Clustering Tourism Destinations in Denpasar City in 2023 Based on Visitor Preferences Using K-Means and DBSCAN Clustering Methods

Komang Satya Dharmawan[a1], Gusti Made Arya Sasmita[a2], I Nyoman Piarsa[b3]
[a]Department of Information Technology, Faculty of Engineering, Udayana University
Bukit Jimbaran, Bali, Indonesia-8036110
e-mail: [1]satyadharmawan@student.unud.ac.id, [2]aryasasmita@unud.ac.id, [3]manpits@unud.ac.id

Abstrak

Dalam era digital, meningkatnya jumlah destinasi wisata di Kota Denpasar menimbulkan tantangan dalam pengelolaan yang efektif dan berkelanjutan. Penelitian ini mengembangkan sistem pengelompokan destinasi wisata tahun 2023 berdasarkan preferensi pengunjung menggunakan metode clustering K-Means dan DBSCAN. Data kunjungan wisatawan dan fasilitas destinasi dari Dinas Pariwisata Kota Denpasar diolah melalui pra-pemrosesan, termasuk pengisian data hilang dan standarisasi fitur. K-Means mengelompokkan destinasi berdasarkan jarak ke pusat klaster, sedangkan DBSCAN unggul dalam mendeteksi klaster berdasarkan kepadatan dan outlier. Evaluasi menggunakan Elbow Method dan Silhouette Score menunjukkan bahwa K-Means lebih optimal dalam membentuk klaster dengan jumlah tertentu, dengan Silhouette Score sebesar 0,654 untuk wisatawan asing, 0,614 untuk wisatawan domestik, dan 0,579 untuk wisatawan gabungan. Sebaliknya, DBSCAN unggul dalam menangani distribusi data tidak teratur dan mengidentifikasi outlier, dengan Silhouette Score sebesar 0,681 untuk wisatawan asing, 0,578 untuk wisatawan domestik, dan 0,529 untuk wisatawan gabungan. Tingkat kesesuaian antara kedua metode mencapai 86,7% untuk wisatawan asing, 80,0% untuk wisatawan domestik, dan 83,3% untuk wisatawan gabungan. Hasil penelitian divisualisasikan melalui dashboard interaktif yang memetakan distribusi klaster, memungkinkan Dinas Pariwisata untuk menganalisis pola kunjungan dan merumuskan strategi pengembangan destinasi wisata secara lebih terarah dan efisien.

Kata kunci: *Data Mining, Clustering, K-Means, DBSCAN, Pariwisata, Kota Denpasar*

Abstract

In the digital era, the increasing number of tourist destinations in Denpasar City poses challenges for effective and sustainable management. This study develops a clustering system for tourist destinations in 2023 based on visitor preferences using K-Means and DBSCAN methods. Data on tourist visits and destination facilities, sourced from the Denpasar City Tourism Office, were processed through preprocessing steps, including missing data imputation and feature standardization. K-Means clusters destinations based on proximity to cluster centroids, while DBSCAN excels in detecting density-based clusters and outliers. Evaluation using the Elbow Method and Silhouette Score indicates that K-Means is more optimal for forming a specific number of clusters, with Silhouette Scores of 0.654 for foreign tourists, 0.614 for domestic tourists, and 0.579 for combined tourists. Conversely, DBSCAN performs better in handling irregular data distributions and identifying outliers, with Silhouette Scores of 0.681 for foreign tourists, 0.578 for domestic tourists, and 0.529 for combined tourists. The agreement rate between the two methods reaches 86.7% for foreign tourists, 80.0% for domestic tourists, and 83.3% for combined tourists. The results are visualized through an interactive dashboard mapping cluster distributions, enabling the Tourism Office to analyze visit patterns and formulate targeted, efficient strategies for destination development

Keywords : *Data Mining, Clustering, K-Means, DBSCAN, Tourism, Denpasar City*

1. Introduction

Advancements in information technology have transformed the tourism sector, encompassing activities such as culinary experiences, accommodation, transportation, and cultural entertainment [1]. Denpasar City, Bali's tourism hub, offers culturally and historically rich destinations, including the Bajra Sandhi Monument, Sindhu Beach, and Bali Museum, attracting both domestic and international visitors. These destinations significantly contribute to the local economy. However, the growing number of tourist sites and intensifying global competition demand effective management strategies. Optimal management relies on a deep understanding of visitor preferences, attraction types, and facilities like parking, restrooms, and information boards to enhance destination appeal and tourist experiences [2]. Without proper destination clustering, promotional and development strategies risk being inefficient, hindering equitable tourism growth in Denpasar.

The Denpasar City Tourism Office faces challenges in systematically grouping destinations based on visit patterns and facilities, such as the number of foreign and domestic tourists and infrastructure availability. This study employs data mining techniques, specifically clustering, to address these issues, utilizing datasets on tourist visits and facilities from the Tourism Office [3]. K-Means and DBSCAN clustering methods are applied to categorize destinations. K-Means partitions data based on proximity to cluster centroids, suitable for well-defined clusters, while DBSCAN excels in identifying density-based clusters and outliers, enabling the analysis of complex data like visit frequency and facility quality [4]. This approach aims to uncover visit patterns and support targeted tourism policy development.

Unlike previous studies, such as [1], which used only K-Means to cluster tourism destinations in Bojonegoro, this research integrates K-Means and DBSCAN for a more comprehensive analysis. The combination of centroid-based and density-based approaches ensures robust handling of complex data, revealing hidden patterns in visitor preferences and destination characteristics. The resulting clusters, visualized through interactive tools, are expected to assist the Tourism Office in crafting marketing strategies, prioritizing facility improvements, and promoting equitable destination development, fostering a sustainable and competitive tourism ecosystem in Denpasar aligned with local cultural values.

2. Research Method

The approach for clustering tourism locations in Denpasar City in 2023 begins with data on monthly tourist visits, then moves on to pre-processing and feature standardization. The Python programming language is used to apply the K-Means and DBSCAN algorithms throughout the clustering phase. K-Means clusters destinations depending on their closeness to the cluster center, and the elbow approach determines the best number of clusters. DBSCAN, on the other hand, uses optimal Eps and Min Sample parameters to detect outliers and identify clusters based on density. The Silhouette Score is used to evaluate clustering performance, and the results are shown on an interactive dashboard. Figure 1 presents an overview.

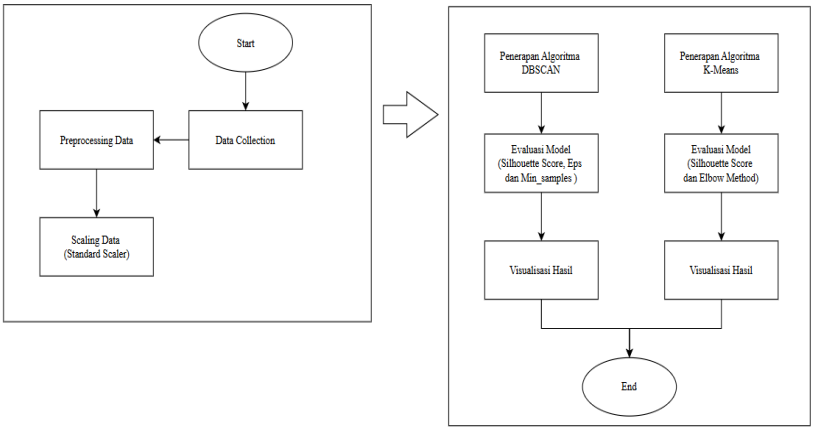

Figure 1. Overview of Research Methods

2.1. Data Collection

The data used in the analysis of clustering tourist destinations in Denpasar City in 2023 includes the number of monthly tourist visits and facilities in 30 tourist attractions (DTW). This data was obtained from the Denpasar City Tourism Office through the data collection method of extracting official documents and stored in CSV format. Tourist visit data includes the number of foreign, domestic, and total combined tourists from January to December 2023, while facility data includes the number of toilets, parking areas, trash bins, sinks, Wi-Fi access, information boards, Tourism Information Centers, rest areas, restaurants, pedestrian paths, and DTW ratings (scale 1-5).

2.2. Data Pre-processing

The goal of this pre-processing step is to prepare the data for use by cleaning it up. Drop columns, missing values, data type conversion, and choosing all numeric columns are among the steps that are completed. The pre-processing stage can be seen in Figure 2.

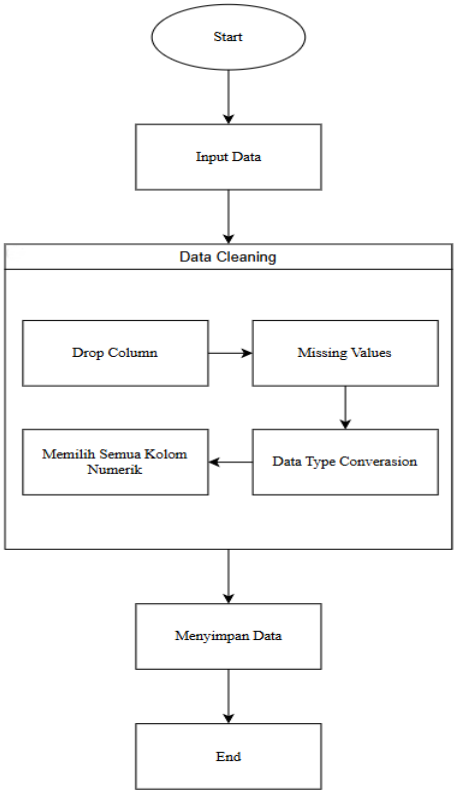

Figure 2. Pre-processing Data

The stages of data preprocessing are as follows:

1. Drop Column
   The drop column stage aims to eliminate columns that are not needed in abstract text summarization research, such as the No and Regency columns.
2. Missing Values
   Filling missing values is a process that aims to overcome missing or empty values in the dataset by replacing them using the median value. This method is used to maintain data consistency and avoid bias that can arise due to incomplete data.
3. Data Type Converasion
   Changing the data type is a data type conversion process, which transforms values from float data type to integer data type.
4. Selecting all Numeric Columns
   Selecting all numeric columns except non-numeric columns (Tourist Attraction Type, and Tourist Attraction Name) because this data has no numeric information.

2.3. Model Testing

Model testing is a process to determine which model or algorithm has better analysis results. This model testing was conducted on the research of tourism destination clustering in Denpasar City by using K-Means and DBSCAN algorithms.

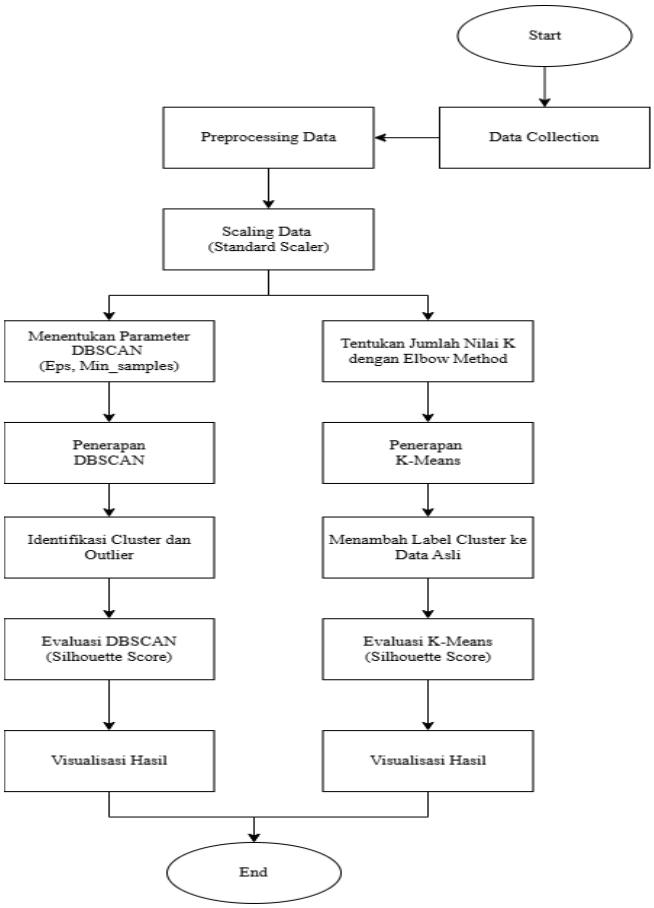

Figure 3. Model K-Means and DBSCAN

Figure 3 illustrates the flow of clustering analysis of tourist destinations in Denpasar City in 2023 using K-Means and DBSCAN. The process begins with the collection of data on tourist visits and facilities, followed by preprocessing to ensure data quality. The data was standardized using StandardScaler to equalize the feature scale. K-Means parameters (number of clusters) are determined via Elbow Method, while DBSCAN uses optimized eps and min_samples parameters. K-Means clusters data based on proximity to the centroid, while DBSCAN is density-based and detects outliers. Performance evaluation is done with Silhouette Score to measure cluster quality. The clustering results are visualized to map the distribution of clusters, with K-Means producing ordered clusters and DBSCAN excelling at detecting outliers, providing insights for tourist destination management.

3. Literature Study

The literature review contains material related to the research, including tourism, data mining, clustering, K-Means, DBSCAN, Silhouette Score.

3.1. Tourism

Tourism is a temporary activity in which one or more individuals travel to a place other than where they live. Based on RI Regulation No. 10 of 2009, tourism includes various activities supported by facilities and services provided by the business world, the community, and the central and local governments. This indicates that tourism is a complex activity involving tourists, tourist attractions, businesses, and the government [5].

3.2. Data Mining

Data mining, also known as knowledge discovery in databases (KDD), is the process of gathering and applying historical data to identify patterns, correlations, and regularities in sizable data sets [6]. The process of applying certain techniques or methodologies, such induction-based learning, to uncover intriguing patterns or information in chosen data is known as data mining. There is a great deal of variation in data mining approaches, methods, or algorithms. The goal classification, clustering, or anomaly detection as well as the KDD process as a whole have a major role in choosing the best approach or algorithm [7].

3.3. Clustering

Clustering, or clustering analysis, is an algorithm that belongs to the unsupervised learning category, as it is able to classify unlabeled data based on the similarity between data. This process divides or groups the data in a set into several clusters, where the data in one cluster has a high level of similarity, while the level of similarity between clusters is relatively low. If the data used has no similarity, then the data will be grouped in different clusters. The main purpose of clustering is to divide data into several similar groups, so as to generate new information regarding the hidden data structure. The potential of clustering is vast, as it can help reveal patterns or structures in data that are useful for various applications, such as classification, clustering, image processing, and pattern recognition [8].

3.4. K-Means

K-Means is a clustering algorithm used in data mining that uses a point-based partitioning method to quickly and efficiently create groups from vast volumes of data. One non-hierarchical data clustering technique that can divide data into two or more groups is K-Means [8]. Using preset criteria, data is divided into multiple segments before being merged into a single cluster [9]. The steps involved in applying optimization with the K-Means algorithm are as follows [10]:

1. Determine the number of clusters (k) contained in the dataset.
2. Determine the center point (centroid) randomly in the first step.
3. Calculate each data's closest distance to the centroid using the formula below:

$$De = \sqrt{(x_i - y_i)^2} + \cdots + (x_n - y_n)^2 \quad (1)$$

Description:

De = Euclidean Distance

(x) = object coordinate

(y) = centroid coordinate

i = number of objects

4. Using the present cluster members, recalculate the cluster points. The average of all the data in a cluster is called the cluster point. The formula can be used to compute it:

$$C_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \quad (2)$$

Description:

x_{ij} =n cluster

p = number of members n cluster

5. Recalculate each object with the Cluster point (new Centroid). If the group calculation does not change again, then the Clustering calculation is complete. But if the group cal-

culation still changes, then the calculation is carried out again like step c until the Cluster members do not move again.

3.5. DBSCAN

DBSCAN a clustering algorithm called DBSCAN (Density-Based Spatial Clustering of Applications with Noise) concentrates on grouping data according to density. It works by identifying data points that are in high-density areas, i.e. points that have many neighbors within a certain radius, and grouping them into clusters [11]. DBSCAN, unlike K-Means, does not require an initial number of clusters, can handle irregular clusters, and ignores noise/outliers. It uses data density, clustering based on minpts within a radius of epsilon (ε), with the number of clusters affected by ε and minpts. The distance between points is generally measured by the Euclidean Distance, although other metrics can be used [12].

The data clustering process using the DBSCAN Algorithm involves several steps, namely:

1. Setting the Minimum Points (minPts) and Epsilon (eps) values.
2. Selecting the starting point randomly.
3. Calculates the distance between points using the Euclidean distance function, with the following equation:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Description:

x and y are two data points

X_i and Y_i are the i feature values of points x and y respectively

n is the number of features

4. Clusters are formed based on data density.
5. A point is classified as a core point and a cluster will emerge if the number of points in radius eps reaches or surpasses minpts.
6. If the point is a boundary point and no other points within the density range can be reached, the process will continue to another point.
7. Steps 3 to 6 will be repeated until all points have been processed.

3.6. Silhouette Score

Silhouette score is an evaluation method used to measure how well each data in a cluster is grouped, and whether the cluster division results from methods such as elbow are correct or not. silhouette score measures the closeness of data to the same cluster compared to other nearby clusters. Silhouette values range from -1 to 1, values close to 0.5 to 1 indicate that the data is in the right cluster, while values close to -1 indicate that the data is more suitable to be in another cluster. With this formula, Silhouette helps in assessing whether the number of clusters produced is truly optimal, thus providing a more accurate picture of the quality of the clustering performed [13]. The formula for Silhouette is as follows:

$$S = \frac{b - a}{\max(a, b)} \quad (4)$$

Keterangan:

S = silhouette

a = average distance of a sample to all points in the same class

b = average distance of a sample and all points in the nearest cluster

4. Result and Discussion

The results of the research conducted by the author based on the research overview process are presented as follows.

4.1. Data Collection

The Denpasar City Tourism Office provided statistics on the number of monthly visitor visits in 2023 as well as details on the city's tourist destination amenities. These data were collected as the initial step of this study and subsequently put through a preprocessing stage. The visitation data includes the number of foreign, domestic, and total visits to 30 tourist attractions (DTW) in Denpasar, Bali, from January to December 2023, reflecting tourist visitation trends. Facility data includes the number of toilets, parking, wifi, TIC, rest areas, restaurants, pedestrian paths, as well as the aggregate total and rating (scale 1-5) of each DTW, which serves as a reference for the evaluation and development of the tourism sector in Denpasar.

4.2. Data Pre-processing

Data preprocessing is a process that aims to clean data from invalid data or unimportant data. The data preprocessing stages include deleting attributes or columns in the dataset that are irrelevant, filling in missing values or empty values and changing the column data type and selecting all numeric columns.

4.3. Model Testing

Modeling with K-Means and DBSCAN grouped the data on foreign, domestic, and combined tourist visit preferences and destination facilities in Denpasar City in 2023 into several clusters. The results show structured preference patterns and facility characteristics. DBSCAN, based on density, recognizes irregular and noise clusters, providing insights into data distribution not detected by K-Means.

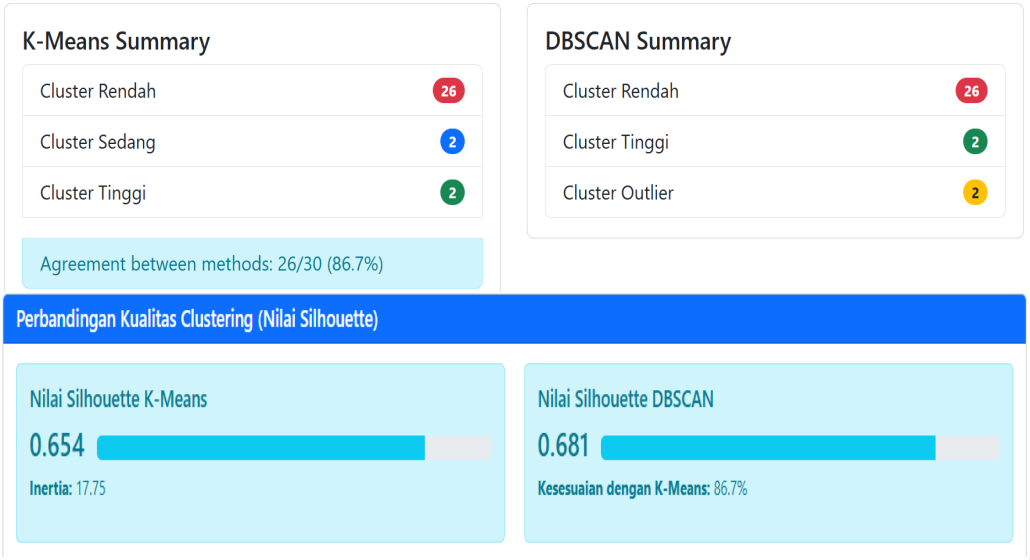

Figure 4 Summary of K-Means and DBSCAN Clustering Comparison of Foreign Tourists

Comparison of the clustering of foreign tourist destinations in Denpasar City in 2023 using K-Means and DBSCAN showed consistent results with an agreement rate of 86.7% (26

out of 30 destinations). K-Means grouped the destinations into three clusters (low: 26 destinations, medium: 2 destinations, high: 2 destinations) with a Silhouette value of 0.654 and inertia of 17.75, reflecting good cluster density. In contrast, DBSCAN produced two clusters (low: 26 destinations, high: 2 destinations) and identified 2 destinations as outliers, with a higher Silhouette value (0.681), indicating better clustering and ability to capture data patterns and unique destinations not detected by K-Means.

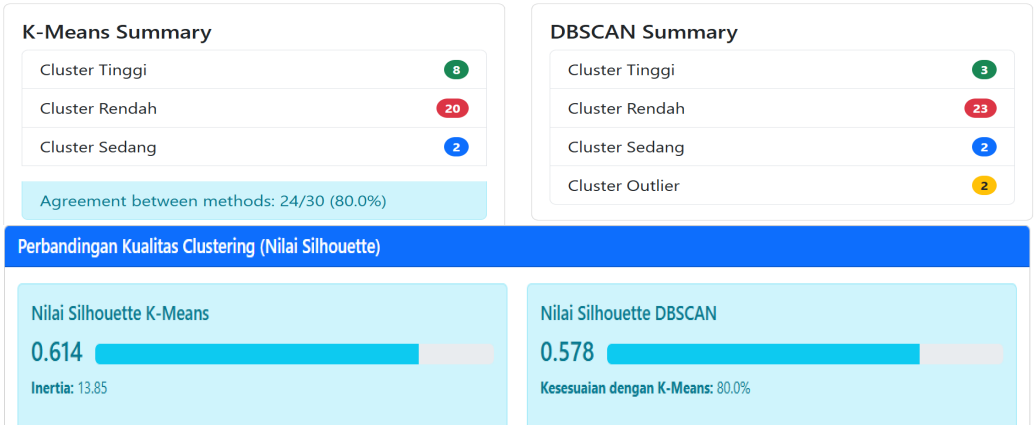

Figure 5 Summary of K-Means and DBSCAN Clustering Comparison of Domestic Tourists

Comparison of the clustering of domestic tourism destinations in Denpasar City in 2023 using K-Means and DBSCAN showed consistency with an agreement rate of 80.0% (24 out of 30 destinations). K-Means produced three clusters (low: 20 destinations, medium: 2 destinations, high: 8 destinations) with a Silhouette value of 0.614 and inertia of 13.85, indicating better clustering quality with tighter and more defined clusters. In contrast, DBSCAN produced three clusters (low: 23 destinations, medium: 2 destinations, high: 3 destinations) with 2 destinations as outliers and a lower Silhouette value (0.578), although it was able to detect unique data patterns and outliers not identified by K-Means.

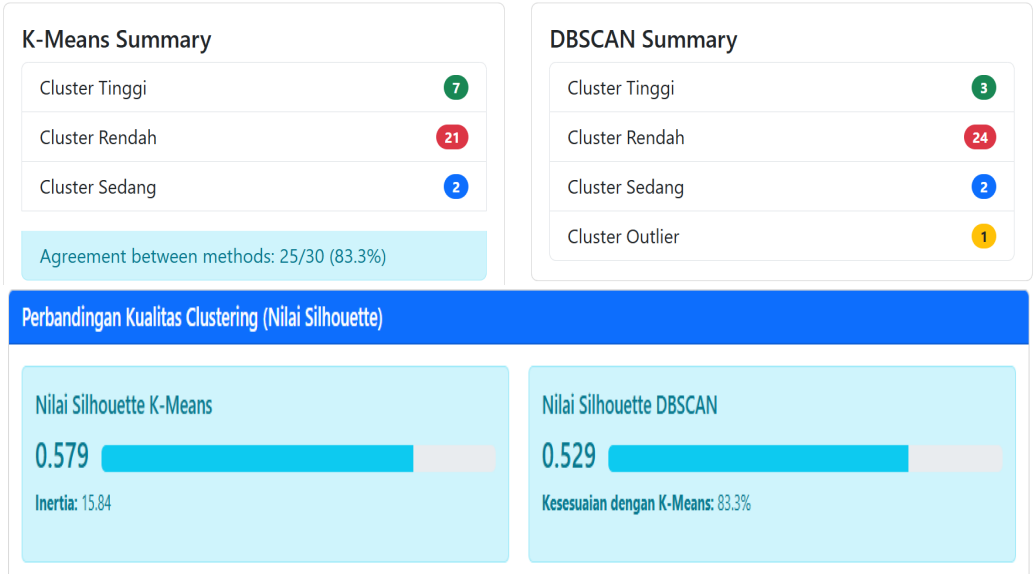

Figure 6 Summary of K-Means and DBSCAN Clustering Comparison of combined Tourists

Comparison of the clustering of combined tourism destinations in Denpasar City in 2023 using K-Means and DBSCAN showed consistency with an agreement rate of 83.3% (24 out of 30 destinations). K-Means produced three clusters (low: 21 destinations, medium: 2 destinations, high: 7 destinations) with a Silhouette value of 0.579 and inertia of 15.84, indicating better clustering quality with tighter and more defined clusters. In contrast, DBSCAN produced three clusters (low: 24 destinations, medium: 2 destinations, high: 3 destinations) with 1 destinations as outliers and a lower Silhouette value (0.578), although it was able to detect unique data patterns and outliers not identified by K-Means.

5. Conclusion

Clustering tourism destinations in Denpasar City in 2023 using K-Means and DBSCAN was successfully conducted with data on tourist visits and facilities from the Tourism Office. The data was pre-processed, including median imputation and feature standardization. K-Means produces three clusters (low, medium, high) with the optimal number of Elbow Method, while DBSCAN clusters by density, identifies outliers and irregular clusters using optimized Eps and Min Samples, enabling effective identification of visitation patterns and facility characteristics. Comparison of clustering evaluation results shows that K-Means is more optimal in forming a predetermined number of clusters, with the highest Silhouette Score values of 0.654 for foreign tourists, 0.614 for domestic tourists, and 0.579 for combined tourists, indicating good cluster density. In contrast, DBSCAN is superior in handling data with irregular distribution and detecting outliers, with the highest Silhouette Score values of 0.681 for foreign tourists, 0.578 for domestic tourists, and 0.529 for combined tourists. The agreement rate between the two algorithms reached 86.7% for foreign tourists, 80.0% for domestic tourists, and 83.3% for combined tourists, indicating high consistency even though DBSCAN provides additional insights through outlier detection.

References

[1] B. M. Al-Fahmi, E. Rahmawati, and T. Sagirani, “Penerapan K-Means Clustering Pada Pariwisata Kabupaten Bojonegoro Untuk Mendukung Keputusan Strategi Pemasaran,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 9, no. 2, pp. 141–149, 2023, doi: 10.25077/teknosi.v9i2.2023.141-149.

[2] N. Putu *et al.*, “Promosi Destinasi Pariwisata Dan Budaya Kota Denpasar Melalui Aplikasi Permainan,” *SENADA (Seminar Nas. Manajemen, Desain dan Apl. Bisnis Teknol.*, vol. 5, pp. 314–320, 2022, [Online]. Available: https://eprosiding.idbbali.ac.id/index.php/senada/article/view/687

[3] L. Maulida *et al.*, “PENERAPAN DATAMINING DALAM MENGELOMPOKKAN KUNJUNGAN WISATAWAN KE OBJEK WISATA UNGGULAN DI PROV. DKI JAKARTA DENGAN K-MEANS,” *J. Inform. Sunan Kalijaga)*, vol. 2, no. 3, pp. 167–174, 2018.

[4] S. Astiti and R. Harman, “Pengelompokan Destinasi Wisata di Batam Berdasarkan Daya Tarik dan Fasilitas Menggunakan Metode K-Means Clustering,” vol. 5, no. 4, pp. 2005–2012, 2024.

[5] A. A. MAHARDIKA, E. N. KENCANA, I. K. G. SUKARSA, K. JAYANEGARA, I. L. WIJAYAKUSUMA, and I. W. SUMARJAYA, “Klasterisasi Karakteristik Wisatawan Mancanegara Menggunakan Metode K-Means Clustering,” *E-Jurnal Mat.*, vol. 12, no. 2, p. 140, 2023, doi: 10.24843/mtk.2023.v12.i02.p411.

[6] H. Gunawan and V. Purwayoga, “Data Mining Menggunakan Algoritma K-Means Clustering Untuk Mengetahui Potensi Penyebaran Virus Corona Di Kota Cirebon,” *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 1, pp. 1–8, 2022, doi: 10.32736/sisfokom.v11i1.1316.

[7] N. Lusianah, A. Irma Purnamasari, and B. Nurhakim, “Implementasi Algoritma K-Means Dalam Pengelompokan Jumlah Wisatawan Akomodasi Di Jawa Barat”.

[8] K. Gustipartsani, N. Rahaningsih, R. D. Dana, and I. Y. Mustafa, “DATA MINING CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS PADA DATA KUNJUNGAN WISATAWAN DI KABUPATEN KARAWANG,” 2023.

[9] L. F. Marini and C. D. Suhendra, “Penggunaan Algoritma K-Means Pada Aplikasi Pemetaaan Klaster Daerah Pariwisata,” *J. Media Inform. Budidarma*, vol. 7, no. 2, p. 707, 2023, doi: 10.30865/mib.v7i2.5558.

[10] C. Mei Hellyana, “Penerapan Algoritma K-Means Terhadap Kunjungan Wisatawan Asing Di Hotel Berbintang di Indonesia,” *J. Sains dan Manaj.*, vol. 11, no. 1, pp. 67–77, 2023.

[11] Y. Syawali, M. Haikal, H. Rangkuti, and K. A. Mayadi, “PENERAPAN ALGORITMA DBSCAN UNTUK ANALISIS DEMOGRAFIS dan PENGELUARAN PELANGGAN MALL”.

[12] R. Wulandari and W. Yustanti, “Analisis Text Clustering Kebijakan Pembukaan Daerah Wisata pada Masa Pandemi Berbasis Densitas Spasial (DBSCAN),” *J. Emerg. Inf. Syst. Bus. Intell.*, vol. 3, no. 2, pp. 1–10, 2022.

[13] M. Adi, P. Firdaus, and Y. Yamasari, “Implementasi Data Mining Untuk Pengelompokan Kunjungan Wisata Di Kabupaten Mojokerto,” *J. Informatics Comput. Sci.*, vol. 06, 2024.