Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Evaluation of K-Means, DBSCAN, and Hierarchical Clustering for Strategic Segmentation of Tourism SMEs in Rembang, Indonesia

Ardiansyah Ramadhan*[1], Fandi Achmad[2], Ibnu Zulkarnain[3], Masayoshi Aritsugi[4]

[1]Computer Engineering, Telkom University, Indonesia
[2]Industrial Engineering, Telkom University, Indonesia
[3]Industrial Engineering, National Taiwan University of Science and Technology, Taiwan
[4]Faculty of Advanced Science and Technology, Kumamoto University, Japan

Email: [1]ardiansyahramadhanar@telkomuniversity.ac.id

Received : Apr 13, 2025; Revised : Jun 23, 2025; Accepted : Jul 1, 2025; Published : Jul 9, 2025

Abstract

Small and Medium Enterprises (SMEs) play a crucial role in job creation, regional competitiveness, and economic equity. In the tourism sector, particularly in ecotourism and cultural tourism, clustering SMEs presents challenges due to complex and interrelated data variables. This study aims to evaluate the effectiveness of three clustering algorithms—K-Means, DBSCAN, and Hierarchical Clustering—in segmenting SMEs based on real-world tourism datasets. A purposive sampling method was applied to 203 valid respondents from SMEs in Rembang Regency, Central Java. Clustering performance was assessed using the Silhouette Coefficient and Davies-Bouldin Index, while computational efficiency and scalability were analyzed through execution time and memory usage. The results show that DBSCAN achieved the best clustering quality (Silhouette Coefficient: 0.5496, Davies-Bouldin Index: 0.3298), effectively managing noise and irregular cluster shapes. Hierarchical clustering offered moderate quality and helped reveal relationships between SMEs. In contrast, K-Means demonstrated the lowest quality (Silhouette Coefficient: 0.2321) due to its limitation in handling non-spherical clusters. For computational efficiency, Hierarchical Clustering required the least memory (0.14 MB) and shortest execution time (5.73 seconds), while K-Means took the longest time (26.00 seconds). DBSCAN consumed more memory due to density-based processing. K-Means was the most stable in scalability testing with increasing dataset sizes, whereas Hierarchical Clustering showed inefficiency. The findings support selecting appropriate clustering methods based on data complexity and size. This study enhances data-driven tourism development strategies and advances clustering methodology for applied informatics. Future work may explore hybrid clustering and predictive models for deeper insights.

Keywords: *Clustering Algorithm, DBSCAN, Hierarchical Clustering, K-Means, Tourism Industry.*

1. INTRODUCTION

The Indonesian economy, particularly in the industrial sector, is advancing swiftly [1], [2]. Small and Medium Enterprises (SMEs) constitute a fundamental component of this industry [3]. Small and medium-sized enterprises (SMEs) are integral to creating employment, improving regional competitiveness, and fostering economic equity in the tourism sector [4], [5]. Despite their importance, a systematic method for delineating the distribution and evolution of SMEs across various tourism sectors remains insufficiently established [6]. Effective mapping is crucial as it offers strategic insights for local governments and industry stakeholders, allowing them to formulate more focused policies that promote the growth and sustainability of SMEs [4], [5]. In tourism, delineating industrial sectors improves managerial efficiency and efficacy [7], [8]. Attributes obtained by clustering these tourism-related SMEs can be utilized to identify prospective industry segments and guide decision-making processes [8], [9].

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Every district in Indonesia possesses distinct social, economic, and topographical attributes, which affect the industries that emerge in each area [10]. Rembang Regency in Central Java possesses significant tourism potential due to its abundant artistic and cultural history [4], [11]. Rembang is home to various SMEs in the arts, performing arts, handicrafts, food and beverage, clothes and fashion, art markets, transportation, and lodging [5], [12]. These SMEs facilitate tourism by providing essential services, distinctive local products, and cultural experiences that augment visitor appeal [13]. Consequently, categorizing sub-districts according to predominant industry types can yield significant insights into regional potential and aid in developing more successful policies [4], [5], [9].

This research examines the clustering of SMEs in Rembang Regency utilizing 2023 data, encompassing variables such as telecommunication infrastructure, transportation, energy sources, waste management, geographical location, clean water availability, supporting industries, spatial distribution, hospitality, safety and security, stakeholder engagement, and environmental dynamism [12]. Clustering regions is intricate [14], [15]. Tourism data frequently comprises many properties, such as numerical, categorical, and spatial variables, necessitating meticulous preprocessing, feature selection, and normalization to achieve significant clustering results [14], [15]. Moreover, absent or erratic data might engender biases, hence exacerbating the precision and dependability of grouping [16].

Various clustering techniques, such as K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Hierarchical Clustering, can be employed to address challenges, biases, and complications in accuracy [17], [18], [19]. Chaudhry (2023) [20] completed a systematic literature review on unsupervised clustering approaches but did not include practical implementation in real-world tourism datasets. Truabasarj and Permadi's (2024) [17] research using K-Means, DBSCAN, and Hierarchical Clustering to evaluate recommendations for fixed broadband sales locations. This study did not investigate clustering with datasets comprising varied features or industry categories.

K-Means is a prevalent technique recognized for its computational efficiency and appropriateness for datasets with predetermined cluster counts, rendering it optimal for delineating tourist sectors based on particular characteristics such as visiting rates or industry classifications [15], [17], [21]. Nonetheless, it is constrained by restrictions, including sensitivity to outliers and the requirement for predetermined cluster quantities, which may affect clustering precision [21].

DBSCAN is beneficial for detecting clusters with odd shapes and diverse densities, rendering it appropriate for tourism areas with unevenly distributed data [17], [22]. In contrast to K-Means, DBSCAN can identify outliers and does not necessitate the predefinition of the number of clusters. Nonetheless, its efficacy may diminish in high-dimensional datasets or when parameter optimization, such as epsilon values, is insufficient [17], [19], [20], [22].

Conversely, hierarchical clustering offers a hierarchical depiction of clusters, facilitating the comprehension of linkages between primary and subordinate tourism destinations [17], [23]. Although flexible, its significant computational complexity renders it less suitable for large datasets [23]. Moreover, G. J. Oyewole and G. A. Thopil (2023) [14] observed that the applications and trends of data clustering have experienced significant expansion, particularly within the tourist sector. Nonetheless, their research did not emphasize the technical execution of clustering algorithms as a strategic instrument or decision-making framework in the tourism industry. Manhoor (2024) [19] addressed the acceleration of clustering as a notable issue and opportunity, necessitating innovative methodologies for future progress.

In addition to algorithm selection, there are methodological obstacles in tourism clustering [19]. Feature selection, parameter optimization, and managing missing or noisy data are essential for achieving reliable clustering results [24]. Inadequate preprocessing or suboptimal parameter selection might result in erroneous clusters, diminishing the generalizability and trustworthiness of results [25]. Furthermore, clustering outcomes must be adeptly presented and evaluated to yield valuable insights for

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

tourism policymakers and stakeholders [26]. Confronting these methodological problems is essential for enhancing the relevance of clustering techniques in tourist research [24], [26].

This research utilized a questionnaire method administered to SMEs in Rembang Regency, Central Java, employing a purposive sample technique [4], [5]. The questionnaire was constructed using a 6-point Likert scale, omitting a neutral option to minimize bias and compel respondents to submit unequivocal answers from "strongly disagree" to "strongly agree" [4], [5]. Nevertheless, the prior studies [4], [5], which concentrated on fostering sustainable performance in SMEs, did not consider clustering. This study primarily concentrates on clustering. Before choosing the most efficacious clustering algorithm as a strategic method for various tourism sectors, it is vital to assess clustering performance. The study utilizes the Silhouette Coefficient and the Davies-Bouldin Index to accomplish this objective. The Silhouette Coefficient examines intra-cluster cohesion and inter-cluster separation, offering insights into clustering quality.

In contrast, the Davies-Bouldin Index measures intra-cluster similarity and inter-cluster differences, assisting in assessing clustering efficacy [27]. These measures were selected for their capacity to assess clustering structures in intricate datasets [27]. Nonetheless, alternative metrics like the Dunn Index and the Calinski-Harabasz Index may be investigated in subsequent studies to refine clustering assessments further and improve result dependability.

This study's research contributions are as follows:

1. A comparative analysis of clustering algorithms (K-Means, DBSCAN, Hierarchical Clustering) applied to real-world tourist datasets, offering practical insights into their efficacy for various data formats and industry demands.
2. This research customizes clustering applications for tourism SMEs, providing strategies to enhance various industry sectors, including arts, performing arts, handicrafts, food and beverage, clothing and fashion, art markets, transportation, and accommodation.
3. The study connects clustering algorithm efficacy with strategic decision-making for tourist stakeholders, merging machine learning with policy development to improve industry advancement.

This work seeks to address theoretical and methodological problems to enhance comparing clustering methodologies in tourist research, thereby promoting informed, data-driven decision-making for regional development.

2. METHOD

This research methodology typically gathers and analyses data through machine learning to categorize the acquired information. Enhancing the identification of the optimal clustering method for small and medium enterprises can be achieved by comparing K-Means, DBSCAN, and Hierarchical Clustering. Each model is subjected to tuning to enhance its performance, and the evaluation outcomes function as recommendations, as depicted in Figure 1. This procedure entails comparing clustering results, assessing regularisation, and analyzing programming interpretation.

Following model tuning, assessments of computational efficiency, strong scalability, and weak scalability are performed to evaluate the feasibility of each clustering approach. Computational efficiency assessments evaluate execution duration and memory consumption, confirming the algorithms' applicability in practical scenarios. Robust scalability testing assesses the reduction in execution time as additional processing cores are employed, offering insights into the efficiency of parallel computing. Weak scalability testing evaluates if the method sustains consistent execution time as dataset size expands alongside proportional computer resources, confirming its efficacy for extensive and expanding tourism datasets. These assessments ascertain the most effective and scalable clustering

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

technique, consistent with the study's aim of enhancing data-driven decision-making in the tourism sector.

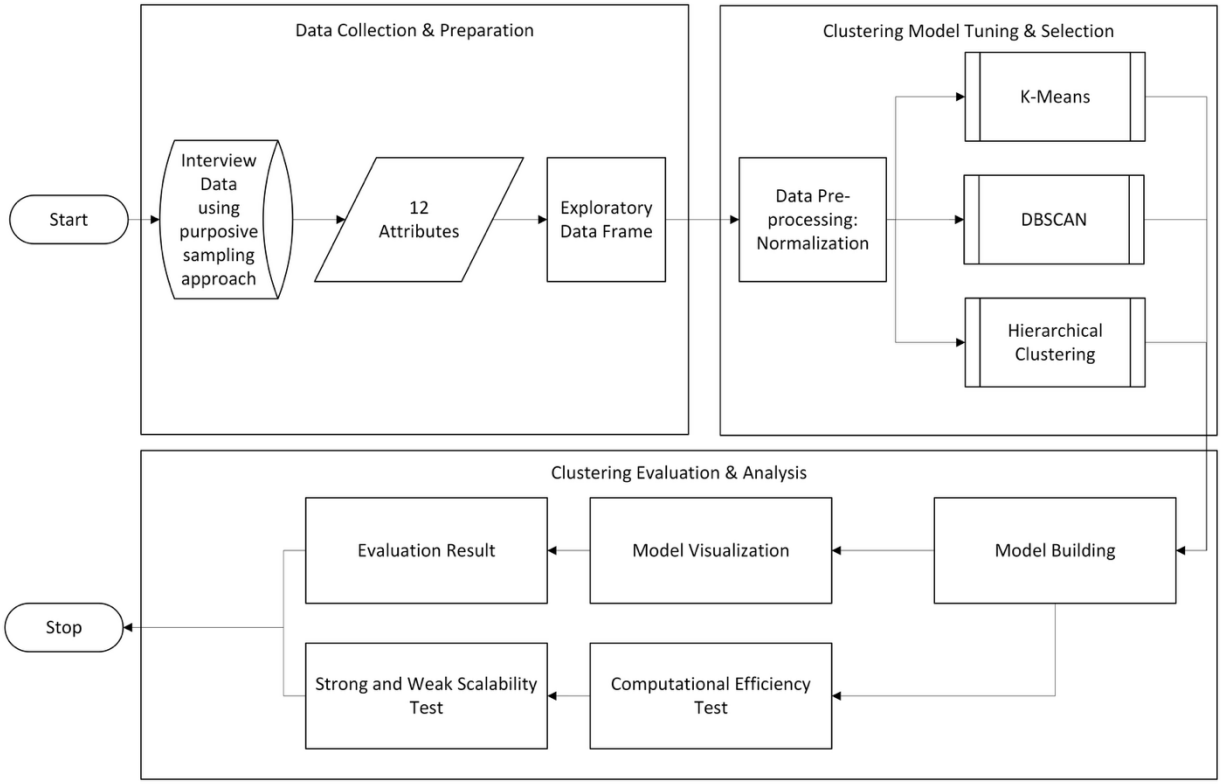

Figure 1. Methodology for evaluating clustering algorithms within the tourism sector

2.1. Data Collection and Preparation

This study's data were gathered by a questionnaire administered to SMEs in Rembang Regency, Central Java, utilizing a purposeful sample methodology [4], [5]. Respondents were selected based on geographic region, employee count, and kind of small and medium-sized enterprise (SME). Before distributing the questionnaire, preliminary research was undertaken with academic specialists in tourism, SME coordinators, and tourist coordinators to ascertain the relevance and clarity of the questions [4], [5]. the questionnaire included a 6-point Likert scale devoid of a neutral option, necessitating respondents to select a definitive response ranging from "strongly disagree" to "strongly agree" [4], [5]. The collected sample size of 203 out of 219 respondents (92.69%) satisfied the validity criteria for data analysis, with proportions representing the various types of SMEs in the region [4], [5].

This research technique seeks to enhance the efficacy of clustering algorithms—K-Means, DBSCAN, and Hierarchical Clustering—in delineating the tourism sector. Data were gathered from 203 respondents, all of whom are proprietors of diverse small and medium enterprises (SMEs). The data-gathering approach was executed via organized interviews and the distribution of questionnaires. Participants were chosen from various sites within the Kaliori, Kragan, Lasem, Pamotan, Pancur, Rembang, Sale, Sedan, Sluke, and Sulang sub-districts.

Table 1. Respondents

No	Respondents	Type of SMEs or Industries
1	Respondent 1	Art and Performance SMEs
	Respondent 2	
	…	

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

	Respondent 17	
2	Respondent 18 Respondent 19 … Respondent 60	Handicraft Industry
3	Respondent 61 Respondent 62 … Respondent 119	Food and Beverage Industry
4	Respondent 120 Respondent 121 … Respondent 162	Apparel and Fashion Industry
5	Respondent 163 Respondent 164 … Respondent 166	Art and Antiques Market Industry
6	Respondent 167 Respondent 168 … Respondent 170	Performing Arts Industry
7	Respondent 171 Respondent 172 … Respondent 203	Transportation and Accommodation Industry

The dataset includes 12 attribute aspects: telecommunications, transportation, electricity resources, waste management, location, clean water sources, supporting industries, spatial planning, hospitality, security and safety, stakeholders, and environmental dynamics [4], [5], [11], [12].

Table 2. Indicators

Attributes	Code	Indicators
Telecommunication	TL 1	Cellular towers are present around the tourist area.
	TL 2	Comprehensive information about tourism destinations is available in print media, on social media, or through websites.
	TL 3	Easy to get the latest information about tourism objects.
	TL 4	There is a stable internet network in the object.
	TL 5	Can make phone calls or send messages inside the tourist area.
	TL 6	An easily accessible list of emergency contact numbers is available.
Transportation	TP 1	The roads leading to the tourist attractions are in good condition.
	TP 2	Tourist destinations are accessible by both private and public transportation.
	TP 3	Affordable public transportation options are available.
	TP 4	Multiple alternative routes provide access to the tourist area for both private and public transportation.
Power Source	PS 1	There is an electricity network in the tourist area.
	PS 2	The electricity network can be used properly in tourist areas.
	PS 3	There are adequate generators to anticipate if there is a problem in the tourist area.
	PS 4	Repairs to the electricity network in the tourist area can be completed quickly (within one hour).

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

	PS 5	Electricity network checks are carried out regularly (once a month).
	PS 6	Daily garbage collection is carried out in the tourist area.
Waste Management	WM 1	Daily garbage collection is carried out in the tourist area.
	WM 2	Waste is sorted by cleaning staff within the tourist area.
	WM 3	Sufficient and easily accessible trash bins are available throughout the tourist area.
	WM 4	Trash bins are provided and categorized by waste type (e.g., paper, plastic, organic).
	WM 5	Tourist attractions are consistently kept clean.
Location	LT 1	Tourist attractions are located near public facilities such as hotels, restaurants, and minimarkets.
	LT 2	Tourist attractions are located close to the city center.
	LT 3	Many tourist destinations are situated on the outskirts of the city.
	LT 4	Both visitors and vendors can easily reach the location.
	LT 5	Tourist attractions offer captivating views.
	LT 6	Attractions provide appealing photo spots.
Clean Water Source	CW 1	Clean water is readily available in the tourist area.
	CW 2	Proper drainage channels exist in the tourist area.
	CW 3	Drainage systems are well-maintained.
Supporting Industry	SI 1	There is effective coordination between the tourist attraction management and government or private entities.
	SI 2	Collaboration with industries supports the development of facilities and infrastructure in tourist attractions.
	SI 3	Local communities are involved in managing the tourist area.
	SI 4	A handicraft center is available within the tourist area.
	SI 5	A culinary center exists in the tourist area.
	SI 6	A souvenir shop selling items unique to the tourist area is available.
	SI 7	Travel agents provide services to the tourist area.
Spatial	ST 1	Unique rides reflecting local culture, traditions, history, or natural scenery are offered.
	ST 2	Tourist attraction managers integrate local wisdom into their innovations.
	ST 3	The location is near other tourist attractions.
Hospitality	HT 1	An information center is available to assist visitors.
	HT 2	Cleaning staff maintain cleanliness in the tourist area.
	HT 3	A designated health area is present within the tourist area.
	HT 4	Ambulance services are available in the tourist area.
	HT 5	Medical personnel are on-site in the tourist area.
	HT 6	Hotels near the tourist area are easily accessible.
	HT 7	Tour guide services are provided.
	HT 8	A large parking area is available for visitors.
	HT 9	Facilities such as rinse areas, clean toilets, and sinks are provided.
Safety and Security	SS 1	Visitor checks are conducted at the entrance to the tourist attraction.
	SS 2	Security officers are present within the tourist area.
	SS 3	Disaster safety procedures are established in the tourist area.
	SS 4	A designated gathering point for emergencies is available in the tourist area.

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Stakeholder	SH 1	Stakeholders, including SMEs, industries, and government entities, actively promote tourism through social media and similar platforms.
	SH 2	Stakeholders organize special programs to boost tourism promotion.
	SH 3	Training programs for local residents are conducted to enhance skills in tourism management and promotion.
	SH 4	Communication between local residents and stakeholders regarding tourism matters is facilitated.
	SH 5	Stakeholders collaborate to develop facilities and infrastructure around the tourist area.
	SH 6	Stakeholder policies support the availability of facilities and infrastructure in the tourist area.
	SH 7	Stakeholders plan programs aimed at enhancing tourism potential.
Enviromental Dynamism	ED 1	Local residents support initiatives to develop tourism potential.
	ED 2	Residents actively participate in tourism development efforts.
	ED 3	Over the past year, the environment in the tourist area has undergone significant changes, particularly during the pandemic.
	ED 4	Changes in the habits of local residents near the tourist area occur annually.
	ED 5	New tourism potentials are added to the area each year.
	ED 6	The government organizes regular programs to promote tourism.
	ED 7	Local residents are supportive of the efforts to enhance tourism potential.

2.2. Cluster Model Tuning and Selection

Prior to the application of clustering techniques, the data is subjected to preprocessing to guarantee consistency and dependability. The preprocessing procedures encompass addressing missing values through deletion and imputation methods. The relevant entries are eliminated if an attribute contains missing values beyond 20% of the total dataset; otherwise, mean or median imputation is utilized according to the data distribution. Outliers are determined by the Interquartile Range (IQR) approach and Z-score analysis. Significant outliers over three standard deviations are eliminated or substituted using winsorization to avert distortion in clustering outcomes. Min-Max normalization addresses scale discrepancies, converting data to a range of [0,1] to ensure equal scaling and mitigate the influence of qualities with broader numerical ranges.

2.2.1 Outlier Detection

All stages in data preprocessing pass through outlier detection using formulas (1), (2), and (3). Outlier detection ensures that the data used is not invalid. KMeans, DBSCAN, and Hierarchical go through this process.

$$IQR = Q_3 - Q_1 \quad (1)$$

$$Lower\ Bound = Q_1 - 1.5\ x\ IQR \quad (2)$$

$$Upper\ Bound = Q_3 + 1.5\ x\ IQR \quad (3)$$

Outlier detection in tourism clusters helps identify unusual visitor numbers, spending, or business performance patterns. It can reveal why specific destinations get more or fewer visitors than expected,

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

whether due to seasonal changes, promotions, or trends. It also helps improve service quality by spotting businesses with very high or low customer satisfaction.

2.2.2 Normalization

Normalization is a crucial preprocessing step in data analysis that ensures all features contribute equally to clustering algorithms. It adjusts data to a standard scale, improving the accuracy of distance-based methods like K-Means, DBSCAN, and Hierarchical Clustering. By transforming data to have a mean (μ) of 0 and a standard deviation (σ) of 1, normalization helps prevent features with larger values from dominating the clustering process. At this stage, normalization follows Equation (4), where:

$$Z = \frac{x-\mu}{\sigma} \quad (4)$$

After outlier detection and normalization, each clustering algorithm proceeds with its process. K-Means initializes centroids, assigns data points based on Euclidean distance, and updates centroids iteratively. DBSCAN identifies dense regions using a radius (ε) and minimum points, forming clusters while marking outliers as noise. Hierarchical clustering builds a dendrogram by merging or splitting clusters step by step, allowing flexible cluster selection. Each method provides unique insights, with K-Means suitable for well-separated clusters, DBSCAN handling noise and varying shapes, and Hierarchical clustering offering a tree-like structure for analysis.

2.2.3 K-Means

The WCSS method determines the optimal number of clusters (k) by identifying the "elbow point" in the WCSS graph, where adding more clusters provides diminishing variance reduction. The method to determine the best k is in equation (5).

$$WCSS = \sum_{i=1}^{k} \sum_{x \epsilon c_i} \|x - \mu_i\|^2 \quad (5)$$

Next, Initialize the initial centroid randomly. Assign each point to the nearest centroid as entered in equation (6). Initial centroids are randomly selected, and each data point is assigned to the nearest centroid based on the squared Euclidean distance, ensuring minimal intra-cluster variance.

$$C_i = \mathrm{x} : \|x - \mu_i\|^2 \leq \left\|x - \mu_j\right\|^2, \forall_j \neq i \quad (6)$$

Finally, the centroid is updated using equation (7) before evaluation. The centroid of each cluster is updated by averaging the positions of all assigned data points, refining cluster boundaries iteratively for better grouping.

$$\mu_i = \frac{1}{C_i} \sum_{x \epsilon c_i} x \quad (7)$$

The clustering quality is assessed using computational efficiency and scalability tests to ensure the method's effectiveness in handling large datasets.

2.2.4 DBSCAN

After outlier detection and normalization, DBSCAN determines the optimal clustering by defining core, border, and noise points based on a distance threshold (ε) and minimum neighbors (q). Equation (8) calculates the neighborhood density to identify clusters, using Euclidean distance for point similarity. The method is then evaluated for computational efficiency and scalability.

$$N_{(p)} = q \; \epsilon \; D \mid \; d_{(p,q)} \leq \varepsilon \quad (8)$$

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

2.2.5 Hierarchical Clustering

Hierarchical clustering is applied to determine the optimal clustering structure by iteratively merging or splitting clusters based on similarity. The best clustering result is evaluated using a specific criterion formulated in Equation (9), which measures the cluster quality to ensure meaningful data grouping.

$$d(C_i, C_j) = \sum_{x \epsilon c_i \cup c_j} \left\| x - \bar{x}_{c_i \cup c_j} \right\|^2 - \sum_{x \epsilon c_i} \left\| x - \bar{x}_{c_i} \right\|^2 - \sum_{x \epsilon c_j} \left\| x - \bar{x}_{c_j} \right\|^2 \quad (9)$$

After preprocessing, three clustering methodologies are implemented: K-Means, DBSCAN, and Hierarchical Clustering. Every algorithm undergoes hyperparameter optimization to improve its efficacy: the ideal number of clusters (k) for K-Means is determined using the Elbow Method and Silhouette Analysis; the epsilon (ε) and minimum points (MinPts) parameters for DBSCAN are adjusted through k-distance graphs; and the most appropriate linkage criterion (single, complete, or average) for Hierarchical Clustering is ascertained through dendrogram analysis [31]. Following parameter optimization, clustering models are developed using the preprocessed data, and the clustering results are analyzed to identify trends in SME distribution. The choice of hyperparameters profoundly affects clustering results. In K-Means, an unsuitable selection of k may result in overfitting due to excessive clusters or underfitting from insufficient clusters, thus impairing the model's generalization capability [32]. In DBSCAN, inappropriate ε and MinPts values can lead to excessive noise if ε is excessively large or MinPts is insufficiently tiny [32].

In contrast, highly fragmented clusters may arise if ε is excessively small or MinPts is excessively large. In Hierarchical Clustering, the linkage criterion influences the morphology and dimensions of clusters; for instance, a single linkage may yield extended clusters, whereas complete linkage typically results in more compact formations [28]. Consequently, precisely adjusting these hyperparameters is crucial for effective clustering, guaranteeing that the recognized patterns appropriately represent the underlying data structure.

2.3. Clustering Evaluation and Analysis

Two assessment metrics are employed to compare the performance of the clustering models: The Silhouette Coefficient assesses clustering quality by measuring cluster cohesion and separation, while the Davies-Bouldin Index examines cluster compactness and separation, with a lower value indicating superior clustering [33]. These metrics are used to efficiently evaluate intra-cluster similarity and inter-cluster separation, ensuring robustness in the assessment of SME groupings. Moreover, data visualization methods, including scatter plots, dendrograms, and heatmaps, are utilized to enhance the interpretation of clustering outcomes, hence fostering an intuitive comprehension of clustering frameworks and yielding significant insights for decision-making [33].

This study evaluates clustering performance, conducts computational efficiency tests, and analyzes strong and weak scalability. Computational efficiency is evaluated based on execution time and memory consumption, confirming that the selected clustering algorithms are viable for real-world applications, especially in extensive tourism data processing [28]. Robust scalability testing assesses the extent to which an algorithm diminishes execution time with the increased utilization of processing cores, demonstrating the efficacy of parallel computing in clustering tasks [29]. Weak scalability testing evaluates the algorithm's capacity to sustain consistent execution time as dataset size expands with computer resources, confirming its relevance for increasing tourism datasets [30]. These assessments offer an in-depth comprehension of the advantages and drawbacks of each clustering technique, facilitating the choice of the most appropriate algorithm based on efficacy, computational expense, and scalability factors [28], [29], [30].

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

In addition to assessing clustering quality, tests for computational efficiency are performed to measure execution time and memory usage, thereby verifying the feasibility of any clustering algorithm [34]. Moreover, robust scalability testing assesses the reduction in execution time as additional processing cores are employed, demonstrating the efficacy of parallelization. Weak scalability testing evaluates if execution time is consistent as the dataset size increases about computational resources [35]. These assessments offer a thorough examination of clustering efficacy, guaranteeing the selection of the most efficient and scalable approach for enhancing strategic decision-making in the tourism sector.

This methodology guarantees the accuracy and significance of results in analyzing SME distributions within the tourism sector by methodically tackling preprocessing, clustering optimization, and evaluation. Incorporating comprehensive preprocessing procedures and a substantiated evaluation improves the dependability and clarity of clustering results.

2.3.1 Evaluation of effeciency

The Silhouette Score assesses clustering quality by comparing the average intra-cluster distance $a_{(i)}$ and the nearest-cluster distance $b_{(i)}$, as defined in Equation (10). A score of 1 indicates well-separated and compact clusters, ensuring high clustering performance.

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{max_{(a_{(i)}, b_{(i)})}} \tag{10}$$

The DBI, defined in Equation (11), evaluates clustering by measuring intra-cluster similarity (s_i) and inter-cluster separation (d_{ij}). Lower DBI values indicate better clustering with well-defined and distinct cluster structures.

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{11}$$

2.3.2 Computanional Effiency Test (Memory & Time)

Equation (12) measures each clustering algorithm's total memory consumption by summing all elements' memory usage (m_i). This assessment ensures the algorithm's feasibility for resource-constrained environments.

$$M = \sum_{i=1}^{n} m_i \tag{12}$$

Equation (13) calculates the total runtime from the start to the completion of each clustering algorithm. This evaluation helps determine computational efficiency and scalability for different dataset sizes.

$$T_{exec} = T_{end} - T_{start} \tag{13}$$

2.3.3 Evaluation of effeciency

Scalability testing assesses how well a clustering algorithm performs as computational resources increase, like equation (14). It is categorized into strong and weak scalability tests. Strong scalability (S_s) measures how execution time (T_n) decreases when more processors (p) are added, assuming a fixed problem size. Higher values indicate efficient parallelization.

$$S_s = \frac{T_n}{T_p} \tag{14}$$

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Weak scalability (S) evaluates performance when the number of processors (p') and workload increase. It ensures that execution time (T_p) remains stable as computational demand grows.

$$S_x = \frac{T_p}{T_{p'}} \quad (15)$$

3. RESULT

K-Means, DBSCAN, and Hierarchical Clustering methods are employed for normalization and outlier detection. The Elbow Method determines the ideal number of clusters for K-Means, represented by scatter plots. DBSCAN identifies dense clusters and outliers, augmented by dimensionality reduction by PCA. Hierarchical Clustering is examined via a dendrogram, illustrating the hierarchical relationships among data points. These methodologies are assessed to determine their efficacy in revealing patterns and anomalies, offering significant insights for decision-making in Small and Medium Enterprises (SMEs).

The visualization of Figure 2 illustrates the distribution of various variables or groups, revealing consistent patterns over most of the data. The interquartile ranges (IQRs), depicted by the boxes and whiskers that include most data, signify consistency among the variables. The medians, represented as horizontal lines within the boxes, exhibit relative stability, indicating that the core tendencies of the groups are comparable. Nonetheless, certain outliers exist, as evidenced by dots located beyond the whiskers, signifying data points that markedly diverge from the primary distribution [36], [37].

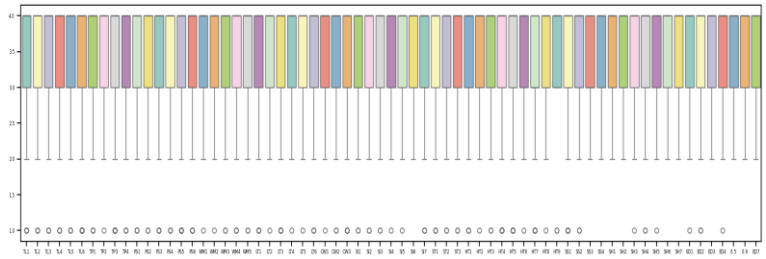

Figure 2. Interquartile Range Prior to Normalization

These outliers may indicate abnormalities, measurement inaccuracies, or significant variances necessitating additional examination. The labels at the bottom, including "TL1," "TP1," and "PS1," denote specific variables or samples, facilitating the identification of groups with distinct traits or atypical behaviors. The graphic underscores a general homogeneity in the data while identifying specific groupings that warrant further scrutiny due to their outliers or anomalies.

Figure. 3 illustrates the distribution of multiple variables (TL1, TL2, ..., ED7) through boxplot visualizations. Each boxplot depicts each variable's median, interquartile range (IQR), whiskers, and outliers. Most variables display approximately symmetric distributions, with medians centrally within the range. However, many variables exhibit skewed distributions.

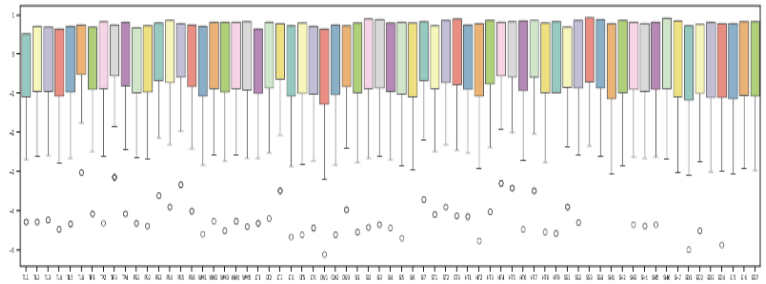

Figure 3. Interquartile Range Post-Normalization

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Outliers, shown by points beyond the whiskers, are evident in almost all variables, with specific variables like CAV1, HT7, and ED1 exhibiting more outliers than others. The placement of outliers distant from the whiskers signifies considerable extreme values in the sample. The distribution range differs among variables, with some, such as CAV1, exhibiting a wider range than others. The variations and outliers indicate significant disparities in the data distribution among variables, necessitating further research based on the dataset's context.

Figure 4 displays a collection of boxplots depicting the data distribution for multiple variables (TL1, TL2, ..., ED7). Each boxplot illustrates the median, interquartile range (IQR), whiskers, and possible outliers. The variables display stable distributions, with medians often located in the middle of their respective boxes. Nonetheless, outliers exist across several variables, evident as points beyond the whiskers, signifying extreme data values [38], [39].

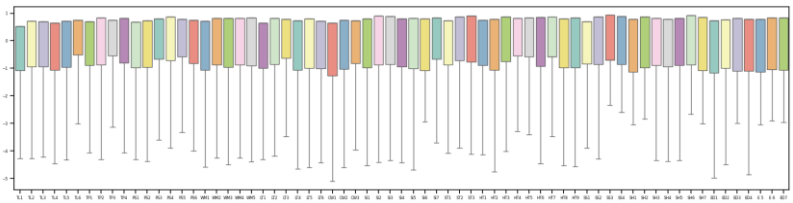

Figure 4. Distribution of various variables following outlier removal

The whiskers exhibit varying lengths among variables, indicating data dispersion or variability disparities. CAV1 and ED1 exhibit a wider range and more extreme values than other variables. This image indicates that although the dataset exhibits a relatively steady central tendency across variables, several variables necessitate more scrutiny due to their extensive ranges and notable outliers. Additional investigation is required to ascertain the context and ramifications of these differences.

Normalization and outlier identification are essential for enhancing clustering precision and stability. Before normalization, scale discrepancies across variables may skew clustering techniques such as K-Means, rendering them more susceptible to high-magnitude features. Post-normalization, the data distribution attains more uniformity, guaranteeing that each variable contributes equitably to the clustering process. Nonetheless, outliers continue to present an issue, potentially skewing cluster placements. Utilizing outlier detection methods, such as Interquartile Range (IQR) filtering or DBSCAN, facilitates the identification and elimination of extreme data, resulting in more uniform and representative clusters. This enhancement improves the clarity of clustering outcomes, especially in strategic decision-making for the tourism sector and small to medium-sized firms (SMEs). Enhanced data quality facilitates precise insights, diminishes noise, and augments pattern identification. Future research may investigate other normalization strategies, such as log transformation or robust scaling, in conjunction with machine learning-based outlier identification techniques to enhance clustering efficacy and decision-making results.

3.1. K-Means

Applying the K-Means clustering algorithm to our dataset entailed determining the ideal number of clusters via the Elbow Method and displaying the resultant clusters through a scatter plot. The Elbow Method was utilized to ascertain the optimal number of clusters by graphing the Within-Cluster Sum of Squares (WCSS) against different cluster quantities [40]. This method determines the "elbow point," where the reduction in WCSS markedly diminishes, achieving a balance between cluster compactness and simplicity. The research revealed that the optimal number of clusters is two, indicating the inherent grouping within the data. The scatter plot offers a clear visual depiction of the clustering outcomes,

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

emphasizing the distribution of data points inside each cluster and illustrating K-Means' efficacy in aggregating similar data points.

Figure 5 depicts the Elbow Method, demonstrating the correlation between the number of clusters and the Within-Cluster Sum of Squares (WCSS). As the number of clusters rises, the Within-Cluster Sum of Squares (WCSS) diminishes owing to a decrease in intra-cluster variance [40]. Nonetheless, beyond a specific threshold, the decrease becomes negligible, creating an "elbow" that signifies the optimal quantity of clusters. The elbow is evident in two clusters, indicating this is the ideal selection. Exceeding two clusters does not markedly enhance clustering compactness, whereas reducing clusters would lead to an oversimplification of the data structure [41].

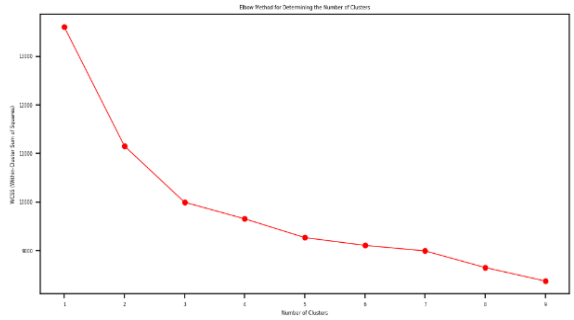

Figure 5. Within-Cluster Sum of Squares

Figure 5's Elbow Method results demonstrate that selecting two clusters strikes a balance between reducing intra-cluster variance and preventing superfluous complexity. This discovery corresponds with the dataset's characteristics, wherein a distinct separation is seen between two principal categories. The significant reduction in WCSS up to two clusters indicates that further clusters would yield only negligible enhancements while augmenting computing complexity. This conclusion aligns with prior research [40], highlighting the necessity of determining an ideal number of clusters by balancing variance reduction and over-segmentation. This concept is essential for practical applications, especially in SME segmentation, where delineating significant groupings without undue fragmentation can result in more effective tactics.

The Figure 6 shows the clustering outcomes produced by the K-Means technique. The data points are allocated in two dimensions (Dim 1 and Dim 2) and are categorized into two clusters, indicated by red (Cluster 0) and blue (Cluster 1). Cluster 0 is located in the left and central regions, whereas Cluster 1 is primarily on the plot's right side. Despite the broad separation of the clusters, overlapping regions exist where the delineation between them is ambiguous, highlighting the algorithm's limitations in differentiation.

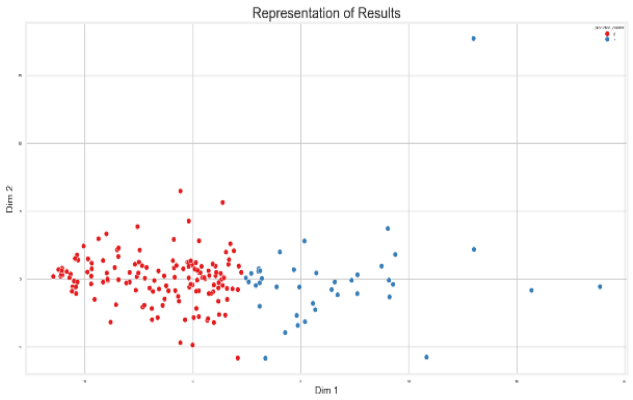

Figure 6. K-Means Clustering Outcome

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Cluster 0 has a more concentrated aggregation of points, indicating superior compactness, whereas Cluster 1 seems more scattered, with points distributed at greater distances. The variability in Cluster 1 may suggest the existence of noise or outliers within the data. K-Means presumes spherical clusters, which may hinder its efficacy in managing irregularly shaped clusters or fluctuating densities, thus accounting for the less obvious separation evident in this graphic.

Multiple modifications could enhance the clustering outcomes. Dimensionality reduction methods, like Principal Component Analysis (PCA), can streamline the data structure and enhance cluster delineation [42], [43]. Furthermore, evaluating alternate algorithms such as DBSCAN, which is more adept at managing clusters with diverse densities and noise, may produce more precise outcomes. Ultimately, reassessing the selection of k (the number of clusters) using techniques such as the elbow approach or silhouette analysis may ascertain if two clusters are optimum for this dataset.

3.2. DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a robust clustering algorithm that identifies clusters of varying shapes and sizes while effectively detecting outliers [44], [45]. Unlike centroid-based methods, DBSCAN does not require specifying the number of clusters in advance and instead relies on two key parameters: the minimum number of points (MinPts) required to form a dense region and the maximum distance (Epsilon) within which points are considered part of the same cluster. This flexibility makes DBSCAN particularly useful for datasets with irregular distributions or noise, as it can separate dense regions from sparse areas [15], [46]. When combined with dimensionality reduction techniques such as Principal Component Analysis (PCA), DBSCAN can efficiently handle high-dimensional data, providing valuable insights into the natural groupings and anomalies within complex datasets [47], such as those related to Small and Medium Enterprises (SMEs).

Figure 7 shows clustering results using the DBSCAN algorithm with data dimensionally reduced through PCA. Most data forms a dense main cluster, represented by yellow points. Meanwhile, one point separated far from the main cluster, shown in purple, is identified as an outlier or noise by the DBSCAN algorithm [46], [47]. These outliers indicate that most data have a relatively homogeneous structure, while the outlier represents significantly different or unique data compared to the majority.

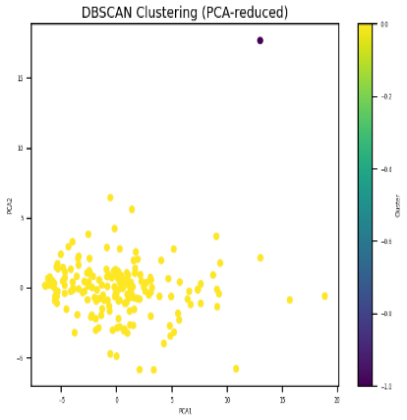

Figure 7. DBSCAN Cluster Result

This outlier may be caused by measurement errors, extreme variations, or unique specific conditions, requiring further analysis to determine whether the data is valid or should be excluded from the analysis [48]. These results suggest that the pattern in the data is relatively simple, with one dominant group, which may reflect the general characteristics of the analyzed data [42]. In the context of research on Small and Medium Enterprises (SMEs), the central cluster could represent the general patterns of the

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

industry. At the same time, the outlier might indicate an anomaly or unique characteristic of a specific unit [46], [47].

3.3. Hierarchical Clustering

This is an example of the use of sub-chapters in a paper. Sub-chapters are allowed to be included in all chapters, except in the conclusion. Hierarchical Clustering is a powerful unsupervised machine-learning technique used to group data points based on their similarity without requiring a predefined number of clusters [23]. Using metrics such as Euclidean distance to measure proximity and linkage methods to define how clusters are joined, hierarchical Clustering provides a visual and systematic approach to understanding the relationships and patterns within a dataset. This technique is particularly advantageous in exploratory data analysis, where the goal is to uncover natural groupings in data, making it highly relevant for applications like analyzing behavioral patterns or identifying distinct subgroups in complex datasets [49], such as those found in Small and Medium Enterprises (SMEs).

Figure 8 illustrates the dendrogram from the hierarchical clustering analysis, illustrating the grouping process of data based the dendrogram from the hierarchical clustering analysis, illustrating the grouping process of data based on their similarity or proximity, using Euclidean distance as the metric. The horizontal axis represents the individual samples, while the vertical axis indicates the distances at which clusters are merged. The dendrogram reveals three main clusters at a higher level of the hierarchy: the orange cluster (samples 9, 2, 17, 1, 8, 7, 5, 3, 15), the green cluster (samples 4, 13, 12, 19), and the red cluster (samples 14, 6, 11, 18, 10, 0, 16).

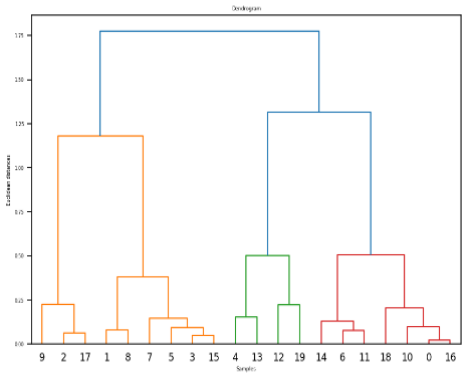

Figure 8. Interquartile Hierarchical Clustering

Within each cluster, the data points exhibit substantial similarity, as indicated by smaller vertical distances, while the separation between clusters—marked by greater vertical distances—suggests significant differences. To identify the optimal number of clusters, the dendrogram could be cut at approximately 1.25, yielding three distinct clusters. This hierarchical structure provides insights into the relationships among the data, where closer groupings imply more substantial similarities [49], [50]. Further analysis of these clusters could uncover unique characteristics or patterns within the data, which is particularly useful in research contexts such as small and medium enterprises (SMEs) for understanding specific group behaviors or anomalies.

3.4. Evaluation Silhoutte Coefficient and davies Bouldin Index

The evaluation results from the three clustering methods (K-Means, DBSCAN, and Hierarchical Clustering) exhibit varying effectiveness in dividing data into two categories, as shown in Table 3. DBSCAN achieved the highest Silhouette Coefficient (0.5496) and the lowest Davies-Bouldin Index (0.3298), indicating that the clusters produced by DBSCAN are more distinctly separated and more compact than those formed by the other two methods. The results demonstrate that DBSCAN excels at

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

defining the data structure, particularly in datasets characterized by changing densities and noise, which corresponds with the algorithm's strengths [17], [45], [51].

In contrast, Hierarchical Clustering had the lowest Silhouette Coefficient (0.2662) and a Davies-Bouldin Index of 1.4886, indicating that its clustering quality was subpar relative to DBSCAN while somewhat superior to K-Means. This outcome suggests that although Hierarchical Clustering offers significant insights into cluster interactions via dendrograms, its efficacy may be compromised by data patterns that do not conform to distance-based approaches like the Ward method [23].

Table 3. Assessment Outcomes K-Means, DBSCAN, Hierarchical

Clustering Type	n_Clusters	Silhouette Coefficient	Davies-Bouldin Index
K-Means	2	0.2321	1.6754
DBSCAN	2	0.5496	0.3298
Hierarchical	2	0.2662	1.4886

K-Means demonstrated a Silhouette Coefficient of 0.2321 and the highest Davies-Bouldin Index of 1.6754, signifying the poorest clustering effectiveness. This result is due to K-Means' presumption of spherical cluster formations, which may be inappropriate for datasets exhibiting complex or irregular distributions [15], [17], [21].

DBSCAN proved to be the superior technique for this dataset, with value hyperparameter ε = 15.88 and MinPts = 68, owing to its capacity to manage clusters with heterogeneous densities and proficiently detect noise [42], [45], [48]. Nevertheless, choosing the best suitable algorithm is contingent upon the particular analytical goals and the properties of the dataset [52], [53]. Hierarchical Clustering can yield further insights if a hierarchical interpretation of the cluster structure is necessary despite its inferior assessment metrics [23], [54]. Conversely, if the dataset has well-delineated spherical clusters, K-Means may remain a feasible choice despite its inferior clustering quality. Subsequent investigations may examine parameter optimization or manipulate cluster amounts to enhance clustering outcomes [14], [18], [19], [48], [55], [56].

3.5. Efficiency and Scalability

Computational efficiency is crucial in clustering methods, particularly when managing extensive data sets. Algorithm efficiency is often assessed based on two primary criteria: execution time and memory consumption [57], [58]. The execution time indicates the speed at which the algorithm processes data and generates clusters, while memory usage denotes the computational resources necessary for data storage and manipulation. Various clustering techniques demonstrate differing efficiencies based on their underlying mechanisms, including iterative optimization, density estimation, or hierarchical merging. Understanding these trade-offs is essential for selecting the most appropriate clustering technique according to dataset attributes and computational limitations. Table 4 presents the findings of computational efficiency.

Table 4. Results of Computational Efficiency

Clustering Type	Execution Time (s)	Memory Usage (MB)
K-Means	17.90	3.84
DBSCAN	11.33	2.38
Hierarchical	13.74	3.71

The performance analysis indicates that the K-Means clustering algorithm exhibits the longest execution time of 17.90 seconds while maintaining a moderate memory usage of 3.84 MB. This is

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

attributed to its iterative nature, where cluster centers are repeatedly refined until convergence. The iterative updates contribute to its computational intensity.

Conversely, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) demonstrates a reduced execution time of 11.33 seconds, with the lowest memory consumption of 2.38 MB. This efficiency stems from its density-based approach, which involves computing distances among data points to determine local density. DBSCAN is particularly effective for datasets with varying densities and outliers, making it a robust option despite its dependency on distance computations.

Hierarchical clustering, on the other hand, exhibits an intermediate execution time of 13.74 seconds and a memory usage of 3.71 MB. This method constructs a hierarchical tree structure through iterative merging or division of clusters. Although hierarchical clustering does not require iterative updates like K-Means or extensive distance calculations as seen in DBSCAN, its computational complexity increases with larger datasets.

The selection of a clustering algorithm must consider the balance between execution time and memory efficiency. Hierarchical clustering offers moderate computational efficiency but may not scale effectively for extensive datasets due to its quadratic time complexity. DBSCAN is advantageous for datasets with noise and varying densities, requiring minimal memory [57], [58], [59], [60]. Despite its computational demands, K-Means remains a widely used method for structured datasets with clearly defined clusters. The choice of an appropriate clustering technique depends on the dataset's characteristics, available computational resources, and specific analytical requirements.

Strong scalability is a crucial principle in parallel computing that assesses the efficiency with which an algorithm decreases execution time as the number of processor cores increases while keeping the problem size constant. Like shown in Table 5, an algorithm with optimal scalability should demonstrate a proportional reduction in execution time as additional cores are employed. However, factors such as synchronization overhead, communication latency, and algorithmic constraints often impede linear improvements in practical applications. Evaluating the strong scalability of clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering provides valuable insights into their computational efficiency in multi-core environments.

Table 5. Results of the Strong Scalability Test

Core	Execution Time (s)		
	KMeans	DBSCAN	Hierarichal
1	2.198334	0.058153	0.949489
2	3.776264	3.123195	3.391078
4	3.664294	3.207053	2.979006
8	5.361955	3.552478	3.700688

Examining the execution times across 1, 2, 4, and 8 cores reveals distinct scalability trends among the clustering algorithms. K-Means exhibits non-linear scalability, with execution time increasing from 2.198 seconds on a single core to 3.776 seconds on two cores, slightly decreasing to 3.664 seconds on four cores, and then rising to 5.362 seconds on eight cores. This irregular pattern suggests that after a certain threshold, the benefits of parallelization are offset by inter-core communication overhead and synchronization costs. The iterative nature of K-Means, which involves repeated centroid updates and global adjustments, makes it susceptible to diminishing returns as more cores are utilized, potentially leading to performance degradation.

In contrast, DBSCAN displays a significant increase in execution time as additional cores are introduced. The execution time starts at 0.058 seconds on a single core and sharply rises to 3.123 seconds on two cores, further increasing to 3.207 seconds on four cores, and reaching 3.552 seconds on eight cores. This pattern suggests that DBSCAN does not substantially benefit from parallel execution. Since

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

its primary operations—neighborhood searches and density-based region expansion—are inherently sequential, increasing the number of cores does not efficiently distribute the computational workload, resulting in minimal performance gains.

Meanwhile, Hierarchical Clustering shows moderate improvements in execution time with increasing core counts, though scalability remains limited. The execution time starts at 0.949 seconds on one core, increases to 3.391 seconds on two cores, decreases slightly to 2.979 seconds on four cores, and then rises to 3.701 seconds on eight cores. This trend indicates that hierarchical clustering can leverage parallelization to some extent, but its efficiency gains plateau due to the computational complexity of pairwise distance calculations and dendrogram construction, which are difficult to distribute evenly across multiple cores.

Overall, K-Means clustering demonstrates limited scalability beyond four cores, as execution time increases due to coordination overhead. DBSCAN experiences significant performance degradation with additional cores, implying that its algorithmic structure is not conducive to parallel execution. Hierarchical clustering achieves marginal efficiency improvements but reaches a saturation point where further core additions provide negligible benefits. These findings highlight that while multi-core architectures can enhance clustering performance, their effectiveness depends on the algorithm's internal structure, workload distribution capabilities, and the trade-off between parallel execution benefits and communication overhead.

Like shown in Table 6, weak scalability is a fundamental concept in parallel computing that assesses an algorithm's ability to manage increasing data volumes while maintaining a consistent execution time when computational resources are proportionally increased [61], [62], [63]. Ideally, as the dataset size doubles with a corresponding augmentation in computing capacity, the execution time should remain stable. This parameter is particularly significant in data science and machine learning, where efficiently handling large datasets is crucial for real-time analytics and decision-making. Unlike strong scalability, which focuses on speed improvement at a constant problem size, weak scalability evaluates an algorithm’s ability to accommodate growing data without incurring exponential computational costs.

Table 6. Results of the Weak Scalability Test

Data size	Execution Time (s)		
	KMeans	DBSCAN	Hierarichal
206	0.518493	1.053068	1.250144
406	0.562756	0.032630	0.367015
812	0.523547	0.015667	0.345609
1624	0.517544	0.015692	0.363602

The weak scalability test results for K-Means, DBSCAN, and Hierarchical Clustering exhibit distinct computational efficiency patterns as the dataset size increases from 206 to 1,624 observations. K-Means demonstrates excellent weak scalability, with execution time remaining nearly constant across different dataset sizes. The execution time starts at 0.5185 seconds for 206 observations and remains stable at 0.5175 seconds for 1,624 observations. This consistency suggests that K-Means efficiently utilizes computational resources, benefiting from parallelization and optimized centroid updates [21], [64]. Minor fluctuations in execution time can be attributed to factors such as centroid initialization and convergence dynamics. Due to its scalability, K-Means is highly suitable for large-scale clustering tasks where processing efficiency is a priority.

DBSCAN exhibits a sharp decline in execution time as the dataset size increases. Initially, it records a high execution time of 1.0531 seconds for 206 observations, but as the dataset expands, the execution time drops significantly to 0.0157 seconds for 1,624 observations. This pattern suggests that

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

DBSCAN benefits from computational optimizations in neighborhood searches and density-based clustering processes. Despite its reliance on distance calculations, which can be computationally expensive, DBSCAN's improved performance with larger datasets suggests its feasibility for handling substantial data volumes, particularly in applications requiring density-based clustering [32], [51].

Hierarchical clustering shows a decreasing execution time trend, starting at 1.2501 seconds for 206 observations and reducing to 0.3636 seconds for 1,624 observations. This decrease may be due to computational optimizations that mitigate some of the expected quadratic or cubic complexity [23], [63]. However, hierarchical clustering remains less scalable than K-Means or DBSCAN due to its intensive pairwise distance calculations and cluster merging operations. Therefore, it is more suitable for smaller datasets where hierarchical relationships need to be preserved.

This study extends prior research by assessing clustering algorithm scalability using real-world datasets. The insights gained have practical implications across various industries, including tourism, arts, handicrafts, food and beverages, fashion, antiques, performing arts, and transportation. Unlike generic studies, this research tailors clustering strategies to specific industry needs, addressing gaps in previous studies. By bridging the gap between algorithmic performance and real-world applications, this study provides valuable insights for data-driven decision-making, reinforcing the importance of machine learning in strategic business planning.

4. DISCUSSIONS

This study explored the comparative effectiveness of K-Means, DBSCAN, and Hierarchical Clustering algorithms in identifying patterns and anomalies within datasets related to Small and Medium Enterprises (SMEs). The research outcomes provide multifaceted insights into the behavior of each algorithm and how their strengths and limitations align with practical applications.

DBSCAN consistently emerged as the most effective method, as evidenced by its superior Silhouette Coefficient (0.5496) and lowest Davies-Bouldin Index (0.3298). This evaluation suggests that DBSCAN generated the most well-defined clusters, even in noise and non-spherical data distributions. The results validate DBSCAN's ability to manage real-world data complexity, such as variations in density and irregular patterns commonly found in SME datasets. This algorithm aligns with previous studies by Ester et al. [44] and newer comparative works [46] highlighting DBSCAN's robustness in complex environments.

Conversely, K-Means demonstrated the weakest performance, with the highest Davies-Bouldin Index (1.6754) and the lowest Silhouette Coefficient (0.2321). This result is attributable to K-Means' assumption of spherical clusters and its sensitivity to outliers and scale variances. Despite its widespread use due to simplicity and computational efficiency in structured datasets [40], the findings reaffirm limitations when dealing with real-world heterogeneity. However, its stable execution times in weak scalability tests indicate potential for large-scale implementations when data is relatively clean and homogenous.

Though less clustering clustering quality, hierarchical clustering offers interpretability through dendrograms. It provides a detailed perspective of relationships within the data, particularly valuable in exploratory analyses. Its moderate performance and visualization strength make it suitable for identifying subgroup hierarchies within SME sectors. Nevertheless, its scalability remains a concern, as reflected in strong scalability tests that showed saturation with higher core usage.

The results underscore how normalization and outlier handling significantly influence clustering precision. Prior to normalization, disparities in feature magnitudes skewed the clustering output, especially for distance-based algorithms like K-Means. Post-normalization, however, the clustering performance improved across all methods. This Clustering supports earlier findings in the literature that emphasize the importance of preprocessing in unsupervised learning tasks [36], [37].

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

From a computational standpoint, DBSCAN again proved advantageous, showing lower memory usage and faster execution in single-core setups. However, its poor performance in multi-core environments—due to its inherently sequential operations—suggests that it may not benefit from parallelization as much as the other algorithms. Similar observations in prior works [57], [58] corroborate this. On the other hand, K-Means showed excellent weak scalability, suggesting suitability for larger datasets where speed and efficiency are critical.

Compared with previous studies, this research extends the existing literature by applying Clustering for data grouping and evaluating algorithmic efficiency and scalability in real-world SME datasets. Prior studies often focus on theoretical or synthetic data [21]. In contrast, this research contributes practical value through domain-specific applications, including sectors like tourism, fashion, food and beverages, and the creative economy. The study addresses a critical gap in data-driven SME analysis by tailoring algorithmic insights to these industries.

The findings suggest no universally best clustering algorithm; instead, the optimal method depends on data characteristics and operational goals. Hierarchical Clustering is recommended for interpretability. DBSCAN is preferred for robust Clustering in noisy and variable datasets. K-Means may still be suitable for scenarios requiring high scalability and efficiency despite lower clustering quality. Future research should explore hybrid clustering techniques, advanced preprocessing, and optimization of clustering parameters to further enhance accuracy and applicability in complex datasets.

The segmentation of SMEs in Rembang Regency can inform regional planning, resource allocation, and targeted policy interventions, particularly in tourism development zones. For example, identifying clusters of SMEs with similar operational challenges or market orientation can help local governments design customized capacity-building programs, infrastructure support, or promotional strategies. These insights empower policymakers to shift from generic SME development policies to more evidence-based, cluster-specific interventions, thereby enhancing the effectiveness of public investment and fostering equitable regional growth.

5. CONCLUSION

A purposive sample strategy was employed based on data collected via a questionnaire administered to SMEs in Rembang Regency, Central Java. Participants were selected based on geographical location, workforce size, and SME classification. Before disseminating the questionnaire, preliminary research was conducted with academic experts in tourism, SME coordinators, and tourism coordinators to evaluate the relevance and clarity of the questions. The questionnaire featured a 6-point Likert scale without a neutral option, necessitating responders to provide a conclusive answer from “strongly disagree” to “strongly agree.” A sample size of 203 out of 219 respondents (92.69%) met the validity criteria for data analysis, reflecting the proportions of different types of SMEs in the region. This study assesses and contrasts the efficacy of K-Means, DBSCAN, and Hierarchical Clustering algorithms in discerning strategic methodologies for various tourism sectors. The results demonstrate that each clustering method possesses unique benefits and drawbacks for clustering quality, computational efficiency, and scalability.

This study evaluates the efficiency, clustering quality, and scalability of K-Means, DBSCAN, and Hierarchical Clustering for tourism-related SMEs in Rembang Regency, Central Java. The findings highlight that each algorithm has distinct strengths and weaknesses, making them suitable for different applications.

Regarding clustering quality, DBSCAN demonstrated the highest performance with a Silhouette Coefficient of 0.5496 and a Davies-Bouldin Index of 0.3298, indicating its effectiveness in handling datasets with varying densities and noise. Hierarchical clustering showed moderate performance, while

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

K-Means had the lowest Silhouette Coefficient of 0.2321, suggesting limitations in capturing complex data structures.

Regarding computational efficiency, the results indicate that DBSCAN had the shortest execution time of 11.33 seconds and the lowest memory usage of 2.38 MB, making it the most efficient algorithm. Hierarchical clustering followed with an execution time of 13.74 seconds and memory usage of 3.71 MB. K-Means had the longest execution time of 17.90 seconds and a memory consumption of 3.84 MB, primarily due to its iterative nature.

Scalability assessments revealed varying performance among the algorithms. Regarding strong scalability, K-Means showed limited efficiency beyond four cores, increasing execution time due to inter-core communication overhead. DBSCAN experienced significant performance degradation, as its execution time rose substantially with additional cores, indicating poor scalability in multi-core environments. Hierarchical clustering demonstrated marginal improvements but reached a saturation point where additional cores provided negligible benefits. In weak scalability tests, K-Means displayed the most stable execution times as dataset sizes increased, making it the most scalable algorithm for large datasets. DBSCAN showed decreasing execution times with larger datasets, suggesting improved computational efficiency. Hierarchical clustering exhibited a declining execution time trend but remained less scalable due to its high computational complexity.

DBSCAN is the most effective technique for clustering tourism-related SMEs, particularly for datasets with irregular patterns such as seasonal demand or niche markets. Hierarchical clustering provides valuable insights into hierarchical relationships among SMEs, making it suitable for market segmentation studies. Although K-Means is less effective in handling complex structures, it remains a practical choice for structured datasets with well-defined clusters due to its scalability. These insights help optimize business strategies, resource allocation, and market segmentation in the tourism industry. Future research should explore hybrid clustering techniques or integrate predictive analytics to enhance decision-making processes within tourism and other sectors.

This paper connects clustering algorithm efficacy with practical applications in tourism, providing a strategic framework for stakeholders. By customizing clustering techniques to the distinct aspects of the tourism industry—such as arts, handicrafts, food and drinks, fashion, antiques, performing arts, and transportation—decision-makers can refine business plans, optimize resource allocation, and foster growth in the tourism sector. Future studies may investigate hybrid clustering methodologies or the integration of clustering with predictive analytics to enhance strategic planning within the tourism sector.

In addition to evaluating traditional clustering methods, future research should explore hybrid or adaptive clustering approaches that combine the strengths of multiple algorithms. For instance, integrating density-based techniques (such as DBSCAN) with hierarchical frameworks can enhance both flexibility and interpretability in analyzing complex SME datasets. Moreover, benchmarking the results against non-tourism datasets or SMEs in other sectors could uncover broader patterns and validate the robustness of the proposed methodology. This comparative lens not only enriches the analytical perspective but also opens opportunities to generalize findings across different domains of applied informatics.

ACKNOWLEDGEMENT

This research was supported by Telkom University under the project NO. 627/LIT06/PPM-LIT/2024. The authors would also like to extend their gratitude to the respondents from Kabupaten

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

Rembang for their valuable participation and insights, which greatly contributed to the completion of this study.

REFERENCES

[1] E. Aminullah, "Forecasting of technology innovation and economic growth in Indonesia," *Technol Forecast Soc Change*, vol. 202, May 2024, doi: 10.1016/j.techfore.2024.123333.

[2] S. P. Dhakal and S. P. Tjokro, "Tourism enterprises in Indonesia and the fourth industrial revolution – are they ready?," *Tourism Recreation Research*, vol. 49, no. 2, pp. 439–444, Mar. 2024, doi: 10.1080/02508281.2021.1996687.

[3] S. Utama, R. Yusfiarto, R. R. Pertiwi, and A. N. Khoirunnisa, "Intentional model of MSMEs growth: a tripod-based view and evidence from Indonesia," *Journal of Asia Business Studies*, vol. 18, no. 1, pp. 62–84, Jan. 2024, doi: 10.1108/JABS-08-2022-0291.

[4] F. Achmad and I. Inrawan Wiratmadja, "Driving Sustainable Performance in SMEs Through Frugal Innovation: The Nexus of Sustainable Leadership, Knowledge Management, and Dynamic Capabilities," *IEEE Access*, vol. 12, pp. 103329–103347, 2024, doi: 10.1109/ACCESS.2024.3433474.

[5] F. Achmad, Y. Prambudia, and A. A. Rumanti, "Improving Tourism Industry Performance through Support System Facilities and Stakeholders: The Role of Environmental Dynamism," *Sustainability (Switzerland)*, vol. 15, no. 5, Mar. 2023, doi: 10.3390/su15054103.

[6] P. Jafarzadeh, T. Vähämäki, P. Nevalainen, A. Tuomisto, and J. Heikkonen, "Supporting SME companies in mapping out AI potential: a finnish AI development case," *Journal of Technology Transfer*, 2024, doi: 10.1007/s10961-024-10122-5.

[7] Q. Liu, J. Gao, and S. Li, "The innovation model and upgrade path of digitalization driven tourism industry: Longitudinal case study of OCT," *Technol Forecast Soc Change*, vol. 200, Mar. 2024, doi: 10.1016/j.techfore.2023.123127.

[8] A. I. Ramaano, "The potential significance of geographic information systems (GISs) and remote sensing (RS) in sustainable tourism and decent community involvement in African-rural neighborhoods," *Journal of Electronic Business & Digital Economics*, vol. 3, no. 3, pp. 341–362, Oct. 2024, doi: 10.1108/JEBDE-03-2024-0006.

[9] A. Ramadhan, H. M. Jumhur, and F. A. Nur, "POLICY FORMULATION FOR ANTICIPATING THE IMPACT OF ACID RAIN ON PADDY PLANTS USING NORMATIVE JURIDICAL ANALYSIS," *INDONESIAN JOURNAL OF URBAN AND ENVIRONMENTAL TECHNOLOGY*, pp. 164–182, Jul. 2024, doi: 10.25105/urbanenvirotech.v7i2.19451.

[10] V. A. Sari and S. Tiwari, "The Geography of Human Capital: Insights from the Subnational Human Capital Index in Indonesia," *Soc Indic Res*, vol. 172, no. 2, pp. 673–702, Mar. 2024, doi: 10.1007/s11205-024-03322-x.

[11] F. Achmad and I. I. Wiratmadja, "Strategic advancements in tourism development in Indonesia: Assessing the impact of facilities and services using the PLS-SEM approach," *Journal Industrial Servicess is*, vol. 10, no. 1, 2024, doi: 10.62870/jiss.v10i1.24494.

[12] F. Achmad, Y. Prambudia, and A. A. Rumanti, "Sustainable Tourism Industry Development: A Collaborative Model of Open Innovation, Stakeholders, and Support System Facilities," *IEEE Access*, vol. 11, pp. 83343–83363, 2023, doi: 10.1109/ACCESS.2023.3301574.

[13] E. Azmi, R. A. Che Rose, A. Awang, and A. Abas, "Innovative and Competitive: A Systematic Literature Review on New Tourism Destinations and Products for Tourism Supply," Jan. 01, 2023, *MDPI*. doi: 10.3390/su15021187.

[14] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif Intell Rev*, vol. 56, no. 7, pp. 6439–6475, Jul. 2023, doi: 10.1007/s10462-022-10325-y.

[15] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf Sci (N Y)*, vol. 622, pp. 178–210, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.139.

[16] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc Sci Res*, vol. 110, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

[17] N. Trianasari and T. A. Permadi, "Analysis of Product Recommendation Models at Each Fixed Broadband Sales Location Using K-Means, DBSCAN, Hierarchical Clustering, SVM, RF, and ANN," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 636–652, May 2024, doi: 10.47738/jads.v5i2.210.

[18] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif Intell Rev*, vol. 56, no. 7, pp. 6439–6475, Jul. 2023, doi: 10.1007/s10462-022-10325-y.

[19] Mahnoor *et al.*, "A Review of Approaches for Rapid Data Clustering: Challenges, Opportunities and Future Directions," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3461798.

[20] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," Sep. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/sym15091679.

[21] H. Liu, J. Chen, J. Dy, and Y. Fu, "Transforming Complex Problems Into K-Means Solutions," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 7, pp. 9149–9168, Jul. 2023, doi: 10.1109/TPAMI.2023.3237667.

[22] M. S. Al-Batah, E. R. Al-Kwaldeh, M. A. Wahed, M. Alzyoud, and N. Al-Shanableh, "Enhancement over DBSCAN Satellite Spatial Data Clustering," *Journal of Electrical and Computer Engineering*, vol. 2024, 2024, doi: 10.1155/2024/2330624.

[23] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif Intell Rev*, vol. 56, no. 8, pp. 8219–8264, 2023, doi: 10.1007/s10462-022-10366-3.

[24] M. Mariani and R. Baggio, "Big data and analytics in hospitality and tourism: a systematic literature review," *International Journal of Contemporary Hospitality Management*, vol. 34, no. 1, pp. 231–278, Jan. 2022, doi: 10.1108/IJCHM-03-2021-0301.

[25] E. Aminullah, "Forecasting of technology innovation and economic growth in Indonesia," *Technol Forecast Soc Change*, vol. 202, May 2024, doi: 10.1016/j.techfore.2024.123333.

[26] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowl Inf Syst*, vol. 66, no. 3, pp. 1575–1637, 2024, doi: 10.1007/s10115-023-02010-5.

[27] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective," Sep. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/sym15091679.

[28] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng Appl Artif Intell*, vol. 110, p. 104743, 2022, doi: https://doi.org/10.1016/j.engappai.2022.104743.

[29] Z. Wu, J. Sun, Y. Zhang, Z. Wei, and J. Chanussot, "Recent Developments in Parallel and Distributed Computing for Remotely Sensed Big Data Processing," *Proceedings of the IEEE*, vol. 109, no. 8, pp. 1282–1305, 2021, doi: 10.1109/JPROC.2021.3087029.

[30] G. Richer, A. Pister, M. Abdelaal, J.-D. Fekete, M. Sedlmair, and D. Weiskopf, "Scalability in Visualization," *IEEE Trans Vis Comput Graph*, vol. 30, no. 7, pp. 3314–3330, 2024, doi: 10.1109/TVCG.2022.3231230.

[31] L. S. Ling and C. T. Weiling, "Enhancing Segmentation: A Comparative Study of Clustering Methods," *IEEE Access*, p. 1, 2025, doi: 10.1109/ACCESS.2025.3550339.

[32] O. Kulkarni and A. Burhanpurwala, "A Survey of Advancements in DBSCAN Clustering Algorithms for Big Data," in *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*, 2024, pp. 106–111. doi: 10.1109/PARC59193.2024.10486339.

[33] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," Jan. 30, 2025, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2025.e41953.

[34] Z. Yuan *et al.*, "Benchmarking spatial clustering methods with spatially resolved transcriptomics data," *Nat Methods*, vol. 21, no. 4, pp. 712–722, 2024, doi: 10.1038/s41592-024-02215-8.

[35] Z. Ma, Y. Xu, H. Xu, Z. Meng, L. Huang, and Y. Xue, "Adaptive Batch Size for Federated Learning in Resource-Constrained Edge Computing," *IEEE Trans Mob Comput*, vol. 22, no. 1, pp. 37–53, 2023, doi: 10.1109/TMC.2021.3075291.

[36] D. Muhr, M. Affenzeller, and J. Küng, "A Probabilistic Transformation of Distance-Based Outliers," *Mach Learn Knowl Extr*, vol. 5, no. 3, pp. 782–802, Sep. 2023, doi: 10.3390/make5030042.

[37] S. Bhattacharya, F. Kamper, and J. Beirlant, "Outlier detection based on extreme value theory and applications," *Scandinavian Journal of Statistics*, vol. 50, no. 3, pp. 1466–1502, Sep. 2023, doi: 10.1111/sjos.12665.

[38] B. Avanzi, M. Lavender, G. Taylor, and B. Wong, "Detection and treatment of outliers for multivariate robust loss reserving," *Annals of Actuarial Science*, vol. 18, no. 1, pp. 102–125, Mar. 2024, doi: 10.1017/S1748499523000155.

[39] A. Azizan, S. Anile, C. K. Nielsen, E. Paradis, and S. Devillard, "Population density and genetic diversity are positively correlated in wild felids globally," *Global Ecology and Biogeography*, vol. 32, no. 10, pp. 1858–1869, Oct. 2023, doi: 10.1111/geb.13727.

[40] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Analysis of an optimal cluster approach: a review paper," in *Journal of Physics: Conference Series*, Institute of Physics, 2023. doi: 10.1088/1742-6596/2421/1/012015.

[41] F. dos S. Silva, J. C. dos Reis, and M. S. Reis, "SERIEMA: A Framework to Enhance Clustering Stability, Compactness, and Separation by Fusing Multimodal Data," in *Natural Language Processing and Information Systems*, A. Rapp, L. Di Caro, F. Meziane, and V. Sugumaran, Eds., Cham: Springer Nature Switzerland, 2024, pp. 394–408.

[42] S. Zeng, T. Wang, W. Lin, Z. Chen, and R. Xiao, "A Patent Mining Approach to Accurately Identifying Innovative Industrial Clusters Based on the Multivariate DBSCAN Algorithm," *Systems*, vol. 12, no. 9, p. 321, Aug. 2024, doi: 10.3390/systems12090321.

[43] A. A. Bushra, D. Kim, Y. Kan, and G. Yi, "AutoSCAN: automatic detection of DBSCAN parameters and efficient clustering of data in overlapping density regions," *PeerJ Comput Sci*, vol. 10, 2024, doi: 10.7717/peerj-cs.1921.

[44] J. Peng and Y. Chen, "Density-based clustering with boundary samples verification," *Appl Soft Comput*, vol. 159, p. 111685, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111685.

[45] T. Z. Abdulhameed, S. A. Yousif, V. W. Samawi, and H. I. Al-Shaikhli, "SS-DBSCAN: Semi-Supervised Density-Based Spatial Clustering of Applications with Noise for Meaningful Clustering in Diverse Density Data," *IEEE Access*, pp. 1–1, Sep. 2024, doi: 10.1109/access.2024.3457587.

[46] C. Retiti Diop Emane *et al.*, "Anomaly Detection Based on GCNs and DBSCAN in a Large-Scale Graph," *Electronics (Switzerland)*, vol. 13, no. 13, Jul. 2024, doi: 10.3390/electronics13132625.

[47] N. Garg and P. Dwivedi, "A Novel Approach for Exploring Data-Driven Nutritional Insights Using Clustering and Dimensionality Reduction Techniques," *SN Comput Sci*, vol. 5, no. 8, Dec. 2024, doi: 10.1007/s42979-024-03397-w.

[48] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "Intelligent mapping of geochemical anomalies: Adaptation of DBSCAN and mean-shift clustering approaches," *J Geochem Explor*, vol. 258, p. 107393, 2024, doi: https://doi.org/10.1016/j.gexplo.2024.107393.

[49] G. Shamim and M. Rihan, "Exploratory Data Analytics and PCA-Based Dimensionality Reduction for Improvement in Smart Meter Data Clustering," *IETE J Res*, vol. 70, no. 4, pp. 4159–4168, Apr. 2024, doi: 10.1080/03772063.2023.2218317.

[50] G. Mischler, Y. A. Li, S. Bickel, A. D. Mehta, and N. Mesgarani, "Contextual feature extraction hierarchies converge in large language models and the brain," *Nat Mach Intell*, vol. 6, no. 12, pp. 1467–1477, 2024, doi: 10.1038/s42256-024-00925-4.

[51] N. Hanafi and H. Saadatfar, "A fast DBSCAN algorithm for big data based on efficient density calculation," *Expert Syst Appl*, vol. 203, Oct. 2022, doi: 10.1016/j.eswa.2022.117501.

[52] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602

[53] A. Ramadhan, I. Mendonça, M. Aritsugi, and I. Chandra, "Enhancing the Accuracy of Conductivity Parameters from Real-Time Rainwater Quality Measurements based on Internet of Things Utilizing Machine Learning," in *2024 10th International Conference on Wireless and Telematics (ICWT)*, 2024, pp. 1–6. doi: 10.1109/ICWT62080.2024.10674689.

[54] A. Ramadhan *et al.*, "Central Tendency Data Real-Time Acid Rain Measurement to Evaluate Tool's Performance Using Statistical Analysis," vol. 14, no. 4, 2024.

[55] P. Pellizzoni, A. Pietracaprina, and G. Pucci, "k-Center Clustering with Outliers in Sliding Windows," *Algorithms*, vol. 15, no. 2, Feb. 2022, doi: 10.3390/a15020052.

[56] A. Fahad *et al.*, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans Emerg Top Comput*, vol. 2, no. 3, pp. 267–279, Sep. 2014, doi: 10.1109/TETC.2014.2330519.

[57] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature," Jun. 01, 2021, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-020-05395-4.

[58] H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, and G. Modwel, "A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets," *Multimed Tools Appl*, vol. 81, no. 24, pp. 35001–35026, 2022, doi: 10.1007/s11042-021-10594-9.

[59] A. E. Ezugwu, A. K. Shukla, M. B. Agbaje, O. N. Oyelade, A. José-García, and J. O. Agushaka, "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature," *Neural Comput Appl*, vol. 33, no. 11, pp. 6247–6306, 2021, doi: 10.1007/s00521-020-05395-4.

[60] U. Fang, M. Li, J. Li, L. Gao, T. Jia, and Y. Zhang, "A Comprehensive Survey on Multi-View Clustering," *IEEE Trans Knowl Data Eng*, vol. 35, no. 12, pp. 12350–12368, 2023, doi: 10.1109/TKDE.2023.3270311.

[61] M. A. Mahdi, K. M. Hosny, and I. Elhenawy, "Scalable Clustering Algorithms for Big Data: A Review," *IEEE Access*, vol. 9, pp. 80015–80027, 2021, doi: 10.1109/ACCESS.2021.3084057.

[62] M. Sun *et al.*, "Scalable Multi-view Subspace Clustering with Unified Anchors," in *Proceedings of the 29th ACM International Conference on Multimedia*, in MM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 3528–3536. doi: 10.1145/3474085.3475516.

[63] N. Monath *et al.*, "Scalable Hierarchical Agglomerative Clustering," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, in KDD '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1245–1255. doi: 10.1145/3447548.3467404.

[64] H. Hu, J. Liu, X. Zhang, and M. Fang, "An Effective and Adaptable K-means Algorithm for Big Data Cluster Analysis," *Pattern Recognit*, vol. 139, p. 109404, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109404.

Vol. 6, No. 3, Juni 2025, Page. 1605-1630
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.3.4602