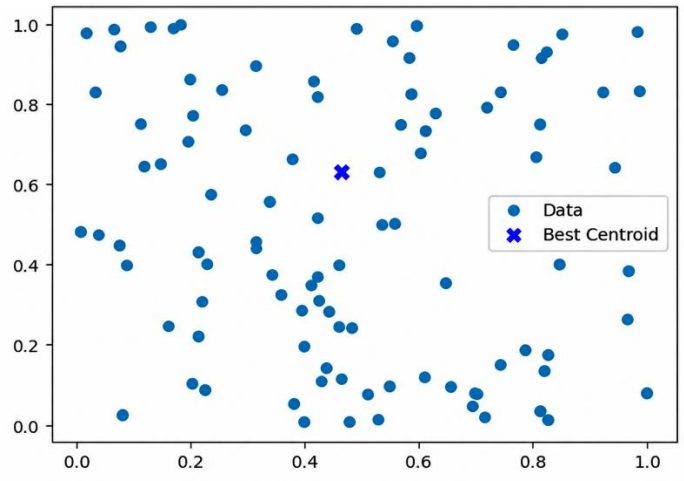


Fig. 6. New best centroid data

6) K-Means Process

Once we have the best food source, the K-Means process begins. The K-Means algorithm will iteratively update the centroid position until the change becomes minimal. The centroid initialization is taken from the RPBCO results, and after several iterations, K-Means successfully groups the MSMEs culinary data into six clusters with stable centroid positions, as shown in Fig. 7.

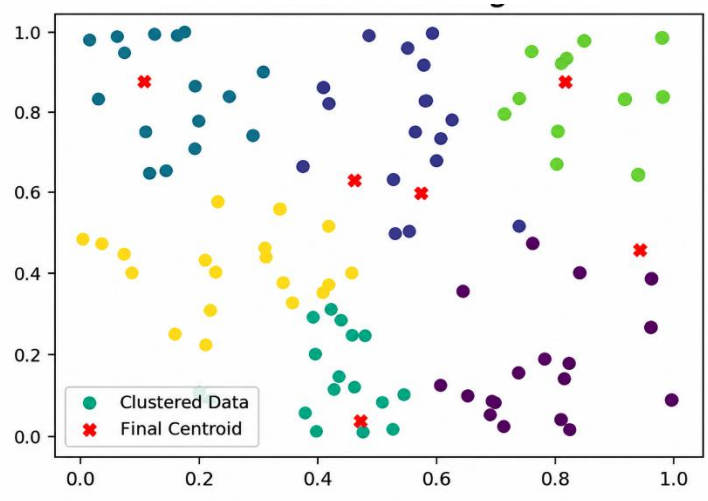


Fig. 7. Cluster data of RPBCO-Kmeans results

The best centroid initialization from RPBCO greatly helps K-Means in finding more accurate and efficient clusters. Usually, K-Means relies on random initialization, which may lead to inconsistent results if the initial centroid is not optimal. However, with the support of RPBCO, the clustering process becomes more stable and accurate.

3.3. Testing Results

In this study, the accuracy is tested using the silhouette coefficient method by trying the number of clusters between 2 and 10. The experimental results show that RPBCO K-Means has higher accuracy at

the number of clusters 6, with the silhouette coefficient value reaching 0.245. Meanwhile, regular K-Means only reaches a value of 0.171. These results confirm previous research, which states that BCO K-Means performs better than conventional K-Means. [4], [26], [27], [28], [29], [30].The results of this test can be seen in Fig. 8.

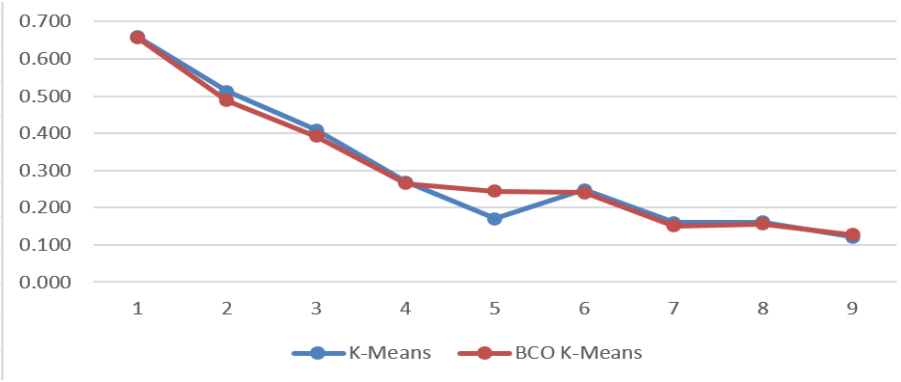


Fig. 8. Graph of K-Means and RPBCO K-Means test results

With this information, it can be concluded that using RPBCO in K-Means gives better results in terms of clustering accuracy, especially when the number of clusters is 6. The accuracy test results were analysed using the Wilcoxon Signed-Rank Test method to compare the accuracy performance between the RPBCO K-Means Algorithm and the K-Means Algorithm. The decision-making process is based on the statistical test results with the following criteria: 1) If the Asymp.Sig value < 0.05, then the alternative hypothesis is accepted, indicating a significant difference; 2) If the Asymp.Sig value > 0.05, then the null hypothesis is accepted, indicating no significant difference. Because the test value shows -1.342, the hypothesis is accepted, so RPBCO can be said to be more effective and better in classifying MSMEs data

3.4. Analysis of MSMEs Data Grouping Result

At this stage, the clustering results obtained using the RPBCO K-Means algorithm were analyzed to evaluate the distribution and characteristics of Micro, Small, and Medium Enterprises (MSMEs) across six clusters. Fig. 9. illustrates the three-dimensional distribution of the clustering results based on three selected features, highlighting the spatial arrangement of data points and variations in cluster membership. The visualization indicates that most data points are concentrated within a relatively narrow feature space, while several observations are more widely dispersed.

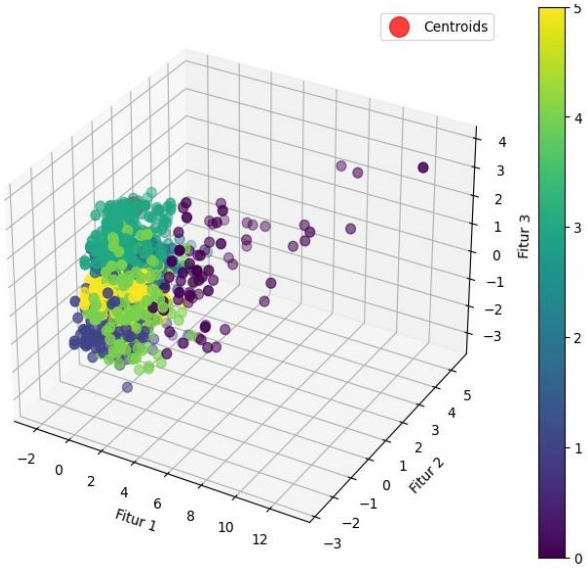


Fig. 9. Distribution of 6 cluster grouping results

Furthermore, the distribution of MSMEs across the six clusters reveals differences in cluster sizes, with Cluster 4 containing the highest number of MSMEs (376 data points), followed by Cluster 5 (302), Cluster 3 (235), Cluster 1 (212), and Cluster 6 (146), whereas Cluster 2 contains the fewest data points (65). These differences indicate an uneven distribution of MSMEs across the identified clusters, reflecting variations in their underlying characteristics. Fig. 10 presents the number of MSMEs assigned to each cluster, providing a quantitative overview of the clustering results. Group 4 recorded the highest number with 376 MSMEs, while Group 2 had the lowest number, which was 65. As shown in Fig. 11, Group 4 is dominated by MSMEs that focus on sales, while Group 5 is more engaged in production. Groups 1, 2, 3, and 6 engage in both production and sales.

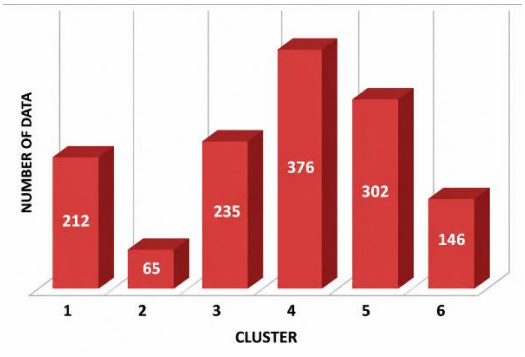


Fig. 10. Graph of the amount of data for each cluster

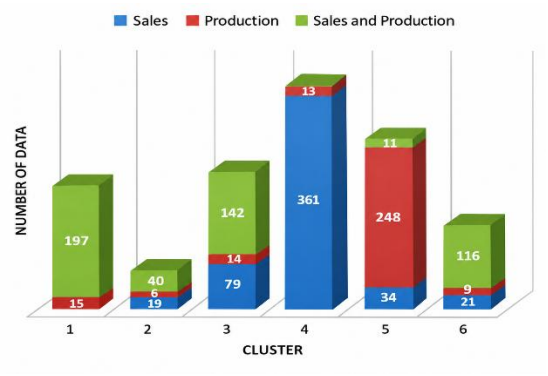


Fig. 11. Graph of MSMEs business activities

Fig. 12 shows that the use of chat media dominates in all clusters, while the use of marketplaces is still limited. This indicates potential that can be further explored. The graph illustrates the distribution of the effectiveness of using marketing media, such as Instagram, Shopee, and WhatsApp, by business cluster. In Cluster 1, Instagram dominates with an effectiveness rate of 80%, while WhatsApp is only 20%, and Shopee has an insignificant contribution. In Cluster 2, Instagram remains the main media with 60% effectiveness, followed by WhatsApp at 50% and Shopee at 30%. In Cluster 3, WhatsApp is the dominant media with 70% effectiveness, followed by Instagram at 40% and Shopee at 30%.

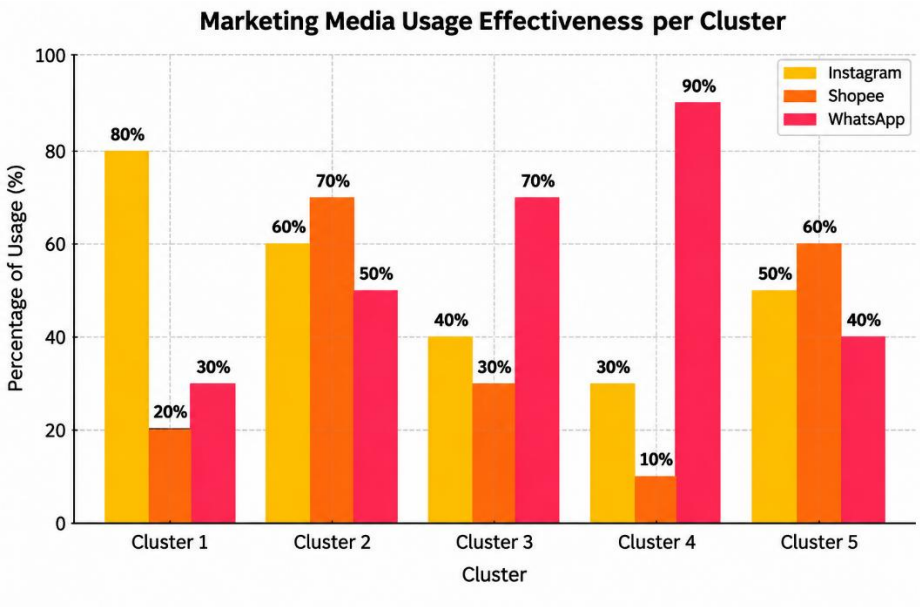


Fig. 12. Graph of electronic media facilities used by MSMEs

In Cluster 4, WhatsApp dominates with 90% effectiveness, while Instagram and Shopee only reach 30% and 10%, respectively. Finally, in Cluster 5, there is a balance between Instagram and Shopee, each having 60% effectiveness, while WhatsApp is used at 40%. This analysis reveals that marketing media preferences vary across clusters, with Instagram and WhatsApp tending to be the top choices, while Shopee is more relevant in certain clusters. These results provide strategic guidance for businesses to choose marketing media that suits the characteristics of each cluster.

Almost all clusters still focus on marketing within the DIY region, with few targeting markets outside Java or even internationally. This is an important note for future MSMEs development, so that marketing can be expanded to a wider area. A graph of marketing objectives can be seen in Fig. 13.

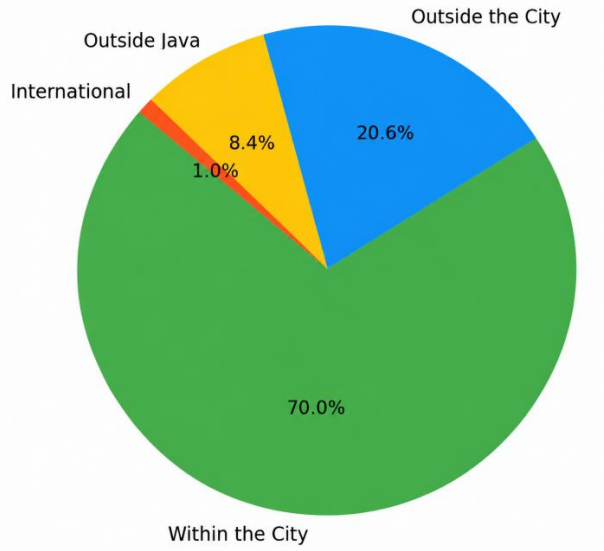


Fig. 13. Marketing objectives of MSMEs

Based on the available data, the distribution of marketing activities across clusters shows that the main focus lies in the 'In-City' category, which accounts for 67.3% of total activities. This indicates that the majority of marketing activities are still oriented towards the local market in the hometown area. The 'Out-of-City' category accounts for 19.8%, indicating that there is a significant effort to expand market coverage to areas outside the city, but still within the same island or province. On the other hand, the 'Outside Java' category only accounts for 8.1%, indicating that marketing to areas outside Java is still relatively limited. Meanwhile, the proportion of marketing activities at the international level is very small, reaching only 1.0%, indicating that efforts to penetrate the global market have not been a top priority. Overall, this data shows that marketing strategies tend to focus on the local market, with some initiatives for regional expansion. However, there is still significant potential to expand beyond Java and expand penetration into international markets in the future

Fig. 14 shows that most clusters have MSMEs with turnover below 10 million in a year. About one-third of the data for each cluster recorded turnover between 10 and 50 million. This suggests that there is still room for growth, and better strategies are needed to increase MSMEs' turnover in the coming years. The graph above illustrates the percentage distribution of turnover by revenue category (< 10 million, 10-50 million, and > 50 million) for each cluster. From the analysis of this graph, it can be inferred that most entities in all clusters have relatively low turnover, with the <10 million revenue category dominating. Cluster 4 recorded the highest proportion at 80%, followed by Cluster 3 with 70%. This finding indicates that most entities in these clusters have relatively small revenues.

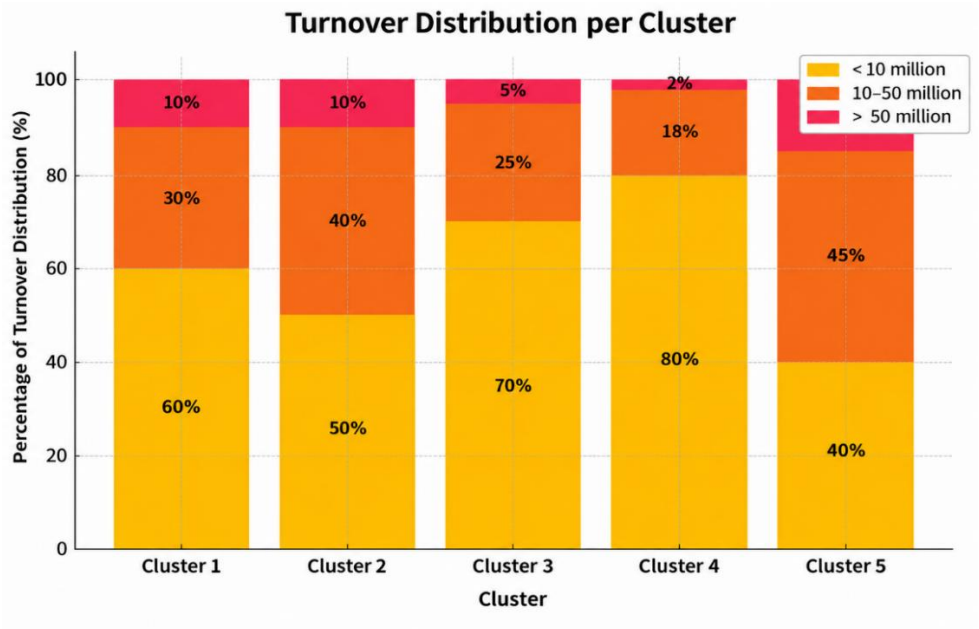


Fig. 14. Graph of turnover obtained by MSMEs in one year

For the 10-50 million turnover category, Cluster 5 showed the highest proportion of 45%, reflecting the dominance of entities with medium income compared to other clusters. Cluster 2 and Cluster 1 recorded proportions of 40% and 30%, respectively, while Cluster 3 and Cluster 4 had lower figures of 25% and 18%. In the >50 million turnover category, Cluster 1 and Cluster 2 recorded the highest proportion, at 10% each. In contrast, Cluster 3, Cluster 4, and Cluster 5 show much smaller contributions, with 5%, 2%, and 5% respectively. This suggests that only a few entities in these clusters managed to achieve high revenues. Overall, the graph reflects that most entities in all clusters have rather low incomes (<10 million), with only a handful of entities achieving high incomes (>50 million). This finding gives an idea of an economic structure that is more oriented towards small to medium-sized entities.

4. Conclusion

The conclusion of this study confirms that the RPBCO K-Means method produces higher accuracy than the traditional K-Means method, with silhouette coefficient values of 0.245 and 0.171 for the 6 clusters analyzed, respectively. This study confirms the effectiveness of RPBCO K-Means in achieving higher clustering accuracy compared to conventional methods, with a 40.9% improvement in the Silhouette Coefficient score. Key findings emphasize the need for MSMEs to adopt tailored marketing strategies based on cluster-specific characteristics. For example, clusters with higher turnover potential should prioritize global e-commerce platforms, while others might focus on regional campaigns. Policymakers should also provide targeted training on digital marketing and offer financial incentives to MSMEs adopting these strategies. By leveraging the insights from this study, MSMEs can improve their market presence and economic resilience, contributing significantly to Indonesia's economic growth. In addition, there is potential for an increase in turnover above 1 million and a shift in marketing media from messaging applications to more effective social media platforms or marketplaces.

Acknowledgment

The authors would like to thank Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat (DRTPM) f or the grant that has been awarded in 2024 with contract number 079/PFR/LPPM-UAD/VI/2024. This research would not have been able to run smoothly without the grant. We would also like to thank our colleagues for their assistance in this research process.

Declarations

Author contribution. Conceptualization: L. Zahrotun; Methodology: L. Zahrotun, I. Arfiani; Software: L. Zahrotun, I. Arfiani; Resources: L. Zahrotun; Writing original draft preparation: L. Zahrotun, I. Arfiani; Translating and editing: D. Normawati. Writing review and editing: All authors.
Funding statement. This research was funded by Direktorat Riset, Teknologi, dan Pengabdian kepada Masyarakat (DRTPM) with contract number 079/PFR/LPPM-UAD/VI/2024 in 2024
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

Data Availability

The dataset is available at: https://huggingface.co/datasets/ikaarfiani/UMKM_Kuliner/tree/main.

References

[1] B. Kusumasari and N. D. Retnandari, "When pandemic knocks, adapt fast: strategy and resilient of MSMEs facing crisis in Indonesia," *Int. J. Glob. Small Bus.*, vol. 12, no. 4, p. 392, 2021, doi: 10.1504/IJGSB.2021.120488.

[2] I. Mahfud, "The role of MSMEs (Micro, Small and Medium Enterprises) in improving the national economy," *J. Econ. Bus. Lett.*, vol. 5, no. 5, pp. 13–23, Oct. 2025, doi: 10.55942/jebl.v5i5.915.

[3] A. S. Wahyudi, I. Yulivan, and A. Rahman, "The Role of Micro, Small, and Medium Enterprises (MSMEs) in Supporting Indonesia's Economic Resilience," *J. Pertahanan Media Inf. tentang Kaji. dan Strateg. Pertahanan yang Mengedepankan Identity, Nasionalism dan Integr.*, vol. 10, no. 2, pp. 297–307, Aug. 2024, doi: 10.33172/jp.v10i2.19519.

[4] R. E. Caraka *et al.*, "Micro, Small, and Medium Enterprises' Business Vulnerability Cluster in Indonesia: An Analysis Using Optimized Fuzzy Geodemographic Clustering," *Sustainability*, vol. 13, no. 14, p. 7807, Jul. 2021, doi: 10.3390/su13147807.

[5] P. Kumar, A. Chandra, S. Shivani, and V. Pillai, "Evaluating the impact of digital and financial literacy on MSME performance," *J. Int. Counc. Small Bus.*, vol. 6, no. 2, pp. 284–296, Apr. 2025, doi: 10.1080/26437015.2024.2405548.

[6] A. sanatan satpathy, S. kumar Sahoo, A. Mohanty, and P. P. Mohanty, "Strategies for enhancements of MSME resilience and sustainability in the post-COVID-19 era," *Soc. Sci. Humanit. Open*, vol. 11, p. 101223, 2025, doi: 10.1016/j.ssaho.2024.101223.

[7] S. Vasani and A. M. Abdulkareem, "MSME market presence and competitiveness in a global economy," *Cogent Econ. Financ.*, vol. 12, no. 1, p. 2416992, Dec. 2024, doi: 10.1080/23322039.2024.2416992.

[8] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-Means and Alternative Clustering Methods in Modern Power Systems," *IEEE Access*, vol. 11, pp. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.

[9] E. Laber and L. Murtinho, "Optimization of Inter-group criteria for clustering with minimum size constraints," in *Advances in Neural Information Processing Systems 36*, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp. 10339–10358. doi: 10.52202/075280-0454.

[10] L.-Z. Tan, Y.-Y. Tan, G.-X. Yun, and C. Zhang, "An improved genetic algorithm based on k-means clustering for solving traveling salesman problem," in *Computer Science, Technology and Application*, WORLD SCIENTIFIC, Dec. 2016, pp. 334–343. doi: 10.1142/9789813200449_0042.

[11] W. Aristika and W. J. Hartono, "Application of K-Means Clustering to Determine the Influence of Facebook Social Media on Micro, Small and Medium Enterprises (MSMEs) in Pekanbaru City District," *J. Ilmu Komput. dan Bisnis*, vol. 11, no. 1, pp. 2389–2395, Oct. 2020, doi: 10.47927/jikb.v11i1.202.

[12] B. J. D. Sitompul, "Improvement of Davies-Bouldin Index Clustering Evaluation Results by Determining the Initial Cluster Center Point of the K-Means Algorithm," Universitas Sumatra Utara, 2018. [Online]. Available: http://repositori.usu.ac.id/handle/123456789/3827

[13] I. Arfiani, H. Yuliansyah, and M. D. Suratin, "Implementation of Bee Colony Optimization in Selecting Centroids (Central Clusters) in the K-Means Algorithm," *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 756–763, Mar. 2022, doi: 10.47065/bits.v3i4.1446.

[14] M. Čubranić-Dobrodolac, L. Švadlenka, S. Čičević, A. Trifunović, and M. Dobrodolac, "A bee colony optimization (BCO) and type-2 fuzzy approach to measuring the impact of speed perception on motor vehicle crash involvement," *Soft Comput.*, vol. 26, no. 9, pp. 4463–4486, May 2022, doi: 10.1007/s00500-021-06516-4.

[15] G. Yao, Y. Wu, X. Huang, Q. Ma, and J. Du, "Clustering of Typical Wind Power Scenarios Based on K-Means Clustering Algorithm and Improved Artificial Bee Colony Algorithm," *IEEE Access*, vol. 10, pp. 98752–98760, 2022, doi: 10.1109/ACCESS.2022.3203695.

[16] W. Xiao, G. Li, C. Liu, and L. Tan, "A novel chaotic and neighborhood search-based artificial bee colony algorithm for solving optimization problems," *Sci. Rep.*, vol. 13, no. 1, p. 20496, Nov. 2023, doi: 10.1038/s41598-023-44770-8.

[17] Satish S. Banait, "Enhancing Big Data Clustering: The Improved K-Means - Artificial Bee Colony Algorithm with MapReduce," *Panam. Math. J.*, vol. 33, no. 3, pp. 125–134, Jul. 2024, doi: 10.52783/pmj.v33.i3.885.

[18] Y. Pratama, E. Sulistianingsih, N. N. Debataraja, and N. Imro'ah, "K-Means Clustering dan Mean Variance Efficient Portfolio dalam Portofolio Saham," *Jambura J. Probab. Stat.*, vol. 5, no. 1, pp. 24–30, Jun. 2024, doi: 10.37905/jjps.v5i1.20298.

[19] K. Sirait, Tulus, and E. B. Nababan, "K-Means Algorithm Performance Analysis With Determining The Value Of Starting Centroid With Random And KD-Tree Method," *J. Phys. Conf. Ser.*, vol. 930, p. 012016, Dec. 2017, doi: 10.1088/1742-6596/930/1/012016.

[20] S. Poulsen *et al.*, "Psychometric Evaluation of the Cybersecurity Concept Inventory," *ACM Trans. Comput. Educ.*, vol. 22, no. 1, pp. 1–18, Mar. 2022, doi: 10.1145/3451346.

[21] D. Sun, H. Fei, and Q. Li, "A Bisecting K-Medoids clustering Algorithm Based on Cloud Model," *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 308–315, 2018, doi: 10.1016/j.ifacol.2018.08.301.

[22] Martanto, S. Anwar, C. L. Rohmat, F. M. Basysyar, and Y. A. Wijaya, "Clustering of internet network usage using the K-Medoid method," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012036, Feb. 2021, doi: 10.1088/1757-899X/1088/1/012036.

[23] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bull.*, vol. 1, no. 6, p. 80, Dec. 1945, doi: 10.2307/3001968.

[24] V. Romanuke, "Random centroid initialization for improving centroid-based clustering," *Decis. Mak. Appl. Manag. Eng.*, vol. 6, no. 2, pp. 734–746, Dec. 2023, doi: 10.31181/dmame622023742.

[25] Tarisya Qurrota A'yuni, Baiq Nina Febriati, Lazuardy Ilham Effendie, Muhammad Muhajir, and R. Yotenka, "MSME Sales Clustering Based on Business Aid Distribution Priority Using K-Affinity Propagation," *Enthusiastic Int. J. Appl. Stat. Data Sci.*, pp. 111–124, Apr. 2023, doi: 10.20885/enthusiastic.vol3.iss1.art10.

[26] M. Nikolić, J. Rakas, and D. Teodorović, "Formulation of the airport collaborative gate allocation problem and the Bee Colony Optimization solution approach," *Eng. Appl. Artif. Intell.*, vol. 128, p. 107433, Feb. 2024, doi: 10.1016/j.engappai.2023.107433.

[27] C. Borgelt and O. Yarikova, "Initializing k-means Clustering," in *Proceedings of the 9th International Conference on Data Science, Technology and Applications*, SCITEPRESS - Science and Technology Publications, 2020, pp. 260–267. doi: 10.5220/0009872702600267.

[28] L. P. K, G. Surya Narayana, M. D. Ansari, and V. K. Gunjan, "Instantaneous Approach for Evaluating the Initial Centers in the Agricultural Databases Using K-Means Clustering Algorithm," *J. Mob. Multimed.*, pp. 1–18, Aug. 2021, doi: 10.13052/jmm1550-4646.1813.

[29] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots," *Appl. Sci.*, vol. 11, no. 23, p. 11202, Nov. 2021, doi: 10.3390/app112311202.

[30] X. Pan, Y. Wang, Y. Lu, and N. Sun, "Improved artificial bee colony algorithm based on two-dimensional queue structure for complex optimization problems," *Alexandria Eng. J.*, vol. 86, pp. 669–679, Jan. 2024, doi: 10.1016/j.aej.2023.12.011.