Scientific Journal of Informatics
Vol. 13, No. 1, Feb 2026
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

K-MEANS WITH PARTICLE SWARM OPTIMIZATION FOR ERROR REDUCTION IN MICRO, SMALL, MEDIUM ENTERPISE CRAFT IN YOGYAKARTA

Athallah Naufal Muthahhar[1*], Lisna Zahrotun[2]
[1,2]Informatics Department, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Indonesia

Abstract.
Purpose: This study aims to optimize the determination of the optimal number of clusters in the segmentation of handicraft-based Micro, Small, and Medium Enterprises (MSMEs) in Yogyakarta to support targeted and data-driven development strategies.
Approach: A quantitative approach was applied to survey data collected from 145 MSMEs. The analytical pipeline consisted of four stages: (1) data acquisition through structured surveys, (2) preprocessing including encoding, mode imputation for missing values, and Min–Max normalization, (3) model development using the K-Means algorithm integrated with Particle Swarm Optimization (PSO) to automatically search for the optimal cluster number (K = 2–10), and (4) performance evaluation using Silhouette Score, Sum of Squared Error (SSE), and Mean Absolute Error (MAE).
Result: The optimization process consistently converged to an optimal configuration of K = 8 clusters. Compared to standard K-Means, the proposed K-Means + PSO model reduced SSE from 54.555 to 51.676 and MAE from 0.124 to 0.116, indicating improved clustering stability and compactness. Semantic centroid analysis further revealed a hierarchical MSME structure consisting of Established Digital Adopters, Developing Potential Enterprises, and Subsistence Micro Enterprises, highlighting disparities in digital maturity and market reach.
Novelty: This study contributes by integrating swarm-based optimization with centroid-driven semantic profiling, bridging algorithmic enhancement and policy-relevant interpretation. The proposed framework provides a robust and interpretable clustering model for MSME segmentation in emerging economic contexts.

Keywords: Clustering, MSME craft, K-Means algorithm, PSO optimization
Received January 2026 / **Revised** February 2026 / **Accepted** February 2026

INTRODUCTION

Micro, Small, and Medium Enterprises (MSMEs) play a vital role in national economic growth, particularly in the handicraft sector of the Special Region of Yogyakarta. Although this region is renowned for its diverse creative industries, many MSMEs still face fundamental obstacles, ranging from limited access to capital and low utilization of digital marketing and production technology to a lack of business training [1]. International studies indicate that MSMEs are highly heterogeneous in terms of scale, resources, and operational capabilities, making uniform development policies ineffective. This complexity requires development strategies that cannot be applied uniformly. Therefore, a technology-based approach such as Data Mining is required to cluster MSMEs based on their business characteristics so that empowerment strategies can be targeted effectively [2]. Data-driven segmentation has been widely recognized as an effective decision-support tool for policymakers in formulating evidence-based MSME development programs.

Data Mining has become a reliable method for extracting knowledge and recognizing patterns from large datasets to produce meaningful information [3]. Generally, there are two main approaches in data mining: predictive and descriptive models [4]. Unlike predictive models aimed at forecasting specific attribute values, descriptive models focus on discovering hidden patterns within data to generate new insights [4], [5]. Descriptive models are particularly suitable for socioeconomic and business-related datasets, where the primary objective is to understand structural patterns rather than prediction. One of the primary techniques in descriptive models is clustering, which is used to group data, in this context, MSMEs based on shared characteristics [6]. Through clustering, MSMEs with similar business profiles can be systematically identified, enabling more precise policy formulation and

*Corresponding author.
Email addresses: 2200018015@webmail.uad.ac.id (Muthahhari), lisna.zahrotun@tif.uad.ac.id (Zahrotun)
DOI: 10.15294/sji.v12i1.40664

targeted resource allocation.

Among various clustering algorithms, K-Means is the most widely adopted due to its computational simplicity and efficiency in handling numerical data [7]. However, this method has significant limitations, particularly its high dependence on the initialization of centroids and the number of clusters (K), which must be determined manually before the process begins. Manual or random determination of K often causes the algorithm to converge prematurely into local optima, resulting in inaccurate and unstable clustering [8]. Several empirical studies have shown that inappropriate selection of K can significantly reduce clustering quality and analytical reliability [9]. To mitigate these conventional limitations, this study proposes the application of Particle Swarm Optimization (PSO) to automate the search for the most optimal number of clusters (K), thereby eliminating the subjectivity associated with manual selection [10].

Although several studies have explored the integration of K-Means with heuristic optimization techniques such as PSO, most of these works focus on benchmark datasets or generic business domains rather than MSME handicraft sectors characterized by heterogeneous data [11]. In addition, the optimization of the number of clusters as a primary objective for clustering error reduction in regional MSME contexts remains underexplored.

This study follows the Knowledge Discovery in Databases (KDD) stages [12], encompassing data selection, preprocessing, transformation, data mining, and evaluation. At the introductory level, this study emphasizes a systematic and reproducible analytical framework rather than detailed technical implementation. The study utilizes a secondary dataset comprising 145 handicraft MSMEs in Yogyakarta with attributes ranging from numerical to categorical and ordinal. The presence of mixed-type attributes in MSME data requires appropriate preprocessing techniques to ensure compatibility with clustering algorithms. To ensure clustering accuracy, all categorical attributes were encoded into numerical forms to suit the algorithm's characteristics [6]. Particle Swarm Optimization is employed to determine the optimal number of clusters based on clustering validity indices, including the Silhouette Score and Sum of Squared Error (SSE) [3], [13].

Previous research has primarily focused on the application of standard K-Means without initial parameter optimization, and studies addressing heuristic optimization for MSME segmentation with mixed data characteristics remain limited. Furthermore, research that explicitly evaluates clustering performance using multiple error metrics in MSME handicraft datasets is still scarce. Therefore, this research intends to group MSME characteristics using the K-Means algorithm optimized with PSO [14], [15]. The main contribution of this study is the application of PSO to optimize the number of clusters in K-Means for MSME handicraft data, resulting in lower clustering error and improved model validity. The novelty of this study lies in the application of PSO to optimize the number of clusters in K-Means, resulting in lower error metrics (SSE and MAE) and improved clustering validity compared to conventional approaches [16], [17], [18].

METHODS

In achieving the research objectives, the activities during the research are illustrated in Figure 1.

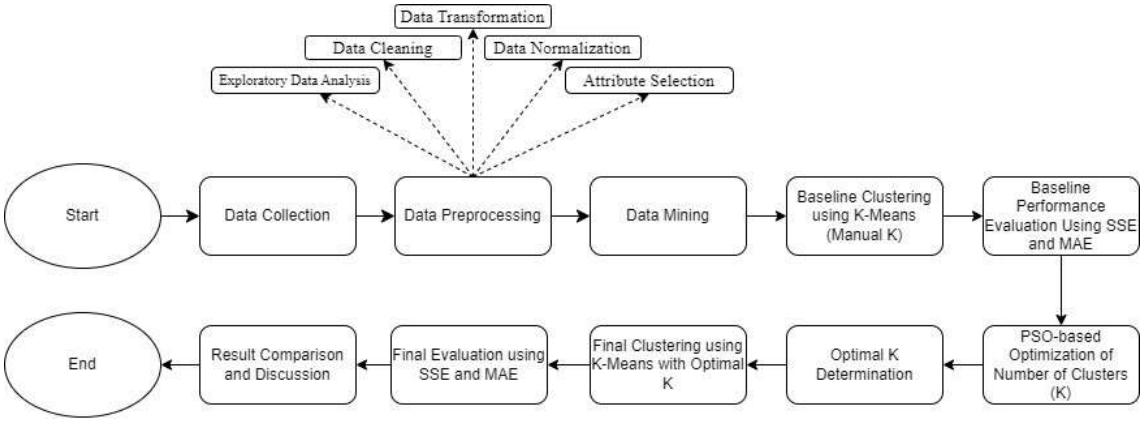

Figure 1. Research framework

Data Collection
The data used in this study was obtained secondarily, originating from a dataset that had been collected and provided by another party, in the form of data on 145 craft MSMEs in the Yogyakarta region. This dataset contains various important attributes that describe business profiles, such as the age of the business owner, education, location, type of product, export status, land ownership, business capital, annual turnover, and use of digital tools. The data was obtained in .xlsx format and then further processed by the researcher to suit the purposes of the analysis.

Data Preprocessing
Data processing is an important stage in the Data Mining process, which is an analysis method that aims to find patterns or hidden information from data [2], [19]. In this study, data processing was carried out through a series of pre-processing stages to ensure that the data was ready to be used in the analysis and grouping process [20]. The stages carried out were as follows:

1) Exploratory Data Analysis: The first process carried out to understand the characteristics and content of the data in depth. This stage includes evaluating the data by checking for data duplication, missing values, or inconsistent data [20], [21].
2) Data Cleaning: The process of correcting data based on preliminary checks to determine whether there is duplicate data or missing values, either stored as NaN or the symbol "-" [22].
3) Data Transformation: The process of converting data into numerical form to suit the K-means and PSO algorithms [22]. The transformation begins with the encoding process of categorical attributes such as Annual Turnover, Business Sector, and Land Ownership Status, so that they can be represented in numerical form. Some attributes use the Label Encoding method, while ordinal attributes such as Annual Turnover are converted manually based on value levels.
4) Data normalization: The process of converting the value of each attribute into a standard distribution [19], [23]. In this study, the Min Max normalization method is applied to transform all attribute values into the range of 0 to 1 [19], [24]. This step is crucial to ensure that each feature contributes proportionally to the clustering process and prevents attributes with larger numerical ranges from dominating the results [23].
5) Attribute Selection: This process is carried out to identify attributes that are relevant to the clustering process [4], [25]. The results obtained were that 11 of the 31 attributes were selected, namely the age of the business owner, the average age of workers, the number of male workers, the number of female workers, annual turnover, health insurance ownership, whether the product is an export commodity, business sector, marketing objectives, use of electronic media, and ownership status of land or business premises.

Data Mining
Data Mining represents the core stage within the Knowledge Discovery in Databases (KDD) framework, which aims to extract meaningful patterns, hidden structures, and valuable knowledge from large datasets [1], [12], [16]. After completing the data preprocessing stage where raw data are cleaned, transformed, normalized, and relevant attributes are selected the processed dataset is ready to be analyzed using clustering techniques [26], [27].

In this study, the data mining process is implemented through an unsupervised learning approach using clustering methods [10], [28], [29]. Clustering is applied to group MSMEs based on similarities across multiple attributes, including the age of the business owner, workforce composition, annual turnover, business sector, export status, use of digital media, and land ownership [6], [26], [30]. The objective of this stage is to identify homogeneous groups of MSMEs such that entities within the same cluster share similar characteristics, while distinct differences are maintained between clusters [29].

To achieve this objective, two clustering schemes are employed standard K-Means clustering and K-Means clustering optimized using Particle Swarm Optimization (PSO) [16]. The standard K-Means method is used as a baseline model, while the PSO-based approach is proposed to enhance clustering performance by determining the optimal number of clusters [31].

Baseline Clustering with K-Means

As illustrated in the methodological framework in Figure 1, the clustering stage constitutes the core model construction process in this study. The initial clustering approach employs the K-Means algorithm, a partition-based unsupervised learning method introduced by MacQueen (1967), which is widely used for grouping numerical data based on similarity [32]. In this research, K-Means serves as the baseline clustering model prior to optimization.

The dataset used for clustering consists of 145 MSME instances represented as multivariate numerical feature vectors obtained after preprocessing [3]. Since no class labels are provided, the clustering process aims to partition the dataset into *k* clusters such that MSMEs within the same cluster exhibit high similarity, while MSMEs in different clusters are well separated.

K-Means clustering relies on distance-based similarity measurement, where each data point is assigned to the nearest cluster centroid. In this study, Euclidean distance is employed as the similarity metric due to its effectiveness for numerical data [13]. The Euclidean distance between a data point and a centroid is calculated using Equation 1 [33]:

$$D_{(x_1,x_2)} = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (1)$$

Where $D(x_1, x_2)$ represents the Euclidean Distance, x_{1i} and x_{2i} are the values of the first and second attributes at index i respectively, and *n* denotes the total number of data attributes.

The clustering process in K-Means is performed iteratively, starting with the random selection of initial centroids [34]. Each data point is then assigned to the cluster with the nearest centroid based on Euclidean distance. After all data is divided into clusters, the centroid is updated by calculating the average position of all cluster members [35]. The centroid point update formula is shown in equation 2 [29].

$$\mu_j - \frac{1}{|C_j|} \sum_{x_1 \in C_j} x_i \quad (2)$$

Where μ_j represents the new centroid for cluster j, C_j denotes the set of data points belonging to cluster j, x_i refers to the i-th data point within cluster j, and $|C_j|$ indicates the total number of data points in that cluster.

The algorithm terminates once the convergence criteria are met or the maximum iteration limit is reached. However, to enhance clustering performance, K-Means is often hybridized with optimization algorithms such as Particle Swarm Optimization (PSO) [36], [37], [38].

Clustering with K-Means + Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is a population-based optimization algorithm developed by Kennedy and Eberhart in 1995, inspired by the social behavior of flocks of birds and schools of fish [21], [39]. PSO is a metaheuristic method capable of finding optimal solutions through the movement of particles in a search space [16].

In this study, PSO was implemented to automatically determine the most optimal number of clusters (K), overcoming the weakness of manual determination of K in K-Means [16]. The process can be seen in Figure 2 [7].

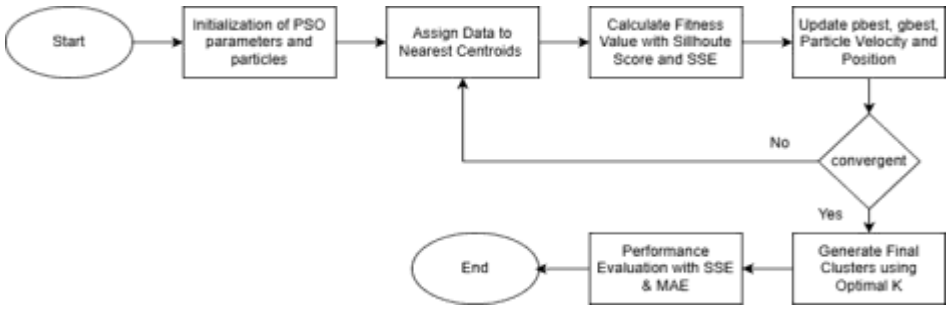

Figure 2. K-Means Optimization with PSO

1) Initialization of PSO Parameters and Particles: Start by initializing PSO algorithm parameters such as acceleration coefficient, number of particles, inertia factor, and maximum number of iterations. In this stage, the specified PSO particles will be calculated randomly, where each particle represents a candidate K value (number of clusters) to be tested.
2) Assign Data to Nearest Centroids: In each iteration, the dataset is assigned to the nearest centroids defined by the current position of each PSO particle. This step determines the cluster membership for every data point based on the particle's configuration.
3) Calculate Fitness Value (Silhouette Score and SSE): Once the data assignment is complete, the quality of the clustering is evaluated using both the Silhouette Score and Sum of Squared Error (SSE). These metrics serve as the fitness function. The Silhouette Score is calculated using equation (3) to measure cluster cohesion and separation [40].

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{3}$$

Where $a(i)$ denotes the cohesion and $b(i)$ represents the separation. A higher Silhouette Score combined with a minimized SSE value indicates distinct separation and high compactness of the clusters, reflecting a better solution quality used to guide the particle updates.

4) Update Pbest, Gbest, Velocity, and Position: Based on the calculated fitness values, the algorithm updates the Personal Best (Pbest) for each particle and the Global Best (Gbest) for the entire swarm. Subsequently, the velocity and position of each particle are updated to explore the search space for the optimal solution in the next iteration.
5) Convergence Check: The algorithm checks whether the termination criteria are met, such as reaching the maximum number of iterations or achieving convergence (where the global best value stabilizes). If the criteria are not met, the process loops back to the data assignment step; otherwise, it proceeds to the final stage.
6) Generate Final Clusters using Optimal K: Upon reaching convergence, the algorithm utilizes the Global Best (Gbest) solution representing the optimal K and centroid positions to perform the final clustering process and generate the definitive cluster structure [15].
7) Performance Evaluation: Finally, the resulting clusters are rigorously evaluated to validate the model's performance using Sum of Squared Error (SSE) and Mean Absolute Error (MAE), ensuring the accuracy and validity of the segmentation.

The PSO calculation used is shown in equation 4 [24].

$$V_i(t) = WV_i + c_1 R_1 \big(X_{pi} - X_i\big) + c_2 R_2 \big(X_{gi} - X_i\big) \tag{4}$$

Where R_1 and R_2 are random values in the range [0, 1], while c_1 and c_2 represent the acceleration coefficients. X_{pi} denotes the personal best position of the i-th particle, and X_{gi} refers to the global best position found by the entire swarm. Furthermore, W is the inertia weight, and X_i indicates the current position of the i-th particle.

PSO standardization employs specific safe ranges for parameter selection to ensure algorithm stability, typically with inertia weights ranging from 0.4 to 1.4 and acceleration coefficients between 1.5 and 2.0 [41]. After determining the particle velocity, the particle position displacement is recalculated using the formula in equation 5.

$$Xi\,(t) \;=\; Xt \;+\; Vt \tag{5}$$

Where X_t denotes the current position of the particle (old position), and V_t represents the newly updated velocity of the particle.

The iterative process continues until the termination criteria are met. The optimal number of clusters (K) is determined by maximizing the Silhouette Score, while the Sum of Squared Error (SSE) is used as a supporting metric to ensure cluster compactness. The algorithm stops when the maximum number of iterations is reached or when convergence is achieved.

Result Evaluation and Comparison

Upon completion of the clustering process, the final stage involves a comparative evaluation between the standard K-Means and the PSO-optimized K-Means using Sum of Squared Error (SSE) and Mean Absolute Error (MAE). SSE quantifies the total squared deviation between each data point and its nearest centroid; a lower value indicates superior cluster compactness [17], [27]. Meanwhile, MAE measures the average absolute deviation between data points and their respective centroids. Lower MAE values signify more accurate and consistent clustering results [32]. The mathematical formulations for these metrics are defined as follows:

1) Sum of Square Error calculation equation 6 [42].

$$SSE = \sum_{k=1}^{K} \sum_{x_i \epsilon S_k} \|x_i - ck\|^2 \tag{6}$$

Where K represents the total number of clusters, and S_k denotes the set of data points belonging to the k- th cluster. Furthermore, x_i refers to the i-th data point within that cluster, C_k is the centroid of the k-th cluster, and $\|x_i - C_k\|^2$ represents the squared Euclidean distance between the data point x_i and its respective centroid.

2) Mean Absolute Error calculation equation 7 [32].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - C_{k(i)}\right|$$

Where n denotes the total number of data points in the dataset. xi represents the i-th data point, while $Ck(i)$ refers to the centroid of the k-th cluster to which the data point belongs. Finally, $|xi - Ck(i)|$ indicates the absolute difference between the data point and its corresponding centroid

RESULT AND DISCUSSION

Data Preprocessing Results

The experimental analysis begins with data preprocessing, which plays a critical role in ensuring the robustness of distance-based clustering. The dataset consists of 145 Micro, Small, and Medium Enterprises (MSMEs) collected from Yogyakarta, encompassing demographic, operational, and digital adoption attributes.

Initial exploratory analysis revealed that the dataset was not directly suitable for clustering due to missing values and heterogeneous feature scales. Missing values were primarily observed in Average Worker Age (29.66%) and Health Insurance Ownership (24.83%), represented by the placeholder symbol "–". These missing entries were treated as null values.

To preserve the dataset size and avoid sampling bias, mode imputation was employed. This method is well- suited for categorical and discrete MSME attributes and has been widely adopted in prior MSME clustering studies to maintain data distribution integrity. Compared to deletion-based approaches, mode imputation ensures that micro-scale enterprises often underrepresented remain included in the analysis.

Feature selection was subsequently conducted based on domain relevance and data completeness. A total of 11 attributes were retained, representing business ownership characteristics, workforce composition, digital adoption, and market orientation. This selection balances dimensional adequacy and interpretability, reducing noise without oversimplifying MSME heterogeneity.

Since K-Means relies on Euclidean distance, categorical attributes were transformed using label encoding, followed by Min Max normalization to scale all features into a uniform range [0,1]. This step prevents dominance of high-magnitude attributes such as annual turnover. A sample of the normalized dataset is presented in Table 1, which serves as the final input for the clustering stage.

Table 1. Normalization data

	Owner Age	Avg Worker Age	Male Workers	Female Workers	Annual Turnover	Health Insurance	Export Status	Business Sector	Marketing Obj	Digital Media	Land/Building Status
0	0.48	0.75	0.1	0.071429	0.111111	0.666667	0	0.636364	0	0.333333	0.666667
1	0.76	1	0.5	0.142857	0.666667	0.666667	0	0	0.421053	0	0.666667
2	0.52	0.5	0.1	0	0.222222	0	0	0.909091	0	0.272727	0.666667
3	0.36	0.5	0.1	0.071429	0	0.666667	0	0.545455	0.263158	0.242424	0.666667
4	0.32	0.25	0.1	0.071429	1	0.666667	0	0.909091	0.421053	0.484848	0
…	...	...	...	...	...	...	...	...	...	...	...
140	0.68	0	0	0	0	0	0	0.636364	0	0.212121	1
141	0.52	0	0.1	0.071429	0.111111	0.666667	0	0.909091	0.315789	0.242424	1
142	0	0	0	0	0	0.666667	0	0.909091	0.947368	0.757576	0.666667
143	0.46	0	0	0	0.222222	0.666667	0	0.636364	0	0.21212	0.666667
144	0.98	0.75	0.2	0.1428557	0.222222	0.666667	0	0.909091	0.789474	0.090909	0.666667

Determination of Optimal Clusters using Particle Swarm Optimization

To overcome the sensitivity of K-Means to predefined cluster numbers, this study integrates Particle Swarm Optimization (PSO) to search for the optimal number of clusters (K). The optimization workflow follows the structure illustrated in Figure 2 and consists of the following steps:

1) Initialization of PSO parameters and particles: The initial stage of the algorithm involves establishing key parameters that govern the swarm's behavior in searching for the optimal solution. The specific PSO configuration adopted in this study is detailed in Table 2.

Table 2. Configuration of PSO parameters

Parameter	Value
Swarm size	50
Max Iterations	30
Inertia Weight (w)	1.0
Cognitive Coefficient (c1)	1.75
Social Coefficient (c2)	1.75
Cluster Range (K)	2 - 10

PSO parameters were initialized according to established theoretical constraints to balance exploration and exploitation. As shown in Table 2, the swarm size was set to 50 particles with a maximum of 30 iterations. The cognitive and social coefficients (c1 = c2 = 1.75) fall within the recommended stability range reported in prior optimization literature, while an inertia weight (w) of 1.0 maintains consistent swarm momentum [41]. The search space for candidate cluster numbers was defined as K = 2 to 10, aligning with typical MSME segmentation practices in the literature.

2) Assign Data to Nearest Centroids and Calculate Fitness Value: In this stage, each particle performs data assignment to the nearest centroids to form temporary cluster structures. Since the particles are still in the initialization phase (Iteration 0), the data grouping is executed using a random number of clusters (K).

Table 3. Results of Data Assignment and Fitness Calculation

	Particle	Candidate K	Silhouette Score	SSE	MAE
0	1	8	0.193	51.676	0.116
0	2	7	0.180	55.835	0.125
0	3	7	0.180	55.835	0.125
…	…	…	…	…	…
30	48	9	0.188	49.166	0.113
30	49	7	0.180	55.835	0.125
30	50	8	0.193	51.676	0.116

At each iteration, particles assign MSME data points to centroids based on the candidate K value and compute fitness scores using Silhouette Score, SSE, and MAE. Early-stage results (Iteration 0), summarized in Table 3, indicate that K = 8 consistently yields a higher Silhouette Score (0.193) compared to nearby candidates such as K = 7, which exhibit higher error values.As the algorithm progresses, a trade-off emerges between compactness (SSE minimization) and separation quality (Silhouette Score). While K = 9 achieves the lowest SSE (49.166), K = 8 maintains superior cluster separation, suggesting a more balanced structure.

3) Update Pbest, Gbest, Velocity, and Position: Following the fitness evaluation, the core mechanism of PSO operates by updating the velocity and position vectors of each particle. Table 3 details the mathematical calculations driving these movements. The 'New Velocity' column is computed based on the particle's distance from its personal best (*pbest*) and the swarm's global best (*gbest*).

Table 4. Result of particle velocity and position update

Iteration	Particle	Prev Position	Prev velocity	New Velocity	New Position	Pbest	Gbest	Silhouette	SSE	MAE
1	1	8.420	0.000	0.000	8.420	8.420	8.420	0.193	51.676	0.116
1	2	7.013	0.000	0.355	7.368	7.013	8.420	0.180	55.835	0.125
1	3	6.719	0.000	1.598	8.137	8.137	8.420	0.193	51.676	0.116
…	…	…	…	…	…	…	…	…	…	…
30	48	6.333	-0.362	2.981	9.314	7.749	8.420	0.188	49.166	0.133
30	49	7.006	-1.120	-0.176	6.830	8.112	8.420	0.180	55.835	0.123
30	50	10.000	0.988	-0.157	8.422	8.285	8.420	0.193	51.676	0.116

The velocity and position updates of particles, detailed in Table 4, demonstrate the swarm's adaptive movement toward the global best solution. Particles dynamically adjust their trajectories based on personal best (pbest) and global best (gbest) positions, ensuring convergence stability.By Iteration 30, the majority of particles converge around K ≈ 8.4, effectively stabilizing at K = 8 after discretization. This convergence behavior confirms that the swarm consistently identifies the same optimal cluster structure across iterations.

4) Convergence Check and Final Cluster Generation: Upon reaching the maximum iteration limit (30 iterations), the convergence check triggers the termination of the optimization process. The algorithm then recapitulates the performance of all candidate cluster counts (K) explored by the swarm. Table 5 presents the final summary of this optimization process.

Table 5. Summary of PSO optimization results and selection of optimal K

Candidate K	Silhouette Score	SSE	MAE	Frequency
2	0.189	87.394	0.168	21
3	0.166	77.566	0.154	12
4	0.177	69.745	0.146	18
5	0.187	63.483	0.136	33
6	0.186	60.034	0.131	133

Candidate K	Silhouette Score	SSE	MAE	Frequency
7	0.180	55.835	0.125	237
8	0.193	51.676	0.116	470
9	0.188	49.166	0.133	319
10	0.155	48.477	0.116	327

The optimization summary presented in Table 5 consolidates all candidate solutions explored during the PSO process. Among all tested configurations, K = 8 emerges as the optimal solution due to the highest Silhouette Score (0.193), competitive SSE and MAE values, and the highest convergence frequency (470 occurrences) among particles. Although K = 9 and K = 10 yield marginally lower SSE values, their inferior Silhouette Scores indicate weaker inter-cluster separation. Therefore, K = 8 is selected as the final clustering configuration.

5) Performance Evaluation: To assess the effectiveness of the proposed method, a comparative evaluation was conducted between standard K-Means and K-Means optimized using PSO across K values from 2 to 10. As shown in Figure 3, both methods exhibit decreasing error trends as K increases; however, K-Means + PSO consistently achieves lower or comparable SSE and MAE values.

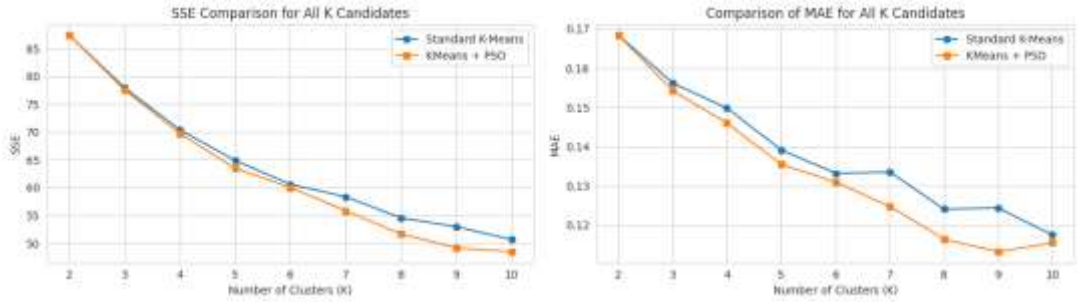

Figure 3. Comparison of SSE and MAE values K-Means vs K-Means + PSO

A focused comparison at the optimal cluster number (K = 8), summarized in Table 6, demonstrates that the proposed method reduces SSE from 54.555 to 51.676 and MAE from 0.124 to 0.116. While the numerical improvement may appear modest, this enhancement is significant given the inherent heterogeneity of MSME data.

Table 6. Comparison of evaluation metrics at K=8

Method	Cluster (K)	SSE	MAE
K-Means	8	54.555	0.124
K-Means + PSO	8	51.676	0.116

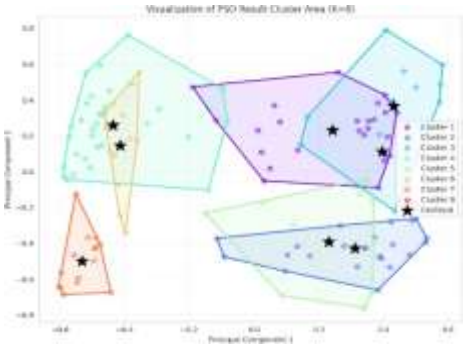

Figure 4. Visual assessment of cluster K-Means + PSO

To visually validate the clustering results, Principal Component Analysis (PCA) was applied to project the multidimensional data into a two-dimensional space. The visualization in Figure 4 illustrates eight

distinct clusters with well-defined centroids and compact intra-cluster distributions.

The use of convex hull boundaries further highlights structural coherence, indicating that the proposed method achieves stable segmentation even under dimensional reduction.

The visualization, titled 'Visualization of PSO Result Cluster Area (K=8)', plots the data along Principal Component 1 and Principal Component 2. In this chart, centroids are represented by star markers, while cluster boundaries are delineated using convex hulls. The method successfully forms eight distinct cluster structures, indicating inherent segmentation patterns within the dataset. While some geometric proximity exists, the K-Means + PSO configuration demonstrates a compact and stable arrangement with improved intra-cluster density. This visual consistency reinforces the quantitative evaluation, confirming that the proposed method effectively organizes high-variance MSME data into coherent and structurally robust segments.

While the previous evaluation confirms the numerical and structural robustness of the proposed clustering approach, quantitative metrics alone are insufficient to support policy-oriented analysis. Accordingly, a semantic interpretation was performed to translate the clustering results into meaningful MSME profiles. As presented in Table 7, the centroid values of the eight optimal clusters produced by the K-Means + PSO method were examined to characterize dominant business attributes. This analysis reveals a hierarchical segmentation of MSMEs into three distinct groups, differentiated primarily by annual turnover, level of digital maturity, and market reach.

The first segment, “Established Digital Adopter”, is represented solely by Cluster 4 and reflects the most mature businesses. This group records the highest annual turnover (IDR 100–120 million), employs a relatively larger workforce, and demonstrates advanced digital readiness through the combined use of e-commerce and social media to reach city- and province-level markets.

The second segment, “Developing Potential”, includes Clusters 1 and 7, which operate in a growth phase with moderate turnovers ranging from IDR 10–25 million. Notably, Cluster 7 achieves national market reach beyond Java Island by leveraging visual-based social media platforms, while Cluster 1 consists predominantly of younger entrepreneurs focused on local markets, indicating strong scalability potential.

The final and largest segment, “Subsistence Micro”, encompasses Clusters 2, 3, 5, 6, and 8. These enterprises generate annual turnovers below IDR 10 million and rely primarily on basic digital tools such as WhatsApp. A substantial portion of this segment consists of self-employed individuals with no employees, highlighting the need for foundational support in digital adoption and business scaling.

Table 7. Profiling of MSME clusters based on key characteristics

Cluster	Business Sector	Avg. Turnover (IDR)	Workforce	Digital Maturity	Market Reach
C1	Trade	10M - 25M	1	Basic (WA Only)	Provincial
C2	Trade	< 10M	2	Basic (WA Only)	City
C3	Trade	< 10M	0	Basic (WA Only)	Provincial
C4	Trade	100M - 120M	2	High (E-Comm + SocMed)	City & Province
C5	Creative	< 10M	2	Basic (WA Only)	City/Provincial
C6	Trade	< 10M	0	Basic (WA Only)	Provincial
C7	Creative	10M - 25M	1	Medium (IG + FB)	National
C8	Creative	< 10M	0	Basic (WA Only)	Provincial

Beyond numerical validation, centroid-based semantic analysis was conducted to translate clusters into meaningful MSME profiles. As summarized in Table 7, the eight clusters form three hierarchical segments: Established Digital Adopters, Developing Potential MSMEs, and Subsistence Micro Enterprises.

This segmentation reveals a clear gradient of digital maturity, market reach, and turnover levels, offering actionable insights for policy formulation and targeted intervention strategies.

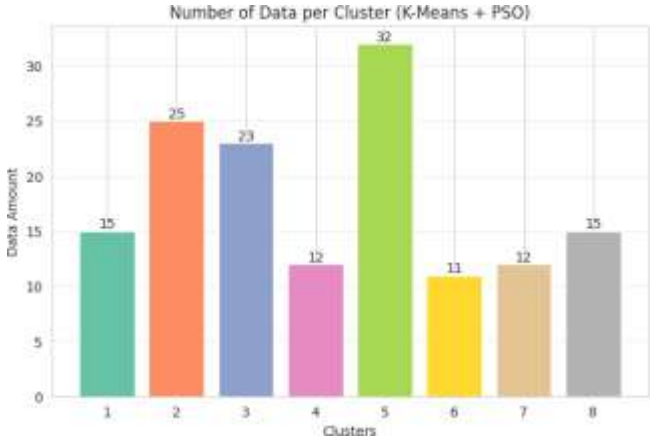

Figure 5. Distribution of data points per cluster

The integration of PSO with K-Means in this study demonstrates measurable improvements in clustering compactness and stability compared to conventional K-Means. Although the reduction in SSE and MAE appears moderate, the consistent convergence toward K = 8 across swarm iterations indicates enhanced optimization reliability.

These findings align with previous studies that reported PSO's effectiveness in improving centroid initialization and reducing local optima issues in K-Means clustering. However, unlike prior research that primarily emphasized numerical optimization, this study extends the contribution by incorporating centroid-based semantic profiling to translate computational clusters into policy-relevant MSME segments.

The hierarchical segmentation into Established Digital Adopters, Developing Potential MSMEs, and Subsistence Micro Enterprises provides a structured framework for digital transformation strategies. This multidimensional interpretation bridges the gap between algorithmic optimization and socio-economic policy application, thereby strengthening the practical significance of the proposed approach.

CONCLUSION

This study aimed to optimize the determination of the optimal number of clusters in MSME segmentation by integrating Particle Swarm Optimization (PSO) into the K-Means algorithm. The results confirm that the proposed K-Means + PSO approach enhances clustering performance, as evidenced by improved SSE and MAE values and consistent convergence toward the optimal cluster configuration (K = 8). The integration of PSO successfully mitigates the sensitivity of conventional K-Means to predefined cluster numbers and reduces the risk of suboptimal local solutions.

Beyond numerical optimization, the centroid-based semantic analysis reveals a hierarchical MSME ecosystem composed of three principal segments: Established Digital Adopters, Developing Potential Enterprises, and Subsistence Micro Enterprises. The dominance of subsistence-level clusters and the limited proportion of digitally mature enterprises highlight structural disparities within the craft-based MSME sector in Yogyakarta.

Scientifically, this research contributes by combining swarm-based optimization with semantic cluster profiling, bridging algorithmic performance improvement and policy-relevant interpretation. This integrated framework not only strengthens clustering stability but also provides a data-driven foundation for designing differentiated and targeted MSME development strategies.

REFERENCES

[1] A. Daniswara, 'Analisis Data Mining Pengelompokkan Umkm Berdasarkan Jenis Usaha Di

Provinsi Jawa Barat Menggunakan K-Means', 2024.

[2] A. A. Arrosyad, A. I. Purnamasari, and I. Ali, 'Implementasi Algoritma K-Means Clustering Untuk Analisis Persebaran Umkm Di Jawa Barat', 2024.

[3] T. Parama Yoga, 'Optimalisasi K-Means Berbasis Particle Swarm Optimization untuk Hasil Produksi Tanaman Sayuran di Indonesia', vol. 17, pp. 2614–5405, Jan. 2023, doi: 10.25134/nuansa.

[4] Y. Setiawan, 'Data Mining berbasis Nearest Neighbor dan Seleksi Fitur untuk Deteksi Kanker Payudara', vol. 8, no. 2, 2023.

[5] J. Khatib Sulaiman, N. Puspitasari, F. Urmila Jannah Helmi Puadi, and U. Mulawarman, 'Klasterisasi Wilayah Penghasil Tanaman Lada Menggunakan Algoritma K-Means', *Indonesian Journal of Computer Science*, 2022.

[6] R. Fadila, A. Rizki Rinaldi, R. Perangkat Lunak, and S. IKMI Cirebon, 'Penerapan Data Mining Dalam Mengelompokan Jumlah Umkm Berdasarkan Kabupaten Kota Menggunakan K-Means Clustering', 2024. [Online]. Available: https://opendata.jabarprov.go.id/.

[7] A. Dedi Hartono, D. Arifianto, and I. Saifudin, 'Implementasi Algoritma K-Means Dengan Particle Swarm Optimization Untuk Pengelompokan Tingkat Kesejahteraan Di Provinsi Jawa Timur', 2022.

[8] J. Penerapan, T. Informasi, D. Komunikasi, A. Franklyn, and Y. Nataliani, 'IT-Explore Pengelompokan Performa Pemain Basket Dengan Seleksi Fitur Nilai Statistik Menggunakan K-Means Dan Fuzzy C-Means', 2022.

[9] I. Faizi and N. P. Sari, 'Integrasi Kansei Engineering dan Data Mining dengan Particle Swarm Optimization pada K-Means untuk Penentuan Konsep Desain Kemasan Dodol', 2025.

[10] U. H. Laili, M. Faisal, and F. Kurniawan, 'Bulletin of Social Informatics Theory and Application Optimization of k-means clustering using particle swarm optimization algorithm for human development index', vol. 8, no. 1, pp. 144–151, 2024, doi: 10.31763/businta.v8i1.678.

[11] R. P. Ferdinandus, H. Nugroho, M. R. Daffa, E. Firmansyah, and G. E. Yuliastuti, 'Implementasi Algortima K-Means Data Penjualan UMKM Hijab Leshamysha 267', 2025.

[12] R. Fauziah and A. I. Purnamasari, 'Implementasi Algoritma K-Means pada Kasus Kekerasan Anak dan Perempuan Berdasarkan Usia', *Hello World Jurnal Ilmu Komputer*, vol. 2, no. 1, pp. 34–41, Mar. 2023, doi: 10.56211/helloworld.v2i1.232.

[13] I. Setiaji, A. Affandy, and A. Z. Fanani, 'Optimasi K-means Clustering Dengan Menggunakan Particle Swarm Optimization Untuk Menentukan Jumlah Cluster Pada Kanker Serviks', *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 3, p. 1463, Jul. 2023, doi: 10.30865/mib.v7i3.6292.

[14] Y. Yuan and Y. Li, 'A Modified Hybrid Method Based on PSO, GA, and K-Means for Network Anomaly Detection', *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/5985426.

[15] A. L. Firmansyah, B. I. Nugroho, and Z. Arif, 'Optimasi K-Means Clustering Pada Data Harga Mangga Menggunakan Particle Swarm Optimization', *Jurnal Teknologi Sistem Informasi*, vol. 6, no. 2, pp. 245–259, Sep. 2025, doi: 10.35957/jtsi.v6i2.13158.

[16] T. M. Dista and F. F. Abdulloh, 'Clustering Pengunjung Mall Menggunakan Metode K-Means dan Particle Swarm Optimization', *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 3, p. 1339, Jul. 2022, doi: 10.30865/mib.v6i3.4172.

[17] A. Yuda, 'Komparasi Algoritma K-Means dan Hierarchical Clustering Untuk Mengetahui Data Customer Dalam Layanan Internet', *INFORMATIKA SAINS TEKNOLOGI*, vol. 2, no. 2, pp. 9–14, Jul. 2024, doi: 10.34005/insit.v2i2.4119.

[18] I. Jayaperwira, A. T. Wibowo, and D. Nurjanah, 'Anime Rekomendasi Menggunakan Collaborative Filtering', 2023. [Online]. Available: www.kaggle.com[11]

[19] M. F. Zahrotun, 'K-Means Centroid Optimization with Genetic Algorithm for Clustering Micro, Small, Medium Enterprises in Yogyakarta', 2025.

[20] R. Gesit Prasasti Alam and Y. Everhard, 'Optimasi K-Means Dengan Particle Swarm Optimization Dalam Penentuan Titik Awal Pusat Klaster Data Telekomunikasi K-Means Optimization with Particle Swarm Optimization In Determining The Starting Point Of Cluster Centers of Telecommunication Data', Feb. 2024.

[21] I Made Satria Bimantara I Made Widiartha, 'Optimization Of K-Means Clustering Using Particle Swarm Optimization Algorithm For Grouping Traveler Reviews Data On Tripadvisor Sites', Jul. 2023.

[22] D. Agustina, Y. I. Mukti, and S. Muntari, 'Integrasi Particle Swam Optimization Menggunakan K-Means untuk Klasterisasi Pengangguran di Kota Pagar Alam', *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 3, no. 1, pp. 34–41, Apr. 2023, doi: 10.55382/jurnalpustakaai.v3i1.543.

[23] M. R. Putri, G. Satya Nugraha, and R. Dwiyansaputra, 'Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Menggunakan Metode K-Means Clustering Grouping Provinces in Indonesia Based on Education Indicators Using the K-Means Clustering Method', 2023. [Online]. Available: http://jcosine.if.unram.ac.id/

[24] D. Surya, D. Toresa, and S. Handayani, 'Optimasi K-Means untuk Pengelompokan Jumlah Sekolah di Riau Menggunakan Particle Swarm Optimization', 2024.

[25] A. Putri Margaretha, N. Ulinnuha, P. Keumala Intan, J. Matematika, F. Sains dan Teknologi, and U. Sunan Ampel, 'Clustering Data Kecelakaan Lalu Lintas melalui Algoritma K-Means dengan Seleksi Fitur Chi-Square 215', 2025.

[26] R. S. Lisudatu, M. Marchelin, and L. K. Wibisono, 'Klasifikasi UMKM Berdasarkan Kinerja Keuangan Menggunakan Algoritma K-Means Clustering', *Jurnal Bisnis Mahasiswa*, vol. 5, no. 2, pp. 891–900, Mar. 2025, doi: 10.60036/jbm.554.

[27] M. Khandava Mulyadien and U. Enri, 'Algoritma K-Means Untuk Pengelompokan Bantuan Langsung Tunai (BLT)', *Jurnal Ilmiah Wahana Pendidikan*, vol. 2022, no. 12, pp. 198–210, doi: 10.5281/zenodo.6944517.

[28] C. Guan, K. K. F. Yuen, and F. Coenen, 'Particle swarm Optimized Density-based Clustering andClassification: Supervised and unsupervised learning approaches', *Swarm Evol. Comput.*, vol. 44,pp. 876–896, Feb. 2019, doi: 10.1016/j.swevo.2018.09.008.

[29] M. Faisal, W. Sri Utami, and N. Pratiwi, 'Bulletin of Information Technology (BIT) Pemanfaatan Data Mining Menggunakan Metode K-Means Untuk Analisa Komoditas Telur Ayam', vol. 5, no. 4, pp. 247–254, 2024, doi: 10.47065/bit.v5i2.1670.

[30] O. Herdiana, S. Maulani, E. A. Firdaus, and K. Kunci, 'Strategi Pemasaran Produk Industri Kreatif Menggunakan Algoritma K-Means Clustering Berbasis Particle Swarm Optimization', vol. 15, 2021, [Online]. Available: https://journal.uniku.ac.id/index.php/ilkom

[31] A. Rosyada and Y. Yamasari, 'Analisis Opini Vaksin COVID-19 menggunakan SVM BerbasisPSO pada Data Twitter', *Journal of Informatics and Computer Science*, vol. 02, 2021.

[32] A. Wahyu and Rushendra, 'Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa', *(Jurnal Edukasi dan Penelitian Informatika*, vol. 8, Apr. 2022.

[33] T. Hidayat, 'Klasifikasi Data Jamaah Umroh Menggunakan Metode K-Means Clustering', *Jurnal Sistim Informasi dan Teknologi*, pp. 19–24, Feb. 2022, doi: 10.37034/jsisfotek.v4i1.115.

[34] AdityaNovita, I. Erwati, and N. Chamidah, 'Klasterisasi Provinsi Di Indonesia Berdasarkan Produktivitas Komoditas Pangan Menggunakan Algoritma K-Means', Jakarta, Aug. 2022.

[35] I. G. M. S. S. Krisna, I. W. Supriana, I. D. M. B. A. Darmawan, A. Muliantara, N. A. S. ER, and L. G. Astuti, 'Perbandingan Pengelompokan Metode PSO K-Means Dan Tanpa PSO Dalam Pengelompokan Data Alert', *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 11, no. 2, p. 283, Jul. 2022, doi: 10.24843/jlk.2022.v11.i02.p07.

[36] N. P. Sutramiani, I. M. T. Arthana, P. F. Lampung, S. Aurelia, M. Fauzi, and I. W. A. S. Darma, 'The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover', *Journal of Information Systems Engineering and Business Intelligence*, vol. 10, no. 1, pp. 13–24, 2024, doi: 10.20473/jisebi.10.1.13-24.

[37] M. A. Hosen, S. H. Moz, S. S. Kabir, S. M. Galib, and M. N. Adnan, 'Enhancing Thyroid Patient Dietary Management with an Optimized Recommender System based on PSO and K-means', in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 688–697. doi: 10.1016/j.procs.2023.12.124.

[38] H. Zhang and Q. Peng, 'PSO and K-means-based semantic segmentation toward agricultural products', *Future Generation Computer Systems*, vol. 126, pp. 82–87, Jan. 2022, doi: 10.1016/j.future.2021.06.059.

[39] Ulumuddin, 'Optimalisasi Algorithma K-Means Menggunakan Metode PSO Pada Penyakit Stunting', vol. 4, no. 1, pp. 87–91, 2024, [Online]. Available:

http://jurnal.bsi.ac.id/index.php/conten

[40] Ira Fazira, Zahratul Fitri, and Risawandi, 'Optimasi Jumlah Cluster Pada K-Means Clustering Menggunakan Particle Swarm Optimization Untuk Pengelompokan UKT Mahasiswa', *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 10, no. 2, pp. 874–886, Jul. 2025, doi: 10.36341/rabit.v10i2.6396.

[41] S. A. Utiarahman, H. Dalai, A. S. Sabudi, and A. History, 'Jurnal Teknologi dan Manajemen Informatika Optimasi K-Nearest Neighbor Dengan Particle Swarm Optimization Pada Klasifikasi Pelanggan Listrik Rumah Tangga Bersubsidi Article Info ABSTRACT', vol. 11, pp. 1–13, 2025, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi

[42] S. B. H. Sakur, 'Analisis Perbandingan Pengukuran Jarak pada Algoritme K-Means Berbasis Sum of Square Error', 2023.