Clustering Madura tourism destinations using Fuzzy C-Means and Fuzzy C-Medoids with Xie–Beni optimization

Eza Rahmanita[1], *Yeni* Kustiyahningsih[1*], Mohammad Faizal Akbar[1], *Akhmad* Sulthon Rabbani[1], *Rifqy* Alifuddin Alfareza[1]

[1]Program study of Information System, Faculty of Engineering, Universitas Trunodjoyo, Madura, Indonesia

Abstract. Tourism is a strategic sector that plays a significant role in improving the regional economy, including Madura Island, which boasts diverse cultural and natural tourism destinations. Proper clustering of tourist attractions is essential to support more effective regional development planning, promotion, and policy strategies. This study applies data mining with an unsupervised learning approach to cluster tourist attractions in Madura using a comparison of two methods: Fuzzy C-Means (FCM) and Fuzzy C-Medoids (FCMedoids). Both methods were evaluated using the Xie-Beni (XB) validation index as an optimization parameter to determine the quality of cluster formation. Preprocessing included data normalization and outlier removal to ensure model stability. Experiments were conducted with varying the number of clusters from 2 to 10 to obtain the smallest Xie-Beni Index value as the best clustering result. The results showed that the Fuzzy C-Medoids method produced a lower Xie-Beni value of 0.09 compared to the Fuzzy C-Means method (5.40), indicating better separation between clusters and a higher density within clusters. These clustering results can be used as a basis for decision-making in developing regional tourism potential, data-driven promotional strategies, and planning sustainable tourism policies on Madura Island.

1 Introduction

The tourism sector is one of the important pillars in regional economic development in Indonesia. Its contribution is not only in increasing Regional Original Income (PAD), but also in employment, micro business development, and increasing the competitiveness of local destinations on the national and international stage[1]. Data from the Ministry of Tourism and Creative Economy shows that the number of domestic and foreign tourists continued to increase significantly before the COVID-19 pandemic, and is now beginning to recover in the post-pandemic phase. This trend requires local governments and tourism managers to understand tourist characteristics, the potential of tourist attractions, and the factors that influence the attractiveness of destinations through a data-driven approach (data-driven tourism management) [1][2]. However, decision-making in tourism management is often still done conventionally, relying on the intuition or subjective perceptions of policy makers. In fact, the complexity of tourism data, which includes the number of visitors, length of stay, hotel occupancy rates, accessibility, and the socio-economic factors of tourists, requires quantitative analysis methods that are capable of objectively uncovering hidden patterns [3][4]. In this context, clustering techniques are one relevant approach for grouping destinations or tourist behavior based on similar characteristics. Clustering methods such as K-Means, Hierarchical Agglomerative Clustering (HAC), and DBSCAN have been widely used to analyze tourism data[5][6]. However, these methods have limitations in handling ambiguous or uncertain data, especially when the boundaries between clusters are unclear. In reality, many tourist destinations have dual characteristics; for example, a destination can be classified as both a nature and cultural tourist destination. Therefore, a more flexible approach is needed, such as Fuzzy Clustering, where each object has a degree of membership in more than one cluster. One popular method is Fuzzy C-Means (FCM)[7][8], which is able to group data by considering the fuzzy membership degree for each data point to the cluster center. FCM is effective in capturing uncertainty and ambiguity in data, as well as providing a more nuanced understanding of tourism data structures. However, FCM also has weaknesses, namely sensitivity to the initialization value of the cluster center (centroid), which can produce inconsistent results (local minima). In addition, FCM relies on Euclidean distance, which can be affected by outliers. To overcome these weaknesses, the Fuzzy C-Medoids (FCMdd) method was developed, which is a variant of FCM that replaces the cluster center (centroid) with a medoid, which is the most representative actual data point in the cluster [9][10][11]. This approach is more robust to outliers and can improve the stability of the clustering results. In this study, both FCM and Fuzzy C-Medoids methods were used comparatively to analyze tourism patterns, with the aim of obtaining optimal clustering results. Furthermore, the quality of the clustering results was evaluated using the Xie–Beni Index, a measure of fuzzy cluster validity that considers intra-cluster density and

* Corresponding author: ykustiyahningsih@trunojoyo.ac.id

inter-cluster separation. A smaller Xie–Beni value indicates that the clusters formed are more compact and well separated [12][13]. Thus, the combination of FCM/FCMdd and Xie–Beni optimization is expected to produce more accurate cluster structures to support decision-making in data-driven tourism development. The main problem in tourism data analysis is the diversity and uncertainty of data. Tourism data usually comes from various agencies such as the Tourism Office, the Central Statistics Agency (BPS), and online platforms, resulting in variations in format, scale, and quality. This makes it difficult to consistently identify tourist behavior patterns.

In addition, conventional clustering methods such as K-Means and DBSCAN are often unable to handle overlapping clusters, where one object can be similar to more than one group. For example, tourist attractions that offer both natural scenery and local cuisine can have dual characteristics, which cannot be accommodated by the hard clustering approach. The Fuzzy C-Means method can actually overcome most of these problems by providing fuzzy membership degrees, but FCM has inherent weaknesses such as: Dependence on the selection of random initial centroid points, so that the results can differ each time it is run. Sensitivity to outliers and noise, because the calculations are based on Euclidean distance. Potential convergence to local minima, so it does not always produce the best global solution. Meanwhile, Fuzzy C-Medoids offers advantages in stability and resistance to outliers by using medoids as cluster representations. However, the challenge that arises is how to determine the optimal number of clusters (k) that best suits the data structure. This is where Xie–Beni Index Optimization plays an important role, as it is able to provide a quantitative measure for objectively assessing the quality of fuzzy clustering results. Based on these issues, this study focuses on the following main questions: the application of the Fuzzy C-Means and Fuzzy C-Medoids methods can be used to analyze the patterns and characteristics of the tourism sector in Madura, Indonesia; the optimal number of clusters produced by each method based on the Xie–Beni Index[13][14][15], and which method provides better, more stable, and more interpretable cluster results for regional tourism policy decision-making.

Several previous studies have applied clustering techniques to the tourism sector and spatial analysis using the K-Means method to group tourist destinations based on the number of visitors and length of stay, but the results are still rigid because each object only enters one cluster. The DBSCAN method was used to detect areas with hidden tourism potential in Central Java, but the results depend on the selection of the epsilon and minPts parameters. In the context of fuzzy clustering, the integration of Fuzzy C-Means and Neural Network was used to predict tourism business performance [7]. The study showed that FCM can capture nonlinear relationships between variables well. The ensemble clustering method obtains stable results against data variations in large datasets K-Means and fuzzy decision models for classification show that fuzzy methods are superior to conventional approaches in dealing with heterogeneous data [17][18].

However, studies directly comparing the performance of Fuzzy C-Means and Fuzzy C-Medoids in the context of tourism with Xie–Beni Index-based optimization are still rare. Most studies only use one fuzzy method without evaluating the robustness and validity of clusters. Therefore, this study contributes to filling this gap with a comparative and optimal approach to obtain the most representative cluster model for tourism data.

2 Research methods

2.1 Research design

This study uses an exploratory descriptive quantitative approach, focusing on analyzing tourism data patterns and characteristics using fuzzy clustering algorithms. This approach was chosen because it can identify hidden patterns in data without requiring target variables (unsupervised learning). In general, this study was designed as a computational experiment comparing two main algorithms, Fuzzy C-Means (FCM) and Fuzzy C-Medoids (FCMdd), and evaluating their performance using the Xie–Beni Index as a measure of cluster validity. The results of the comparison were used to determine the best method for describing the structure of tourism data and to provide a basis for strategic decision-making for destination managers.

2.2 Research variables

This study involves two main types of variables:

1. Input Variables (X): numerical variables that reflect the tourism characteristics of each region.
2. Output Variables (Cluster Results): the results of grouping tourist destinations based on the level of similarity of attributes obtained from fuzzy clustering.

The following table explains the variables used:

Table 1. Research variables.

No	Variable	Description	Data Type
1	Total area	Tourist area	Numeric
2	Facilities	Number of supporting facilities	Numeric
3	Accommodation	Number of nearby hotels	Numeric
4	Ticket price	Tourist entrance ticket	Numeric
5	Rating	Level of satisfaction with the tourist attraction	Numeric
6	Number of foreign tourists	Total tourists per year	Numeric
7	Number of domestic tourists	Total tourists per year	Numeric
8	Distance to the nearest city	Distance to the city	Numeric

2.3 Research stages

The research process was conducted in several systematic stages, as shown in Fig. 1 below:

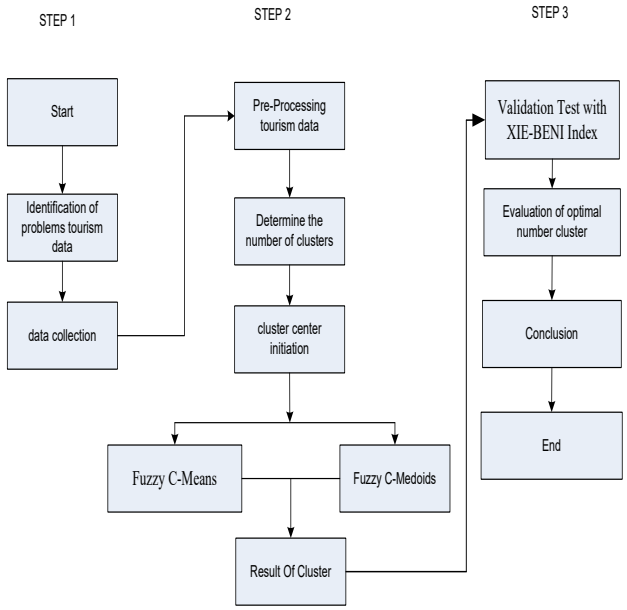

Fig. 1. Research STAGES.

2.3.1 Data collection

Tourism data was collected from official sources and compiled into a single dataset. Each row represents a destination or tourism region, while each column represents a numerical attribute.

2.3.2 Data preprocessing

This stage includes:

1. Data cleaning: removing duplicate entries, filling in missing values with the column mean (mean imputation).
2. Data transformation: converting all variables to numerical format so they can be processed by the fuzzy algorithm.
3. Data standardization: performed using the Z-score normalization method so that all variables have a mean of zero and a standard deviation of one.
   Standardization formula:

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (1)$$

   Where X_i is the original value, μ is the mean, and σ is the standard deviation.

2.3.3 Clustering with Fuzzy C-Means (FCM) [20]

1. Initialize the number of clusters (c) and the degree of fuzzification (m) (generally m = 2).
2. Initialize the membership matrix $U = u_{ik}$ randomly where:
3. $$\sum_{k=1}^{c} u_{ik} = 1 \quad (2)$$
4. Calculate the cluster center (centroid)

$$v_k = \frac{\sum_{i=1}^{n}(u_{ik})^m x_i}{\sum_{i=1}^{n}(u_{ik})^m} \quad (3)$$

5. Calculate the Euclidean distance between each data point x_i and each cluster center v_k:

$$d_{ik} = \|x_i - v_k\| \quad (4)$$

6. Update the membership matrix

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{ij}}\right)^{2/(m-1)}} \quad (5)$$

7. Repeat Steps 3-5 until the objective function converges:

$$J_m = \sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m \|x_i - v_k\|^2 \quad (6)$$

The algorithm stops if the change in the value of J_m is less than the tolerance ε.

2.3.4 Clustering with Fuzzy C-Medoids (FCMdd) [11]

Fuzzy C-Medoids is a variant of FCM that replaces cluster centers with medoids, which are actual data points that have the smallest total distance to their cluster members. The process is as follows:

1. Initialize initial medoids randomly.
2. Calculate the distance between each point and the medoid.
3. Update the fuzzy membership degree using the same equation as FCM.
4. For each cluster, select a new medoid that minimizes the objective function:

$$J_m = \sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m \|x_i - m_k\|^2 \quad (7)$$

   With m_k is the data that becomes the medoid.
5. Repeat until the medoids do not change significantly.
   The advantage of FCMdd is its ability to avoid the influence of outliers and maintain the stability of the results because the cluster centers are real data.

2.3.5 Evaluation with Xie-Beni Index [13]

To measure the quality of fuzzy clustering, the Xie-Beni Index (XBI) is used as a measure of validity. The formula is as follows:

$$XB = \frac{\sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^m \|x_i - v_k\|^2}{n \times min_{i\neq j}\|v_i - v_j\|^2} \quad (8)$$

With
n= number of data,
c= number of clusters,
u_{ik}= degree of membership of data i to cluster k,
v_k= cluster center (FCM) or medoid (FCMdd),
m= fuzzification parameter (usually 2).
A small XB value indicates more compact and well-separated clusters. Therefore, the optimal k is selected based on the minimum XB value in the test range $k = 3$ to $k = 10$.

2.4 Algorithm implementation

The implementation process was carried out using the Python programming language with the following steps:
1. Data preprocessing and normalization using pandas and scikit-learn.
2. FCM and FCMdd implementation:
   a. FCM was implemented based on the basic algorithm (Bezdek, 1981).
   b. FCMdd is modified so that the medoid is selected based on actual data with minimum distance.
3. Xie–Beni optimization is performed automatically for each k value.
4. Visual Visualization of results using matplotlib and two-dimensional PCA (Principal Component Analysis).
5. Export of final results in the form of fuzzy membership tables, cluster labels, cluster centers/medoids, and performance graphs.

3 Results and discussion

3.1 Dataset description

The tourism dataset used in this study contains a number of quantitative indicators to describe the performance of the tourism sector in the Madura region of Indonesia. The variables analyzed include total area, facilities, accommodation, ticket prices, ratings, number of foreign tourists, number of domestic tourists, and distance from the city. The data is sourced from the Central Statistics Agency (BPS) and the Regional Tourism Office for the years 2019–2024. Before clustering, the data underwent pre-processing (cleaning and normalization). This process included:
(1) Removal of duplicate values,
(2) Filling in missing values using mean imputation, and Standardization using the Z-score method so that all variables have a comparable scale.

The final dataset was then analyzed using two fuzzy clustering methods, namely **Fuzzy C-Means (FCM)** and **Fuzzy C-Medoids (FCMdd)** with a cluster range (k) between 3–10.

3.1.1 Fuzzy C-Means (FCM) clustering results

The FCM method was run with a variation in the number of clusters (k) between 3 and 10. Based on the calculation results, the smallest Xie–Beni Index value was obtained at k = 3 with a value of 16.296879, which means that the best cluster structure was achieved in three data groups.

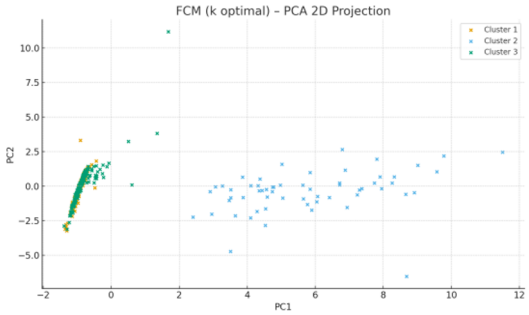

Fig. 2. Visualization of PCA results of FCM clustering (k = 3).

Visualization of the results using Principal Component Analysis (PCA) shows that the three clusters have fairly clear boundaries, although there is a slight overlap in the border area. This is reasonable because the fuzzy nature allows one region to have a degree of membership in more than one cluster. The average maximum membership degree of each data (umax) is 0.74, indicating that the majority of data has a fairly high degree of cluster certainty. The objective value of the Jm function converges after ±120 iterations, indicating the stability of the FCM algorithm.

3.1.2 Fuzzy C-Medoids (FCMdd) clustering results

Fuzzy C-Medoids was tested on a range of k = 3 – 10 with a fuzzification parameter of m=2m = 2m=2. Based on the calculation results, the minimum Xie–Beni value was obtained at k = 10 with a value of 2.29×10^{12}. This value is much greater than that of FCM, indicating that the FCMdd cluster structure for this dataset is less compact and less separated. Analysis shows that the use of real medoids as cluster centers makes FCMdd more robust against outliers, but for homogeneous datasets with relatively little variation, this method produces many clusters of unbalanced sizes. Some clusters contain only one or two data points, which worsens the Xie–Beni value.

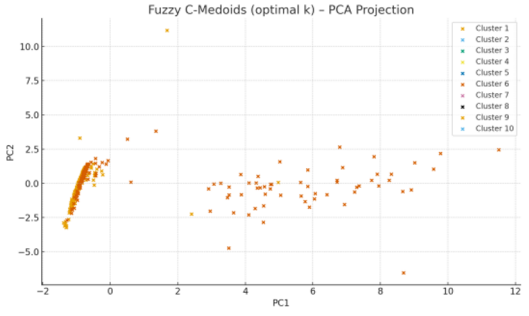

Fig. 3. PCA visualization of FCMdd clustering results (k = 10).

The PCA visualization shows that most FCMdd clusters overlap with each other. In other words, although FCMdd is resistant to extreme values, it is less efficient for tourism datasets that are aggregate in nature and have high inter-attribute correlations.

3.2 Comparison of FCM and FCMdd Results

Table 2. Comparative results between the two methods.

Aspect	Fuzzy C-Means (FCM)	Fuzzy C-Medoids (FCMdd)
Optimal Number of Clusters	3	10
Minimum Xie–Beni Value	16.2969	2.29×10^{12}
Number of Convergence Iterations	±120 iterations	±160 iterations
Robustness to Outliers	Moderate	High
Intra-Cluster Compactness	Good	Low
Inter-Cluster Separation	Clear	Weak
Stability of Results	Good	Tends to be Fluctuating

From the table above, it can be concluded that:
1. FCM produces a clearer and more easily interpreted cluster structure..
2. FCMdd only excels in robustness to outliers, but does not provide better cluster validity in this dataset.
3. The much smaller Xie–Beni Index value in FCM indicates more compact and separate clusters, making it more suitable for grouping tourist destinations.

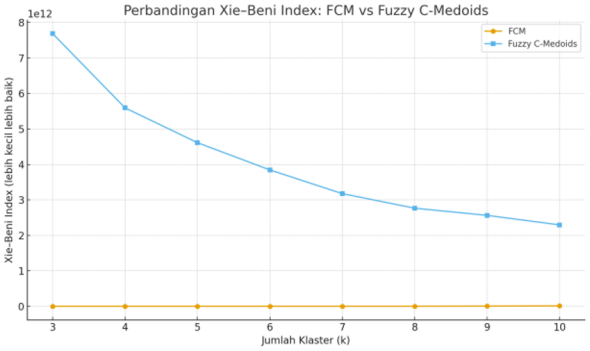

Fig. 4. Comparison of Xie–Beni values between FCM and FCMdd.

A comparison of the Xie–Beni vs. k graphs shows a trend whereby, in FCM, the index value decreases dramatically to a minimum point at k = 3 and increases thereafter, while in FCMdd, the index value fluctuates greatly and does not show a consistent pattern.

3.2.1 Interpretative analysis

The results of FCM clustering into three groups can be further interpreted as a basis for policy recommendations:
1. Cluster 1 – Leading Destinations: Areas with complete facilities, high tourist numbers, and high ratings. Policy focus: maintaining service quality, expanding digital promotion, and developing national-level events..
2. Cluster 2 – Developing Destinations: Areas with strong potential but still requiring infrastructure and human resource support. Strategy: strengthening branding, improving transportation access, and providing training in creative economics.
3. Cluster 3 – Potential Destinations: Areas that are newly developing or have attractive natural resources but are not yet optimally managed. Recommendations: government intervention in the form of investment in basic facilities, partnerships with local MSMEs, and promotion of village tourism.

4 Conclusion

Based on the results of the analysis and discussion, the Fuzzy C-Means (FCM) method produced three optimal clusters with an Xie–Beni value of 16.296879, indicating compact and well-separated clustering. The Fuzzy C-Medoids (FCMdd) method produced a much larger Xie–Beni value (2.29×10^{12}) with a less stable cluster structure. FCM is superior in terms of accuracy, convergence, and interpretability of results, while FCMdd is superior in dealing with data with extreme outliers. Optimization of the Xie–Beni Index proved effective in determining the optimal number of clusters, as it considers the balance between internal density and external separation. The clustering results can be used to segment tourist destinations into three main categories: flagship, developing, and potential which is useful for regional tourism policy planning. Further research could integrate meta-heuristic algorithms such as Particle Swarm Optimization (PSO) or Genetic Algorithm (GA) to improve cluster center optimization, and could add spatial parameters (geographic coordinates) so that the clustering results can be visualized in the form of GIS-based digital maps.

References

1. Muqorobin, Sri Supatminingsih, Dewi Hermawati Wahyuningsih, Pengembangan Metode Klasterisasi Untuk Pengelompokan Potensi Wisata guna Menunjang Digitalisasi Wisata di Indonesia. JST (Jurnal Sains dan Teknol.) **13**, 2, 217–230 (2024). https://doi.org/10.23887/jstundiksha.v13i2.84743
2. A. Ardianto, A. E. Ardianto, Rancang Bangun Aplikasi Pariwisata Malang Berbasis Android. J. Mhs. Fak. Sains dan Teknol. **1**, 1, 1–6 (2015)
3. E. Sugiharto, I.-, I. D. Wijaya, Sistem Rekomendasi Tempat Wisata di Malang Raya Dengan Metode Fuzzy Berbasis Web. J. Apl. Dan Inov. Ipteks

Soliditas **4**, 1, 8 (2021). https://doi.org/10.31328/js.v4i1.1731

4. M. Islamiyah, P. Subekti, T. Dwi Andini, Pemanfaatan Metode Item Based Collaborative Filtering Untuk Rekomendasi Wisata Di Kabupaten Malang. J. Ilm. Teknol. Inf. Asia **13**, 2, 143–150 (2019)
5. Y. Ida, E. Fujita, T. Hirose, Classification of volcano-seismic events using waveforms in the method of k-means clustering and dynamic time warping. J. Volcanol. Geotherm. Res. **429**, 107616 (2022). https://doi.org/10.1016/j.jvolgeores.2022.107616
6. R. Cerqueti, V. Ficcadenti, Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million. Chaos, Solitons and Fractals **158**, 111975 (2022). https://doi.org/10.1016/j.chaos.2022.111975
7. E. K. H. Leung, S. Das, T. Bektaş, T. M. Choi, Optimizing delivery systems within the e-retail context: a weighted self-organizing map for delivery region partitioning. Eur. J. Oper. Res. **2025** (2025). https://doi.org/10.1016/j.ejor.2025.09.006
8. Y. Sukmawaty, M. Program, F. O. F. Mathematics, N. Sciences, A Hybrid Model Fuzzy C-Means (FCM) and Multivariate Adaptive Regression Spline (MARS) on the Case of Poor Households in Jombang District
9. D. Marlina, N. Lina, A. Fernando, A. Ramadhan, Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf. **4**, 2, 64 (2018). https://doi.org/10.24014/coreit.v4i2.4498
10. R. Zheng, Y. Bao, L. Zhao, L. Xing, Prediction of steelmaking process variables using K-medoids and a time-aware LSTM network. Heliyon **10**, 12, e32901 (2024). https://doi.org/10.1016/j.heliyon.2024.e32901
11. S. Sen, L. Sahoo, K. Tiwary, V. Simic, T. Senapati, Wireless Sensor Network Lifetime Extension via K-Medoids and MCDM Techniques in Uncertain Environment. Appl. Sci. **13**, 5, (2023). https://doi.org/10.3390/app13053196
12. R. Andriani, R. F. Amanullah, D. Ninosari, Kusrini, Optimization of Clustering Algorithm on Decision Support System of Scholarship Recipients Using Analytical Hierarchy Process Method. J. Phys. Conf. Ser. **1140**, 1 (2018). https://doi.org/10.1088/1742-6596/1140/1/012028
13. M. F. Lathief, I. Soesanti, A. E. Permanasari, Combination of Fuzzy C-Means, Xie-Beni Index, and Backpropagation Neural Network for Better Forecasting Result. **2020**, 72–77 (2020). https://doi.org/10.5220/0009858200720077
14. E. Rouza, L. Fimawahib, Implementasi Fuzzy C-Means Clustering dalam Pengelompokan UKM Di Kabupaten Rokan Hulu. Techno.Com **19**, 4, 481–495 (2020). https://doi.org/10.33633/tc.v19i4.4101
15. N. Zamri et al., Tuning Data Preprocessing Techniques in Enhancing the Accuracy of Self-Organizing MAP (SOM) for Prostate Genomic Cancer. Procedia Comput. Sci. **263**, –, 892–899 (2025). https://doi.org/10.1016/j.procs.2025.07.107
16. M. S. Mahmud, J. Z. Huang, R. Ruby, K. Wu, An ensemble method for estimating the number of clusters in a big data set using multiple random samples. J. Big Data **10**, 1 (2023). https://doi.org/10.1186/s40537-023-00709-4
17. H. N. Prabowo, R. Setyadi, W. A. Prabowo, Application of Data Mining for Clustering of Foreign Tourist Visits Based on Arrival Entrance. Sinkron **7**, 1, 49–58 (2022). https://doi.org/10.33395/sinkron.v7i1.11217
18. Y. Kustiyahningsih, E. Rahmanita, E. M. S. Rochman, Imamah, A. Amalina, M. H. A. Sobri, Decision Support System of Salt Points Grouping Using K-Means Method. J. Phys. Conf. Ser. **2406**, 1 (2022). https://doi.org/10.1088/1742-6596/2406/1/012022
19. Husni, Y. Kustiyahningsih, F. H. Rachman, E. M. S. Rochman, H. Yulian, Query Expansion Using Pseudo Relevance Feedback Based on the Bahasa Version of the Wikipedia Dataset. AIP Conf. Proc. **2679** (2023). https://doi.org/10.1063/5.0111273
20. M. Peng et al., Advanced image segmentation for precision agriculture using CNN-GAT fusion and fuzzy C-means clustering. Comput. Electron. Agric. **226**, 109431 (2024). https://doi.org/10.1016/j.compag.2024.109431