Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

Clustering of MSMEs Based on Assets and Turnover Using the K-Means Algorithm

Supriatiningsih[1], Ahmad Nouvel[2], Joko Dwi Mulyanto[3], Ubaidillah[4]
Universitas Bina Sarana Informatika. Jl. Kramat Raya No.98 Jakarta, Indonesia

Article Info
Keywords:
MSMEs, K-Means algorithm, clustering, assets, turnover, business classification.

ABSTRACT
Effective UMKM assistance requires business identification and grouping. Officially, UMKM in Indonesia are divided into Micro, Small, and Medium based on assets and turnover. This research aims to group UMKM in a Regency of South Sumatra by applying the K-Means Clustering algorithm using these two variables. The research stages include business understanding, data understanding, data processing, modeling, evaluation, and dissemination. From 15 test data, this study successfully applied K-Means to classify UMKM. The result was the formation of 3 clusters, consisting of 8 data (53%) in Cluster 1, 6 data (40%) in Cluster 2, and 1 data (7%) in Cluster 3. This result has been validated using RapidMiner and shows identical outcomes. This grouping can serve as a basis for stakeholders to provide more effective assistance

Corresponding Author:
Supriatiningsih
Universitas Bina Sarana Informatika. Jl. Kramat Raya No.98 Jakarta, Indonesia
supriatiningsih.stq@bsi.ac.id

INTRODUCTION

One of the key drivers of Indonesia's economic growth is the Micro, Small, and Medium Enterprises (MSMEs) sector. Currently, there are approximately sixty-four million business units categorized as MSMEs, contributing around sixty-one percent to the national Gross Domestic Product (GDP), equivalent to Rp. 8,573.89 trillion (Kementerian Koordinator Bidang Perekonomian Republik Indonesia, 2021). In addition, MSMEs are capable of absorbing about ninety-seven percent of the total workforce (Kementerian Koperasi dan UKM, 2021). However, greater challenges remain. In analyzing the competitiveness of MSMEs, several factors need to be considered, including productivity and innovation levels, ease of doing business, access to financing and marketing, infrastructure support, and business cycles (Umkm & Indonesia, n.d.). These aspects require the involvement of all parties; not only the government but also society as a whole must play a role in improving MSME competitiveness.

Various policies, assistance programs, and training activities have been implemented to enhance MSME capacity and competitiveness. However, these efforts are often not optimal due to redundant programs, mistargeted initiatives, discontinuity, and ineffective execution. One of the main causes is the lack of comprehensive data collection and mapping of existing MSMEs. Moreover, there is no robust database to support policy design and implementation for specific MSME targets. Therefore, it is essential to group MSMEs based on their business

Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

characteristics. Through such clustering, policymakers and relevant institutions can better identify the characteristics and needs of MSMEs.

Information derived from MSME clustering is crucial, and one effective approach to obtain it is through data mining. Data mining refers to the analysis of observational data sets to discover unexpected relationships and summarize data in new, understandable, and useful ways for data owners (Setiawan & Susanto, 2022). Data mining is categorized by its tasks, one of which is clustering. Clustering is the process of partitioning a set of data objects into subsets (Daniel T. Larose, 2014). Each subset, known as a cluster, contains objects that are similar to one another but different from those in other clusters.

One commonly used clustering method is the K-Means algorithm. Clustering differs from classification because it does not rely on a target variable (Susanti et al., 2023). K-Means is one of the oldest and most widely applied clustering algorithms. The use of the K-Means algorithm for MSME grouping has been explored by several researchers, including Puntoriza and Fibriani (2020), who classified MSMEs based on their types across five districts in Malang City. Their computation using K-Means produced three clusters (Puntoriza & Fibriani, 2020). Similarly, Magdalena and Fahrudin (2020) applied the K-Means algorithm to cluster cooperatives in West Java based on internal capital, external capital, and business volume, resulting in three clusters: high, medium, and low (Magdalena & Fahrudin, 2020).

In this study, MSME grouping is conducted based on asset and turnover variables, referring to Law of the Republic of Indonesia Number 20 of 2008 concerning Micro, Small, and Medium Enterprises (Republik Indonesia, 2008). The results of the K-Means clustering process are expected to serve as valuable input for policymakers and stakeholders, such as the Department of Cooperatives and MSMEs, in guiding the assistance, development, and management of MSMEs. Ultimately, this approach aims to foster stronger and more sustainable MSME growth.

METHODS

This study adopts a quantitative descriptive approach supported by data mining techniques, focusing on the application of the K-Means algorithm to classify Micro, Small, and Medium Enterprises (MSMEs) based on their financial characteristics. The research aims to identify natural groupings among MSMEs by analyzing asset and turnover variables. This method is chosen because it allows the researcher to uncover patterns and similarities within the data that may not be visible through traditional statistical approaches. The study emphasizes data-driven insights to support policymakers and institutions in designing targeted empowerment programs for MSMEs in Indonesia.

The research process follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which consists of six sequential stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This framework provides a structured and systematic approach, ensuring that the analysis is both accurate and meaningful.

Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

Business Understanding Phase

In this phase, the researcher identifies the main objective, which is to map MSMEs based on their financial capacity to support more effective policy formulation. The research seeks to determine how MSMEs can be grouped according to their asset size and turnover, in line with Indonesia's Law No. 20 of 2008 on Micro, Small, and Medium Enterprises. The classification results are expected to inform government strategies for financing and training MSMEs according to their respective financial strengths.

Data Understanding Phase

The study collects and examines secondary data from government and institutional records related to MSME performance in South Sumatra Province. The dataset includes information on:

a. Enterprise Name
b. Total Assets (A)
c. Turnover (T)

These variables were selected based on their alignment with MSME classification standards under Indonesian law. Before analysis, a preliminary exploration is conducted to identify outliers, missing values, or inconsistencies that may bias the clustering results.

Data Preparation Phase

The data undergo several cleaning and transformation processes to ensure consistency and reliability. Non-numerical data such as enterprise names are removed, and numerical attributes (assets and turnover) are standardized to eliminate scale bias between variables.

The Z-score normalization technique is applied to standardize data as follows:

$$Z_i = \frac{X_i - \mu}{\sigma}$$

This normalization ensures that both variables contribute equally to the clustering analysis. The standardized data are then compiled into a structured format suitable for processing by the K-Means algorithm.

Modeling Phase

The K-Means clustering algorithm is used to classify MSMEs into groups by minimizing the distance between data points and their corresponding cluster centroids. The objective function of K-Means can be expressed as:

$$\min_{C} \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

The Euclidean distance is calculated as:

$$d(x, \mu) = \sqrt{\sum_{k=1}^{m} (x_k - \mu_k)^2}$$

where mmm is the number of attributes (in this case, two: assets and turnover).

Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

a. The K-Means algorithm proceeds through the following iterative steps:
b. Initialize KKK centroids randomly.
c. Assign each data point xjx_jxj to the nearest centroid using the Euclidean distance formula.

Update each centroid to the mean of all data points within that cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

Repeat steps (2) and (3) until convergence is achieved, meaning cluster memberships no longer change significantly. In this study, the value of KKK is set to three (3), corresponding to the micro, small, and medium classifications as defined by national MSME policy.

Evaluation Phase

To assess the quality of the clustering results, internal validation is performed using the Silhouette Coefficient (S), which measures the degree of cohesion and separation among clusters. It is calculated as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Values of S(i) close to +1 indicate that the object is well clustered, while values near 0 suggest overlapping clusters.

In addition, visual inspection of the cluster distribution is conducted through scatter plots to ensure the clusters are well separated. The clustering process and validation are conducted using RapidMiner software, which provides computational precision and visualization tools to interpret results effectively.

Interpretation and Deployment Phase

After validation, the characteristics of each cluster are analyzed to interpret the defining attributes of micro, small, and medium enterprises. Descriptive statistics such as mean, minimum, maximum, and range for both asset and turnover variables are computed for each cluster to highlight the financial differences among them.

This analytical stage produces an empirical basis for understanding which MSMEs may require financial assistance, training programs, or technological support. The results are then discussed in relation to national MSME development policies, aligning with prior studies that emphasize data-driven classification as a foundation for sustainable empowerment and policy formulation.

Software and Validation Tools

All data processing, modeling, and evaluation procedures are performed using RapidMiner Studio 10.0, which provides integrated tools for data mining, statistical computation, and visualization. The results are cross-validated by comparing computational outputs and cluster statistics to ensure reproducibility and accuracy.

Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

RESULTS AND DISCUSSION

Application of the K-Means Clustering Algorithm to MSME Data

Based on the MSME data that have been transformed (as explained in Table 4 of the Methodology section), the next step is the modeling process using the K-Means clustering algorithm. The procedures are carried out as follows:

1. Determining the Number of Clusters

The number of clusters is determined by referring to the classification of MSMEs as regulated in Law of the Republic of Indonesia Number 20 of 2008 concerning Micro, Small, and Medium Enterprises. In this regulation, MSMEs are categorized according to their assets and turnover for each business scale, as presented in Table 1.

Table 1. Classification of Micro, Small, and Medium Enterprises (MSMEs)

Business Scale	Net Worth / Assets	Sales Revenue / Turnover
Micro Enterprise	Maximum of Rp 50 million	Maximum of Rp 300 million
Small Enterprise	More than Rp 50 million – up to Rp 500 million	More than Rp 300 million – up to Rp 2.5 billion
Medium Enterprise	More than Rp 500 million – up to Rp 10 billion	More than Rp 2.5 billion – up to Rp 50 billion

Based on the classification presented in the table, the number of clusters (k) used in this study is three.

2. Initialization of Random Centroids

The initial centroids are determined randomly, with the number of centroids corresponding to the total number of clusters to be formed. In this study, the initial cluster centers are defined as m_1 = b (2,2), m_2 = f (1,2), and m_3 = o (3,3). These points serve as the preliminary reference positions for the clustering process before iteration begins.

3. Calculating the Distance to Each Centroid

To calculate the distance between each data point and the centroids, the K-Means distance formula, previously defined in the methodology sectio, is applied. The Euclidean distance is used to measure similarity between data points and centroid positions. Table 2 presents the calculation of distances for each MSME to its corresponding centroid, along with the results of data grouping from the first iteration.

Table 2. Distance Calculation and Data Grouping in the First Iteration

No.	MSME Name	Transformed Turnover	Transformed Assets	Distance to Centroid 1	Distance to Centroid 2	Distance to Centroid 3	Cluster
1	Warung Tati	2	2	0	1	2	1
2	Toko Zhi-Zhi Bengkel Las	1	2	1	1	2	2
3	Kemang Tanjung	1	2	2	1	2	2
4	Toko Tasya	2	2	0	1	2	1

Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

No.	MSME Name	Trans-formed Turnover	Trans-formed Assets	Distance to Cen-troid 1	Distance to Cen-troid 2	Distance to Cen-troid 3	Clus-ter
5	Fotokopi Na-ura	1	1	1	0	3	2
6	Bengkel Las Sony	1	2	1	1	2	2
7	Toko Grosir 3 Saudara	3	2	2	2	3	1
8	Warung Wahyu	2	1	1	2	3	2
9	Toko Baju Yossy YM	1	2	1	1	2	2
10	Rudi / Toko Tasiya	2	1	0	1	2	1
11	H. Sidi Zai-nudin	2	2	1	2	3	1
12	Bambang	2	2	0	1	3	1
13	Aida	1	1	1	0	3	2
14	M. Yusuf / Bakso Kepala Sapi	2	2	0	1	2	1
15	Nin / Yoga Mobilindo	3	3	2	3	0	3

In the first iteration, the results differed from the initial data grouping; therefore, the process continued to the next iteration. The subsequent step involves recalculating the centroid values based on the data points assigned to each cluster in the previous iteration. Using the previously defined formula, the new centroids are obtained as follows:

a. Centroid 1: (2,2) → (2.125, 1.625)
b. Centroid 2: (1,2) → (1, 1.66667)
c. Centroid 3: (3,3) → (3,3)

After determining the new centroid values, the distance for each data point to its corresponding centroid is recalculated using the same Euclidean distance formula applied in the first iteration. The results of the distance computation and the updated data grouping for the second iteration are presented in Table 3.

Table 3. Distance Calculation and Data Grouping in the Second Iteration

No.	MSME Name	Trans-formed Turnover	Trans-formed Assets	Distance to Centroid 1	Distance to Centroid 2	Distance to Cen-troid 3	Clus-ter
1	Warung Tati	2	2	0.26563	1.11111	2	1
2	Toko Zhi-Zhi	1	2	0.26563	1.11111	2	1

Jurnal Multidisiplin Sahombu
Volume 5, Number 07, 2025, DOI 10.58471/jms.v5i07
ESSN 2809-8587 (Online)
https://ejournal.seaninstitute.or.id/index.php/JMS/index

No.	MSME Name	Trans-formed Turnover	Trans-formed Assets	Distance to Centroid 1	Distance to Centroid 2	Distance to Cen-troid 3	Clus-ter
3	Bengkel Las Kemang Tanjung	1	2	1.51563	0.44444	1	2
4	Toko Tasya	2	2	0.26563	1.11111	2	1
5	Fotokopi Naura	1	1	1.26563	0.11111	3	2
6	Bengkel Las Sony	1	2	1.51563	0.44444	2	2
7	Toko Grosir 3 Saudara	3	2	2.26563	2.44444	3	3
8	Warung Wahyu	2	1	0.51563	1.44444	3	1
9	Toko Baju Yossy YM	1	2	1.51563	0.44444	2	2
10	Rudi / Toko Tasiya	2	1	0.51563	1.44444	3	1
11	H. Sidi Zai-nudin	2	2	1.26563	1.11111	3	1
12	Bambang	2	2	0.26563	1.11111	3	1
13	Aida	1	1	0.26563	0.11111	3	2
14	M. Yusuf / Bakso Kepala Sapi	2	2	1.26563	1.11111	3	1
15	Nin / Yoga Mobilindo	3	3	2.76563	3.77778	0	3

The data grouping results obtained in the second iteration show that there were no cluster changes, indicating convergence. Therefore, the calculation of distances was stopped. The final grouping results of MSME data using the K-Means clustering algorithm are as follows:

a. Cluster 1: Consists of eight MSMEs — Warung Tati, Toko Zhi-Zhi, Toko Tasya, Toko Grosir 3 Saudara, Warung Wahyu, Rudi / Tk. Tasiya, Bambang, and Aida.
b. Cluster 2: Consists of six MSMEs — Bengkel Las Kemang Tanjung, Fotokopi Naura, Toko Baju Yossy YM, H. Sidi Zainudin, and M. Yusuf / Bakso Kepala Sapi.
c. Cluster 3: Consists of one MSME — Nin / Yoga Mobilindo.

Analysis and Discussion

The results of the clustering process using the K-Means algorithm based on assets and turnover converged in the second iteration, producing three distinct clusters. Cluster 3, which contains only Nin / Yoga Mobilindo, clearly represents the Medium Enterprise category, with transformed turnover and asset values of (3,3), meaning it has a turnover of more than 2.5

billion and assets exceeding 500 million. Cluster 2, comprising six MSMEs, represents the Small Enterprise group, with most data points having transformed values of (1,2) or (2,1) for turnover and assets. Cluster 1, consisting of eight MSMEs, corresponds to the Micro Enterprise group.

These findings, successfully identifying three MSME clusters, are consistent with previous studies that also utilized the K-Means algorithm for mapping purposes. Puntoriza and Fibriani (2020) similarly identified three clusters in their analysis of MSME distribution in Malang, although their classification was based on business types. Likewise, Magdalena and Fahrudin (2020) identified three clusters, high, medium, and low, when clustering cooperatives in West Java. These consistent outcomes reinforce the validity and robustness of the K-Means algorithm as an effective tool for segmentation and mapping of MSME characteristics.

The novelty of this study lies in its use of turnover and asset variables that directly reference Law of the Republic of Indonesia Number 20 of 2008 as the basis for data transformation and classification parameters. The resulting clusters have significant managerial implications. With the established grouping, policymakers and relevant institutions can now design more precise and targeted MSME development programs. Assistance for Cluster 1 (Micro Enterprises) can focus on business legality and initial capital support, while programs for Cluster 2 (Small Enterprises) can emphasize managerial improvement and marketing development. Consequently, this clustering approach enables more accurate, efficient, and impactful support for MSME growth and sustainability.

CONCLUSION

This study demonstrates the effectiveness of the K-Means clustering algorithm in classifying Micro, Small, and Medium Enterprises based on asset and turnover variables. The clustering process successfully identified three distinct groups that align with the official business classifications established by national regulations. The findings reveal that the Micro Enterprise group dominates the dataset, followed by the Small Enterprise group, while the Medium Enterprise category is represented by only one unit. These results indicate a significant concentration of small-scale businesses within the studied region, emphasizing the need for targeted strategies to enhance their capacity and competitiveness. The study also confirms the consistency of the K-Means algorithm with previous research, validating its reliability as a data-driven decision support tool. The use of asset and turnover indicators, as outlined in the Indonesian MSME Law, provides an accurate and relevant foundation for classification. From a managerial perspective, the clustering outcomes can serve as a practical reference for policymakers and development agencies to design more focused programs according to each group's specific needs. Overall, this research reinforces the importance of data mining techniques in supporting evidence-based policymaking and promoting inclusive economic development through MSME empowerment.

REFERENCE

Aggarwal, C. C. (2021). *Data mining: The textbook* (2nd ed.). Springer.

Daniel T. Larose, & Larose, C. D. (2014). *Discovering knowledge in data*. John Wiley & Sons.

Fahmi, D., Halimah, I., & Yusuf, Y. (2024). Sosialisasi Penyusunan Laporan Keuangan Guna Meningkatkan Usaha Umkm di Pokdarwis Ekowisata Keranggan Tangerang Selatan. *Jurnal Abdi Masyarakat Multidisiplin*, *3*(2), 13-18.

Garcia, L., & Patel, R. (2021). A review of data preprocessing techniques for machine learning. *Journal of Data Science, 19*(2), 113–130.

Hidayah, A. (2021). Implementing Data Clustering to Identify Capital Allocation for Small and Medium Sized Enterprises (SMEs). *ASEAN Marketing Journal*, *10*(1), 5.

Jain, A. K. (2022). Advancements in clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 20–35.

Jonathan, J. K., Herwindiati, D. E., Ferdinand, K., & Jong, F. (2024, October). Implementation of Fuzzy C-Means for Clustering MSMEs in Jambi Province. In *2024 Ninth International Conference on Informatics and Computing (ICIC)* (pp. 1-5). IEEE.

Kementerian Koordinator Bidang Perekonomian Republik Indonesia. (2021). *UMKM menjadi pilar penting dalam perekonomian Indonesia*. https://www.ekon.go.id/publikasi/detail/2939/dukungan-pemerintah-bagi-umkm-agar-pulih-di-masa-pandemi

Magdalena, L., & Fahrudin, R. (2020). Penerapan data mining untuk koperasi se-Jawa Barat menggunakan metode clustering pada Kementerian Koperasi dan UKM. *Jurnal Digit, 9*(2), 190. https://doi.org/10.51920/jd.v9i2.120

Puntoriza, P., & Fibriani, C. (2020). Analisis persebaran UMKM Kota Malang menggunakan cluster K-Means. *JOINS (Journal of Information Systems), 5*(1), 86–94. https://doi.org/10.33633/joins.v5i1.3469

Republik Indonesia. (2008). *Undang-Undang Nomor 20 Tahun 2008 tentang Usaha Mikro, Kecil, dan Menengah.*

Schröder, M., & Wirth, R. (2022). The CRISP-DM methodology in practice: A systematic review. *International Journal of Data Science and Analytics, 13*(1), 1–24.

Sukmadewanti, I., Arifudin, R., & Sugiharti, E. (2018). Use of K-means clustering and analytical methods hierarchy process in determining the type of msme financing in semarang city. *Sci, J. Inform*, *5*(2), 148-158.

Susanti, N., Prasetyo, H., & Lestari, M. (2023). Kajian potensi e-commerce sebagai solusi ekspansi pasar UMKM pasca pandemi COVID-19. *Jurnal Ekonomi Digital, 5*(1), 42–50.

Sutramiani, N. P., Arthana, I., Aurelia, S., Fauzi, M., & Surya Darma, I. (2024). The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover. *Journal of Information Systems Engineering & Business Intelligence*, *10*(1).

Tan, P.-N., Steinbach, M., & Karpatne, A. (2024). *Introduction to data mining* (3rd ed.). Pearson.

Widanengsih, E., & Yusuf, Y. (2025). Design of an Application-Based Sales Information System for Koperasi XYZ. *Jurnal Multidisiplin Sahombu*, *5*(02), 538-552.