Journal of Applied Informatics and Computing (JAIC)
Vol.8, No.2, December 2024, pp. 509~515
e-ISSN: 2548-6861

Comparison of Clustering Algorithms: Fuzzy C-Means, K-Means, and DBSCAN for House Classification Based on Specifications and Price

Dhendy Mardiansyah Putra [1]*, Ferian Fauzi Abdulloh [2]**
* Informatika, Universitas AMIKOM Yogyakarta
dhendy.putra@students.amikom.ac.id [1], ferian@amikom.ac.id [2]

Article Info

Article history:

Received 2024-10-09
Revised 2024-11-12
Accepted 2024-11-14

Keyword:

Clustering,
K-Means,
Fuzzy C-Means,
DBSCAN,
Housing Classification,
Real Estate Analysis.

ABSTRACT

This study aims to compare the performance of three clustering algorithms, namely Fuzzy C-Means, K-Means, and DBSCAN, in grouping houses based on their specifications and prices. The data used includes features such as price, building area, land area, number of bedrooms, number of bathrooms, and availability of garages. The performance of these algorithms was evaluated using Silhouette Score and Davies-Bouldin Score to determine the quality of cluster separation. The results indicate that K-Means achieved the best performance with the highest Silhouette Score of 0.7702 for two clusters, followed by Fuzzy C-Means, which excelled in handling overlapping clusters. DBSCAN, while effective in detecting outliers, showed suboptimal performance for this housing dataset. These findings suggest that K-Means is the most suitable clustering method for housing data, while Fuzzy C-Means and DBSCAN can serve as alternatives depending on the data characteristics. This research is expected to assist in making the house searching and classification process more efficient and provide additional insights for developers in shaping housing market strategies.

I. INTRODUCTION

Indonesia is the world's fourth most populous country, with over 270 million people spread across more than 17,000 islands [1]. Every resident needs a home as a place to live and find protection. A house is not only a basic human necessity but also a symbol of stability, security, and well-being. Amid the rapid population growth and urban development in Indonesia, the demand for adequate housing continues to rise. However, the process of finding a suitable home often presents challenges for the public [2]. Factors such as price, location, and house specifications are key considerations in determining the right choice.

When choosing a home, prospective buyers need to consider several important aspects, including price, location, size, design, and technical specifications, which encompass building materials, available facilities, and technology used [3]. Price is often a primary factor influencing decisions, as it directly relates to the financial capacity of potential buyers [4]. Additionally, the price of a house is a determinant of the quality of its specifications, comfort, safety, and energy efficiency. Thus, homebuyers need to obtain complete and accurate information about the house's specifications and price to make well-informed decisions.

House specifications include various crucial aspects such as land area, building area, number of bedrooms, number of bathrooms, and availability of additional facilities like garages and gardens [5]. Furthermore, building materials, construction technology, and location also affect the house's price and value. A home situated in a city center, a suburban area, or a rural region will have different price ranges, even if the specifications are similar [6]. Location is a critical factor influencing the appeal of a property, as it relates to accessibility, public facilities, and the surrounding environment [6]. Therefore, classifying homes based on specifications and price is essential to provide a clearer picture for potential buyers and property developers alike.

This research aims to categorize houses based on their specifications and prices using clustering methods [7]. Clustering is a technique in machine learning used to group data into clusters based on similarities or specific characteristics [8]. In this context, the study will compare the

performance of several clustering algorithms, namely Fuzzy C-Means, K-Means, and DBSCAN. These three algorithms are chosen because of their distinct characteristics in data grouping, which are expected to provide comprehensive results in house classification. Fuzzy C-Means (FCM) is a clustering method that allows data points to belong to multiple clusters with certain membership degrees. FCM is suitable for handling data with uncertainty or data that may belong to multiple groups [9]. K-Means is one of the most widely used clustering methods, where data is divided into clusters based on proximity to cluster centers [10]. K-Means is known for its speed and simplicity, but it has a drawback in requiring manual determination of the number of clusters [11]. On the other hand, DBSCAN is an algorithm that groups data based on density, capable of identifying clusters with irregular shapes [12]. Unlike K-Means, DBSCAN does not require a predefined number of clusters, making it more flexible.

By applying these clustering approaches, the process of finding a home that matches the buyer's criteria is expected to become more efficient and straightforward [7]. Additionally, this research aims to provide valuable insights for property developers in understanding housing market segments in Indonesia, enabling them to develop strategies that better address consumer needs. Clustering housing datasets not only facilitates the house-search process but also offers insights for developers in identifying pricing patterns and specifications preferred by the market [13]. As a result, the findings of this research could have a wide-reaching impact on various stakeholders in the property sector, including prospective buyers, developers, and even the government in formulating housing policies.

II. METHOD

This research utilizes three methods: Fuzzy C-Means, K-Means, and DBSCAN, to determine which is the most effective in clustering housing data that will be tested by the researcher [14]. The aim of this study is to assist in checking housing clusters so that the appropriate model can be directly identified. The research flow can be seen in Figure 1.

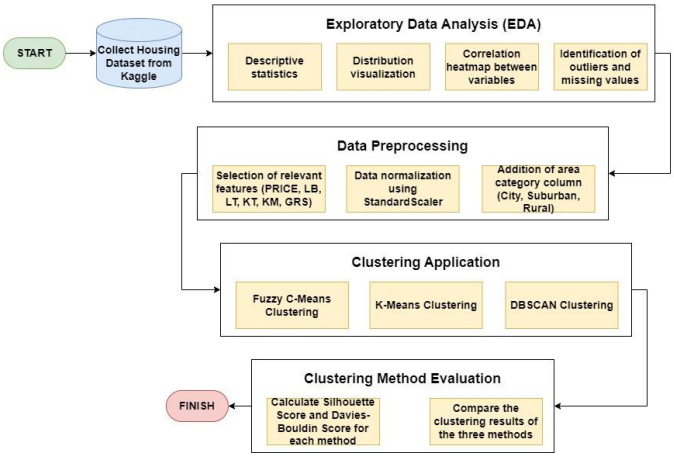

Figure 1. Research Flow

A. *Import Dataset*

In this study, the dataset containing housing-related information was sourced from Kaggle [15]. To import the data into Google Colab, the first step involved mounting Drive using the *google.colab* library, allowing access to files stored in Drive. The dataset file in *.csv* format was then read using the pandas library, specifying the path to the file location in Google Drive. Once the data was successfully loaded, relevant columns for the study were selected using the *.iloc* method, focusing on columns at indices 3, 4, 5, 6, dan 7. This selection ensures that only necessary and relevant data is extracted, preparing it for further analysis and processing in subsequent stages of the research.

B. *Exploratory Data Analysis (EDA)*

The Exploratory Data Analysis (EDA) process was carried out to gain a deeper understanding of the structure and characteristics of the housing dataset [16]. The first step involved obtaining an overview of the data using *df.head()*, which displays the first few rows of the data, including information on price, building area (LB), land area (LT), number of bedrooms (KT), bathrooms (KM), and garage (GRS). Next, descriptive statistics of the dataset were analyzed using df.describe() to understand the distribution of each column, including mean values, standard deviation, and the rage of minimum and maximum values [17]. Information regarding data types and the presence of missing values was also checked using *df.info()*.

To understand the distribution of house prices, a histogram was created, revealing that most houses fall within a specific price range, with a few houses priced significantly higher, indicating the presence of outliers [18]. Further analysis was conducted by creating a correlation heatmap between the numerical variables in the dataset, illustrating the relationships between various features such as price, building area, and land area. A pair plot was also utilized to visually depict the relationships and distributions of these variables, providing a deeper insight into their interactions [19].

Additionally, missing values in the dataset were checked and recorded using the *df.isnull().sum()* method [20]. Finally, a box plot for the house price variable was created, providing a clear visualization of the price distribution and helping to identify potential outliers. This EDA process offered a comprehensive overview of the data to be analyzed, guiding the subsequent steps in this housing-related research. The EDA workflow is illustrated in Figure 2.

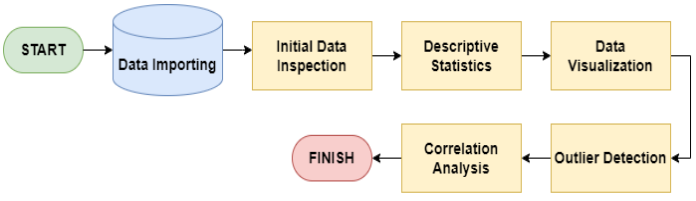

Figure 2. Exploratory Data Analysis

C. Preprocessing Data

In the data preprocessing stage, the variables selected for analysis were based on relevant features, namely PRICE, LB (Building Area), LT (Land Area), KT (Bedrooms), KM (Bathrooms), and GRS (Garage). The data normalization process was carried out using StandardScaler from sklearn to transform the data scale into a normal distribution with a mean of zero, allowing these features to be compared equally during the clustering analysis [21].

Next, the data was categorized based on a combination of LT (Land Area) and PRICE to provide additional insights regarding the area. This categorization determines whether a property falls into the "Urban," "Suburban," or "Rural" category based on the range of land area and house price. This category was then added as a new column in the dataset, enriching the analysis with spatial context in Figure 3.

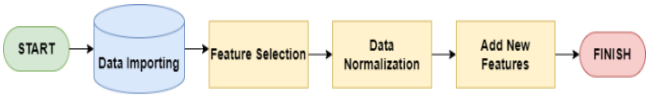

Figure 3. Preprocessing

D. Clustering

Three different clustering methods were tested on this data Fuzzy C-Means Clustering, K-Means Clustering, and DBSCAN Clustering, to group houses based on the available features. Each method was applied to analyze how the houses can be categorized into clusters according to their characteristics, such as price, land area, building area, and other features [22].

1) Fuzzy C-Means: Clustering is conducted with a variation in the number of clusters, ranging from 2 to 5. For each cluster count, the Fuzzy C-Means algorithm attempts to group the data and evaluates its performance using the Silhouette Score, which provides insight into how well the data is clustered [23]. The resulting clusters are then labeled based on predefined categories of house sizes, such as "Small House," "Medium House," and "Luxury House." Cluster names are updated within the dataframe according to these defined categories. The basic formula of Fuzzy C-Means aims to minimize an objective function that involves the membership degree of each data point to a specific cluster. The objective function of Fuzzy C-Means can be expressed as follows.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} U_{ij}^{m} \parallel xi - cj \parallel^2$$

Explanation :

- J_m = Objective function to minimize the distance
- U_{ij} = Membership degree of the *i*-th data point in the *j*-th cluster
- m = Fuzziness parameter for the objective function
- X_i= The *i*-th data point
- C_j= Centroid of the *j*-th cluster
- $\parallel xi - cj \parallel^2$ = Euclidean distance between data point x and centroid c

2) K-Means Clustering: Similar to the Fuzzy C-Means method, K-Means Clustering is performed with a varying number of clusters, ranging from 2 to 5. The K-Means algorithm groups the data based on computed centroids, and the results are evaluated using the Silhouette Score [24]. Each cluster is then labeled according to house size categories, similar to the categories used in Fuzzy C-Means.

$$J = \sum_{i=1}^{N} \parallel x_i - C_{k(i)} \parallel^2$$

Explanation :

- J = Objective function to minimize the distance
- x_i = The *i*-th data point
- $C_{k(i)}$ = Centroid of the cluster *k* assigned to data point X*i*
- $\parallel xi - C_{k(i)} \parallel^2$ = Euclidean distance between data point *x* and the centroid $C_{k(i)}$

3) DBSCAN Clustering: The DBSCAN algorithm is employed with various combinations of parameters, specifically eps (the maximum distance between two samples for them to be considered neighbors) and min_samples (the minimum number of samples required to form a cluster) [25]. The clustering results are then analyzed using the Silhouette Score to assess the quality of the clusters formed. The resulting clusters are labeled based on house size characteristics, and the clustering outcomes are added to the dataframe.

4) Euclidean distance.

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}$$

Explanation :

- $d(p,q)$ = Euclidean distance between two points *p* and *q*
- $p_x,\ q_x,\ p_y,\ q_y$ = Coordinates of the two points *p* and *q*

After clustering is completed, the Davies-Bouldin Score is calculated for each clustering method (Fuzzy C-Means, K-Means, and DBSCAN) [26] to evaluate the quality of the clusters formed. The Davies-Bouldin Score provides an indication of the degree of separation between clusters; a lower score suggests better clustering quality [27]. However, with DBSCAN, calculating the score can sometimes be challenging due to instances of noise or when only a single cluster is formed.

This entire process enables the identification of groups of houses with similar characteristics, facilitating further analysis of the profile of each property group.

III. Result and Discussion

A. Clustering Results Using the Three Methods

This study employs three clustering methods Fuzzy C-Means, K-Means, and DBSCAN to segment housing data based on features such as price, building area (LB), land area (LT), number of bedrooms (KT), number of bathrooms (KM),

and garage availability (GRS). Each method is evaluated using the Silhouette Score and the Davies-Bouldin Score to assess the quality of cluster separation.

1) Fuzzy C-Means

In the *Fuzzy C-Means (FCM)* method, tests were conducted with clusters ranging from 2 to 5. The results indicate that *FCM* effectively segments the housing data, with *Silhouette Scores* ranging from 0.6268 to 0.7600. The highest Silhouette Score of 0.7600 was achieved with 2 clusters, suggesting that this configuration provides the most optimal cluster separation. The clustering results are summarized in Table 1.

TABLE I
SILLHOUTTE SCORE FUZZY C-MEANS

Cluster	Silhouette Score
Cluster 2	0.7600
Cluster 3	0.6566
Cluster 4	0.6305
Cluster 5	0.6268

Additionally, the Davies-Bouldin Score for Fuzzy C-Means is 0.5105, indicating a good level of separation between clusters. However, there is some overlap between clusters, particularly when the number of clusters increases. there was some overlap between clusters, which is expected given the flexibility of the FCM approach.

One of the key advantages of Fuzzy C-Means is its flexibility in grouping data. Unlike hard clustering methods like K-Means, Fuzzy C-Means allows each house to have partial membership in more than one cluster. This provides a "softer" clustering approach, which is particularly useful in cases where properties exhibit characteristics that might fall into more than one category. For example, a house with a price and size that are borderline between medium and luxury categories can be represented by partial memberships in both clusters, allowing for a more nuanced classification.

2) K-Means Clustering :

The *K-Means* method applied to this housing dataset showed relatively consistent clustering results. Tests with 2 to 5 clusters were performed, and the highest *Silhouette Score* achieved was *0.7702* with *2 clusters*, indicating very good separation at this cluster count. However, as observed in *Fuzzy C-Means*, the quality of clustering decreased as the number of clusters increased. For instance, with *5 clusters*, the Silhouette Score dropped to *0.6299*. The results for each cluster configuration are presented in Table 2.

TABLE II
SILLHOUTTE SCORE K-MEANS

Cluster	Silhouette Score
Cluster 2	0.7702
Cluster 3	0.6931
Cluster 4	0.6364
Cluster 5	0.6299

In addition, the *Davies-Bouldin Score* for *K-Means* was recorded at 0.5669, which is slightly higher than that of *Fuzzy C-Means*, suggesting that the clusters are somewhat closer together. This indicates that *K-Means* may not be as effective when handling data with more complex distributions or with several outliers. The main advantage of *K-Means* lies in its speed and simplicity. However, its major drawback is the requirement to manually specify the number of clusters. In the context of housing data, where patterns are not always clearly defined, this can become a limitation. *K-Means* may struggle to capture the underlying structure of the data if the optimal number of clusters is not carefully determined. In this study, methods such as the *Elbow Method* and *Silhouette Analysis* were used to ensure that the most suitable cluster count was selected, thus addressing this limitation.

3) DBSCAN Clustering :

DBSCAN, while powerful for identifying irregularly shaped clusters and outliers, is highly sensitive to parameter settings such as epsilon and min_samples. In this study, DBSCAN struggled to form well-defined clusters due to the relatively uniform distribution of the housing dataset. Specifically, the housing features like price and area do not exhibit significant density variations that DBSCAN excels at identifying. We attempted various configurations for epsilon and min_samples (as shown in Table 3), but even the best parameters produced suboptimal results compared to K-Means and Fuzzy C-Means.

TABLE III
SCORE DBSCAN

Cluster	Silhouette Score	DavisB Score
eps=0.3 min_samples=3	-0.1001	1.3596
eps=0.5 min_samples=5	-0.0388	1.4003
eps=0.7 min_samples=7	-0.0451	1.5684
eps=0.8 min_samples=8	0.6268	1.5047
eps=1.0 min_samples=7	0.3108	1.4634

The Silhouette Scores produced by DBSCAN range from -0.1001 to 0.3108, indicating that in some configurations, the algorithm struggles to identify well-defined clusters. The Davies-Bouldin Score also varies, with the highest value reaching 2.6659, suggesting that the clusters formed are not always well-separated. This variability highlights DBSCAN's sensitivity to parameter selection, especially in data sets with complex structures or significant noise.

The advantage of DBSCAN lies in its ability to detect clusters with irregular shapes and its robustness in handling outliers effectively. However, the primary challenge of this method is the difficulty in selecting appropriate parameters, particularly when the data exhibits significant density

variations. This sensitivity to parameter choice can limit DBSCAN's effectiveness in scenarios with diverse data distributions.

B. Evaluation Metrics

In this study, we employed two main evaluation metrics *Silhouette Score* and *Davies-Bouldin Score* to assess the quality of the clustering results. These metrics were chosen because they offer a comprehensive evaluation of both within-cluster similarity and between-cluster separation, which are essential in classifying housing data based on multiple features such as price and area.

1) Silhoutte Scorec: The Silhouette Score measures the similarity of each data point to its own cluster (cohesion) compared to other clusters (separation). It is calculated as follows

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

Where :

- *a(i)* is the average distance between the data point iii and other points in the same cluster.
- *b(i)* is the minimum average distance between the data point iii and points in any other cluster.

2) Davis-Bouldin Score: The Davies-Bouldin Score measures the average similarity ratio between each cluster and the cluster most similar to it. The formula is.

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left(\frac{s(i) + s(j)}{d(i,j)}\right)$$

Where:

- *s(i)* is the average distance between data points within cluster *i*.
- *d(i,j)* is the distance between the centroids of clusters *i* and *j*,
- *n* is the number of clusters.

Lower *Davies-Bouldin* values indicate better cluster separation. This score was particularly important for evaluating how well-separated clusters were, especially in cases where house features (such as price and area) might overlap across different clusters. *A lower Davies-Bouldin Score* suggests that the clusters are compact and well-separated from each other, which is crucial for effective classification of housing data.

3) Alternative Metrics: While Silhouette Score and Davies-Bouldin Score provide a robust assessment of clustering quality, additional metrics such as the Adjusted Rand Index (ARI) and the Calinski-Harabasz Index can offer further insights. ARI is commonly used to measure the similarity between the predicted clustering and a ground truth classification, while the Calinski-Harabasz Index measures the ratio of the sum of between-cluster dispersion and within-cluster dispersion. Although these metrics were not included in this study, they could be valuable in future work to provide a more comprehensive evaluation of clustering performance.

C. Comparison of Clustering Performance

To determine which clustering method is most suitable, the following analysis is based on performance criteria.

1) *Fuzzy C-Means* performs well in cases where the data exhibits uncertainty or overlap between clusters. This makes it ideal when houses have characteristics that fall between categories, such as price and land area near the boundary of categories.
2) *K-Means* offers consistent and straightforward results, especially in situations where the number of clusters is predefined. The highest *Silhouette Score* was achieved by *K-Means* with 2 clusters, making it suitable for simpler data distributions.
3) *DBSCAN :* excels in detecting outliers and clusters with irregular shapes. However, its performance is highly dependent on the selected parameters and often yields negative *Silhouette Scores*, indicating that the clusters formed may not always be well-defined.

D. Best Method

Based on the clustering results and performance metrics, K-Means stands out as the most suitable method for clustering housing data in this study. K-Means achieved the highest Silhouette Score (0.7702 with 2 clusters), demonstrating its ability to effectively group data points into well-separated clusters. This strong performance can be attributed to K-Means' simplicity and efficiency in handling datasets with more uniform distributions. In scenarios where the optimal number of clusters is known or can be estimated, K-Means proves to be an effective tool, as it ensures clear segmentation between different housing categories, such as small, medium, and luxury houses. The method's ability to maintain cluster coherence while ensuring adequate separation makes it particularly useful for real estate data where clear divisions between categories are necessary for market analysis and decision-making.

However, *K-Means* does have its limitations. One major drawback is its reliance on the user to predefine the number of clusters, which can be problematic in cases where the natural structure of the data is unknown. If the number of clusters is not accurately determined, the performance of *K-Means* can degrade, leading to poorly defined clusters. This limitation was mitigated in this study through the use of *Elbow Method* and *Silhouette Analysis*, which helped to identify the optimal number of clusters.

Fuzzy C-Means emerges as a strong alternative to *K-Means*, particularly in cases where the data exhibits overlapping characteristics between clusters. Unlike *K-Means*, which assigns each data point to a single cluster, *Fuzzy C-Means* allows for partial membership in multiple clusters, offering a more flexible and nuanced classification. This method is particularly advantageous when dealing with housing data that straddles category boundaries. For instance, houses that are positioned between medium and luxury

categories in terms of price and size benefit from this soft clustering approach, allowing for more accurate classification and better insights into market segmentation. Although the *Silhouette Score* for *Fuzzy C-Means* was slightly lower than that of *K-Means* (with a maximum score of 0.7600 for 2 clusters), its ability to handle data overlap offers significant advantages in specific contexts, particularly when the housing market exhibits blurred boundaries between segments.

Despite these strengths, *Fuzzy C-Means* also has certain limitations. The most notable is the increase in overlap as the number of clusters grows, which can reduce the clarity of segmentation. Additionally, *Fuzzy C-Means* is computationally more expensive compared to *K-Means*, especially in large datasets where calculating the degree of membership for each data point becomes resource-intensive. Nevertheless, its flexibility in handling ambiguous data makes it a valuable tool for certain clustering tasks, particularly in markets where properties do not neatly fall into distinct categories.

On the other hand, *DBSCAN* proves useful in specific situations where the dataset contains significant outliers or irregularly shaped clusters. The density-based approach of *DBSCAN* allows it to identify clusters of arbitrary shape, making it especially effective in detecting outliers that may not conform to the general distribution of the data. In this study, however, *DBSCAN* struggled due to the relatively uniform density of the housing data. While it excelled in identifying outliers, the method's performance was hindered by the lack of significant density variations in key features such as price and area. Additionally, the sensitivity of *DBSCAN* to its parameters, such as epsilon and min_samples, proved to be a limitation, as small changes in these parameters led to significantly different results, as indicated by the negative *Silhouette Scores* in some configurations.

The primary advantage of *DBSCAN* lies in its ability to handle noise and outliers effectively, making it particularly suitable for datasets with significant variability in density. However, the method's dependence on the correct selection of parameters and the uniform nature of the housing dataset in this study limited its effectiveness. In more complex datasets where clusters are not well-defined or where outliers play a major role, *DBSCAN* could be a valuable clustering tool. Yet, in this particular housing dataset, where the focus was on clear segmentation based on price and area, *DBSCAN* did not perform as well as *K-Means* and *Fuzzy C-Means*.

IV. CONCLUSION

The conclusion of this study indicates that among the three clustering methods used Fuzzy C-Means, K-Means, and DBSCAN. K-Means delivered the best performance in terms of cluster separation. With the highest Silhouette Score of 0.7702 for two clusters, K-Means was able to provide clear segmentation of the housing data, particularly when the required number of clusters can be predefined. This method also proved to be more stable and easier to apply to data with simpler distributions.

Fuzzy C-Means also delivered strong performance, particularly in cases where the data exhibits uncertainty or overlap between clusters. This allows Fuzzy C-Means to handle more complex and not entirely distinct data. With a Davies-Bouldin Score of 0.5105, this method provides reasonably good cluster quality, despite some overlap between clusters.

Meanwhile, DBSCAN demonstrated superior ability in handling outliers and irregularly distributed data, but its clustering results are highly dependent on parameter selection. Although DBSCAN can identify clusters with irregular shapes, the low Silhouette Score and high Davies-Bouldin Score indicate that this method is not always optimal for the housing data tested in this research.

Overall, *K-Means* emerged as the most suitable clustering method for this housing dataset, achieving the highest Silhouette Score of 0.7702 with two clusters. However, *Fuzzy C-Means* provided valuable insights in cases where the housing data exhibited overlapping characteristics, offering a more flexible clustering approach. *DBSCAN* demonstrated strength in handling outliers, but its performance was limited due to the relatively uniform density of the dataset, which made parameter selection challenging. Future work may explore fine-tuning DBSCAN parameters further or incorporating alternative clustering methods for datasets with more complex structures.

REFERENCES

[1] I. M. Adnyana and H. Iswanto, "Open Access Indonesia Journal of Social Sciences," *Open Access Indones. J. Soc. Sci.*, vol. 4, no. 1, pp. 132–142, 2021, [Online]. Available: https://journalsocialsciences.com/index.php/OAIJSS

[2] H. Ward, "Foreword," in *Revitalizing Residential Care for Children and Youth*, Oxford University Press, 2022, pp. xi–xvi. doi: 10.1093/oso/9780197644300.002.0008.

[3] B. Harsanto, *Dasar-Dasar Manajement Operasi: Konsep, Batang Tubuh Ilmu dan Industri 4.0*, 2nd ed. Jakarta: KENCANA, 2022.

[4] J. Ilmiah and U. Muhammadiyah, "Sang pencerah," pp. 504–516, 2024.

[5] A. Widyastuti, "Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Metode Multiple Linear Regression," vol. 4, no. 1, pp. 30–35, 2024, [Online]. Available: http://ejurnal.unim.ac.id/index.php/submit/article/download/3343/1556

[6] I. Mirzaya Putra, *Pengembangan Wilayah*, Pertama. Medan: CV. Prokreatif, 2023.

[7] T. L. Putri *et al.*, "Penerapan data mining pada clustering data harga rumah dki jakarta menggunakan algoritmak-means," vol. 8, no. 1, pp. 1174–1179, 2024.

[8] N. Hendrastuty, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa," vol. 3, pp. 46–56, 2024.

[9] J. Saputra, M. Iqbal, A. Aksha, and L. Maryani, "EXPLORE – Volume 14 No 2 Tahun 2024 Terakreditasi Sinta 5 SK No : 23 / E / KPT / 2019 Analisis Perbandingan Efektivitas Metode Fuzzy C-Means dan K-Means dalam Mengelompokkan Buku Berdasarkan Frekuensi Peminjaman di Perpustakaan SMKN 1 Mandau EXPLORE – Vol," vol. 14, no. 2, pp. 87–92, 2024.

[10] S. Butsianto and N. T. Mayangwulan, "Penerapan Data Mining Untuk Prediksi Penjualan Mobil Menggunakan Metode K-Means Clustering," vol. 3, no. 3, pp. 187–201, 2020.

[11] M. A. Pryono, S. H. Wijoyo, and F. A. Bachtiar, "Analisis

Sentimen Terhadap Program Merdeka Belajar Kampus Merdeka Pada Sosial Media Twitter Menggunakan K-Means Clustering , Support Vector Machine (SVM) dan Syntethic Minority Oversampling Technique (SMOTE),” vol. 1, no. 1, pp. 1–10, 2017.

[12] F. M. Pranata, S. H. Wijoyo, and N. Y. Setiawan, “Analisis Performa Algoritma K-Means dan DBSCAN Dalam Segmentasi Pelanggan Dengan Pendekatan Model RFM,” vol. 1, no. 1, pp. 1–9, 2017.

[13] R. F. Almahdy and W. M. P. D, “Prediksi Harga Rumah Di Kabupaten Bantul Menggunakan Algoritma Support Vector Regression,” vol. 11, no. 2, pp. 152–165, 2024.

[14] I. H. Zahro, U. A. Rosyidah, and L. Handayani, “Implementasi Algoritma Fuzzy C-Means untuk Pengelompokkan Provinsi di Indonesia Berdasarkan Kualitas Perguruan Tinggi,” *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 1, pp. 80–86, 2024, doi: 10.37148/bios.v5i1.102.

[15] W. Anggara, “Daftar Harga Rumah.” Accessed: Jul. 09, 2024. [Online]. Available: https://www.kaggle.com/datasets/wisnuanggara/daftar-harga-rumah/data

[16] R. Dalam, M. Anggaran, B. Manajemen, and P. S. Informasi, “Komparasi Multiple Linear Regression dan Random Forest Regression Dalam Memprediksi Anggaran Biaya Manajemen Proyek Sistem Informasi,” vol. 3, no. 2, pp. 86–97, 2024.

[17] Nasution, A. Lestari, and R. N. S. Fatonah, *Klasifikasi Kondisi Peralatan Elektronik Metode Gaussian Naïve Bayes*. Penerbit Buku Pedia, 2023.

[18] M. Boull, “Two-level histograms for dealing with outliers and heavy tail distributions,” 2023.

[19] A. Erdely and M. Rubio-sánchez, “Visual analysis of bivariate dependence between continuous random variables”.

[20] S. Shah, M. Telrandhe, P. Waghmode, and S. Ghane, “Imputing missing values for Dataset of Used Cars,” in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, 2022, pp. 1–5. doi: 10.1109/ASIANCON55314.2022.9908600.

[21] A. L. Nogueira and C. S. Munita, “The effect of data standardization in cluster analysis,” pp. 1–15, 2021.

[22] T. Malatesta, Q. Li, and J. K. Breadsell, “Distinguishing Household Groupings within a Precinct Based on Energy Usage Patterns Using Machine Learning Analysis,” 2023.

[23] O. N. Purba, D. N. Sitompul, T. H. Harahap, S. R. Dewi, and R. F. Siregar, “Application of Fuzzy C-Means Algorithm for Clustering Customers,” pp. 0–10, 2023.

[24] T. D. Pangestu, V. Y. Ardila, M. Suteja, and S. P. Barus, “Klasterisasi Hewan berdasarkan Morfologi dengan K-Means Klastering untuk Memudahkan Pemahaman Taksonomi Hewan Klastering Animals based on Morphology with K-Means Klastering to Facilitate Understanding of Animal Taxonomy,” vol. 14, no. 2, pp. 10–20, 2024.

[25] 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.

[26] B. E. Adiana, I. Soesanti, A. E. Permanasari, J. G. No, J. G. No, and J. G. No, “Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering,” no. 2, pp. 23–32, 2018, doi: 10.21460/jutei.2017.21.76.

[27] A. Nowak-brzezi, “How the Outliers Influence the Quality of Clustering?,” 2022.