Mall Customer Segmentation Using DBSCAN with K-Distance Plot

Sayyida Naila Alia[a1], Nyimas Nayla Deswitha[a2], Rendra Gustriansyah[a3]
[a]Informatics Engineering study program, Faculty of Computer Science and Science , Indo Global Mandiri University, Indonesia
e-mail:[1]2023110114@students.uigm.ac.id, [2]2023110106@students.uigm.ac.id , [3]rendra@uigm.ac.id

Abstrak

Segmentasi pelanggan merupakan strategi penting bagi pengelola mall dalam memahami perilaku pengunjung dan merancang program pemasaran yang tepat sasaran. Penelitian ini bertujuan mengimplementasikan dan mengoptimasi algoritma DBSCAN (Density-Based Spatial Clustering of Applications with Noise) pada dataset Mall Customer Segmentation yang terdiri dari 200 data pelanggan dengan lima atribut. Optimasi parameter dilakukan secara sistematis melalui analisis k-distance plot dan eksperimentasi grid search pada 15 kombinasi parameter. Parameter optimal yang diperoleh adalah ε = 0,3 dan minPts = 4, menghasilkan 8 klaster pelanggan dengan persentase noise 11,5% dan Silhouette Coefficient sebesar 0,520 yang mengindikasikan struktur klaster yang cukup baik. Setiap klaster diinterpretasikan secara deskriptif menghasilkan profil bisnis yang bermakna, mulai dari pelanggan muda impulsif hingga segmen ultra-elite yang konservatif. Analisis terhadap 23 data outlier mengungkap pelanggan berpenghasilan tinggi (rata-rata 77,48 k$) dengan perilaku belanja yang sangat heterogen, merepresentasikan segmen potensial bernilai tinggi yang memerlukan pendekatan pemasaran sangat personal. Rekomendasi strategi pemasaran yang spesifik dirumuskan untuk setiap segmen yang teridentifikasi..

Kata kunci: *DBSCAN; segmentasi pelanggan; k-distance plot; pelanggan mall; Silhouette Coefficient.*

Abstract

Customer segmentation is a crucial strategy for mall management in understanding visitor behavior and designing targeted marketing programs. This study aims to implement and optimize the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm on the Mall Customer Segmentation dataset consisting of 200 customer records with five attributes. Parameter optimization was conducted systematically through k-distance plot analysis and grid search experimentation across 15 parameter combinations. The optimal parameters obtained were ε = 0.3 and minPts = 4, producing 8 customer clusters with a noise percentage of 11.5% and a Silhouette Coefficient of 0.520, indicating a reasonable clustering structure. Each cluster was interpreted descriptively to produce meaningful business profiles, ranging from young impulsive buyers to ultra-elite conservative segments. Analysis of 23 outlier data points revealed customers with high income (average 77.48 k$) and highly inconsistent spending behavior, representing a high-value potential segment requiring a highly personalized marketing approach. Specific marketing strategy recommendations were formulated for each identified segment.

Keywords : *DBSCAN; customer segmentation; k-distance plot; mall customer; Silhouette Coefficient.*

1. Introduction

In an era of increasingly fierce business competition, a deep understanding of customer characteristics becomes one of the key success factors for a company. Companies that can recognize and understand their customers' needs and behaviors will have a significant competitive advantage [1]. Customer segmentation, the process of grouping customers based on similar characteristics such as demographics, purchasing behavior, or preferences, enables companies to design more targeted and personalized marketing strategies [2]. In the context of shopping malls, customer segmentation becomes very important given the high variation in

visitor behavior, ranging from loyal buyers to mere window shopping. Without good segmentation, marketing efforts tend to be ineffective [3].

However, the effort to understand each individual customer is not an easy task. Customers have highly diverse characteristics, ranging from age, gender, annual income, to spending habits. These behavioral dynamics constantly change over time, influenced by seasonal factors, trends, and economic conditions. Along with the development of digital technology, the main challenge is utilizing unstructured data, full of noise, and having a non-uniform distribution [4]. Research by Hussein et al. (2024) also confirms that analyzing customer behavior is a very broad and complex area. Phenomena such as customers with high income but low spending score are often found in mall data, thus requiring data analysis methods capable of capturing hidden patterns from complex customer data sets.

So far, various customer segmentation studies have been dominated by the use of the K-Means Clustering algorithm due to its simplicity and computational efficiency. However, K-Means has several significant limitations. First, this algorithm requires manual determination of the number of clusters (k) at the beginning, which is often subjective [5]. Second, K-Means assumes that clusters are spherical and relatively equal in size, whereas in real data, cluster shapes can be very diverse and irregular [2][6]. Third, K-Means is very sensitive to the presence of outliers or noise data, where extreme values can pull the cluster centroid and damage the overall grouping results [7][8]. Based on these limitations, the main problem raised in this study is the suboptimal segmentation of mall customers due to the inability of conventional methods such as K-Means to handle irregular cluster shapes and the presence of outliers.

As an alternative, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm offers several advantages. DBSCAN groups data based on the density of points in an area, thus not requiring the determination of the number of clusters at the beginning and being able to identify clusters of any shape (arbitrary shape) [5][2]. Diyabi et al. (2025) in their research explicitly state that DBSCAN shows advantages in identifying non-spherical clusters, which is a major weakness of K-Means. More importantly, DBSCAN automatically classifies low-density points as noise or outliers, thus being robust to extreme data that can damage clustering results in K-Means[4][8]. Research by Zhang et al. (2025) also developed a better DBSCAN to produce more accurate bank customer segmentation. Therefore, this study chose DBSCAN as the main method for three fundamental reasons: (1) no need to subjectively determine the number of clusters; (2) flexible to any cluster shape; and (3) able to identify outliers as potential customer segments.

Although DBSCAN offers various advantages, research applying it to mall customer segmentation datasets is still limited and shows a major challenge in parameter sensitivity. Wardani et al. (2023) showed that standard DBSCAN produces a low Silhouette Coefficient (0.298) if the parameters are not appropriate. Conversely, Astina et al. (2026) showed that DBSCAN can indeed produce a high Silhouette score (0.666), but produces 21 clusters with 91.3% of data classified as noise, making it less relevant for business applications. Thus, there is a research gap, namely: (1) Most studies only use standard DBSCAN without parameter optimization or hybrid approach. (2) Lack of systematic exploration of epsilon (ε) and minPts parameters on mall customer datasets; (3) The ability of DBSCAN to identify outliers as potential customer segments of business value has not yet been utilized.

Based on this background and research gap, this study aims to apply the DBSCAN algorithm for customer segmentation in shopping malls using the Mall Customer Segmentation dataset from Kaggle containing 200 customers with attributes of age, gender, annual income, and spending score. Specifically, this study will: (1) systematically experiment with epsilon (ε) and minPts parameters using k-distance plot analysis; (2) simulate cluster formation based on the optimal parameter variations found; (3) evaluate cluster quality using the Silhouette Coefficient metric; (4) identify and analyze outliers as a potential segment; and (5) interpret the characteristics of each customer segment along with marketing strategy recommendations. This research is expected to provide theoretical benefits in enriching the literature on DBSCAN application for customer segmentation, particularly in handling irregular cluster shapes and

outlier identification [1][6]. Practically, the results of this study can be used by mall managers to design more personal and efficient marketing strategies based on the profiles of the formed customer segments.

2. Research Method / Proposed Method

This study uses a quantitative experimental approach with unsupervised machine learning methods, specifically the DBSCAN algorithm, to analyze shopping mall customer segmentation. The research flow consists of five main interconnected stages, namely: (1) dataset collection and understanding; (2) data preprocessing; (3) DBSCAN parameter optimization; (4) clustering modeling; and (5) cluster evaluation and interpretation. The overall framework of the research methodology is illustrated in Figure 1.

Figure 1. Research Methodology Flow Framework

2.1. Dataset Collection and Understanding

This study uses secondary data in the form of the Mall Customer Segmentation dataset obtained from the Kaggle public repository [16]. This dataset represents actual customer data from a shopping mall and is commonly used as a benchmark in customer segmentation research. The dataset consists of 200 records with five attributes as detailed in Table 1 [9].

Table 1. Description of Attributes of the Mall Customer Segmentation Dataset

Attribute	Data Type	Description	Value Range
CustomerID	Integer	Unique customer identifier	1-200
Gender	Category	Customer's gender (Male/Female)	Male, Female
Age	Integer	Customer's age in years	18-70 years
Annual Income (k$)	Integer	Customer's annual income in thousands of dollars	15-137 k$
Spending Score (1-100)	Integer	Spending score assigned by mall management (1=low, 100=high)	1-99

All computational processes were performed using the Python 3.10 programming language in the Google Colaboratory environment. The main libraries used include pandas and numpy for data manipulation, scikit-learn for modeling and evaluation, and matplotlib and seaborn for visualization.

2.2. Data Preprocessing

The data preprocessing stage aims to ensure data quality before being input into the clustering model [10]. Preprocessing was carried out through the following three steps. (1) Data Quality Check. A check was performed to detect missing values and duplicate data using the isnull() and duplicated() functions from the pandas library. If missing values were found, imputation would be performed using the mean (for numerical data) or mode (for categorical data). (2) Feature Selection. The CustomerID attribute was removed because it is a unique identifier with no informative value for the clustering process. The Gender attribute was encoded using Label Encoding (Male = 1, Female = 0) so it could be processed by the algorithm. Based on business relevance and initial exploratory results, two main attributes were selected for the clustering process: Annual Income and Spending Score, because they best represent patterns of customer spending behavior. (3) Data Standardization. Given the significant differences in

units and value ranges between Annual Income (15-137 k$) and Spending Score (1-99), Standardization was performed using the Z-score method via the scale() function in R, with the formula:

$$z=(x-\mu)/\sigma z=(x-\mu)/\sigma \quad (1)$$

where x is the original value, μ is the feature mean, and σ is the feature standard deviation. Standardization ensures both features have comparable scales so that no single feature dominates the distance calculation.

2.3 Clustering (DBSCAN)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that works by grouping data points that are densely packed together. This algorithm has two main parameters: (1) ε (Epsilon): the radius of a data point's neighborhood. All points within radius ε of point p are considered neighbors of p. (2) minPts: the minimum number of points (including the point itself) that must be within radius ε for a point to be classified as a core point. The recommended minPts value is ≥ dataset dimensions + 1, with a minimum value of 3. Based on these two parameters, each data point is classified into three categories: (1) Core Point: a point that has at least minPts neighbors within ε radius, including itself. Formally, point p is a core point if

$$|N\varepsilon(p)| \geq \text{minPts}, \text{ where } N\varepsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \varepsilon\} \quad (2)$$

(2) Border Point: a point that does not meet the minPts requirement (not a core point), but lies within ε radius of at least one core point. (3) Noise Point: a point that is neither a core point nor a border point. This point is labeled -1 and considered an outlier. The distance between data points is calculated using Euclidean distance. For two data points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ in two-dimensional space, Euclidean distance is defined as:

$$\text{dist}(p,q)=(((p1-q1)2+(p2-q2)2)dist(p,q)= (((p1-q1)2+(p2-q2)2) \quad (3)$$

The DBSCAN clustering process starts from an unvisited data point, then recursively expands the cluster by adding all points that are density-reachable from an existing core point. This process repeats until all data points have been processed.

2.4. Parameter Optimization (Grid Search)

Determination of optimal DBSCAN parameters was carried out through two systematic stages. (1) K-Distance Plot Analysis. The k-distance plot is used to determine candidate ε values objectively. This method calculates the distance of each data point to its k-th nearest neighbor using the k-Nearest Neighbors (kNN) algorithm [11], then sorts these distances in descending order and visualizes them in a graph. The k value is synchronized with the minPts value to be tested. The elbow point on the graph, where the curve starts to flatten significantly, indicates the data density threshold and is set as the candidate for the optimal ε value. (2) Experimental Grid Search. To validate ε candidates from the k-distance plot and comprehensively explore parameter combinations, a grid search experiment was conducted on all combinations of ε and minPts values as shown in Table 2. Each combination was run on the standardized data and evaluated based on three criteria: (a) number of clusters formed; (b) percentage of noise data; and (c) Silhouette Coefficient value. The best parameter combination was selected based on the highest Silhouette Coefficient value with a noise percentage below 10% and a number of clusters that is business-meaningful (3-10 clusters)[12].

Table 2. Range of Parameter Values Tested in the Grid Search

Parameter	Nilai yang Diuji
ε (Epsilon)	0,1-0,5
minPts	3-5
Total Kombinasi	15 kombinasi (5×3)

Total parameter combinations tested: 15 combinations.

2.5. Evaluation and Interpretation of Cluster Results

The quality of the clustering results was evaluated using the Silhouette Coefficient, a metric that measures how well a data point is clustered within its own cluster compared to the nearest neighboring cluster [13]. The Silhouette Coefficient for each data point *i* is calculated using the formula:

$$S(i)=(b(i)-a(i))/\max\{a(i),b(i)\}\mathbf{S}(\mathbf{i})=(\mathbf{b}(\mathbf{i})-\mathbf{a}(\mathbf{i}))/\max\{\mathbf{a}(\mathbf{i}),\mathbf{b}(\mathbf{i})\} \quad (4)$$

where: *a(i)* = the average distance from point *i* to all other points in the same cluster (measuring cluster cohesion). *b(i)* = the average smallest distance from point *i* to points in the nearest neighboring cluster (measuring separation between clusters).

The Silhouette Coefficient value ranges from -1 to 1. The average Silhouette Coefficient of all data points belonging to clusters (excluding noise) is used as an overall indicator of clustering quality. Guidelines for interpreting the Silhouette Coefficient value are presented in Table 3.

Table 3. Guidelines for Interpreting the Silhouette Coefficient

Silhouette Value Range	Interpretation
0.71 - 1.00	Strong cluster structure
0.51 - 0.70	Reasonable cluster structure
0.26 - 0.50	Weak cluster structure
< 0.25	No substantial structure

After clusters are formed, each cluster is interpreted descriptively based on the average values of Age, Annual Income, Spending Score, and Gender distribution using the original data before standardization. The interpretation process follows the approach commonly used in data-driven customer segmentation, namely: (1) determining a segment name or label that reflects its dominant characteristics; (2) identifying the typical spending behavior of the segment; and (3) formulating relevant and measurable marketing strategy recommendations for mall managers based on the profile of the identified segment.

3. Literature Study

Customer segmentation in shopping malls requires methods capable of handling complex data [1]. K-Means dominates research but is sensitive to outliers and assumes spherical cluster shapes [2][5]. Research in JBI by Saputra et al. [14] also confirmed these weaknesses. As an alternative, DBSCAN [15] excels in noise resistance and the ability to form arbitrary shape clusters without needing to determine the number of clusters in advance [2][4]. Yolandari et al. [16] compared K-Means and DBSCAN on travel review data, resulting in a DBSCAN SI of 0.272 and DBI of 0.838. However, the application of DBSCAN on mall datasets is still limited. Wardani et al. (2023) reported a low SI (0.298) due to incorrect parameters, while Astina et al. (2026) obtained a high SI (0.666) but 91.3% of data became noise [2][7]. Putra et al. [17] compared OPTICS and DBSCAN on mall customers, showing DBSCAN is too strict with a small epsilon and less accurate with a large epsilon. Kumar [18] used RFM and Bisecting K-Means for retail segmentation, emphasizing the importance of proper parameter selection. Ibadirachman et al. [19] optimized DBSCAN parameters with Differential Evolution for bank transaction anomaly detection, achieving 98.41% accuracy and an SI of 0.7916. Lee & Kim [20] developed HTC-DBSCAN with automated parameter tuning based on k-distance graph and moving average.

This study identifies five gaps: (1) ε and minPts parameters are determined by trial-and-error without a systematic approach [2],[4],[7]; (2) noise overproduction up to >90% of data discarded [17]; (3) the business value of outliers as a potential segment is ignored, whereas [19]

shows outlier identification has business value; (4) simultaneous exploration of age, income, and spending score in malls is still minimal, unlike Kumar's RFM approach[18]; (5) there is no validation of SI stability against parameter fluctuations in the n=200 dataset, while [20] developed a similar methodology for trajectory data. Different from previous studies, this study proposes DBSCAN parameter optimization using k-distance plot and Grid Search, and utilizing outliers as a business-valuable segment. This approach was chosen because it eliminates parameter subjectivity, maximizes noise identification, and provides a structured experimental framework. Contains all libraries used as references in this study. All references in the text are written with an [x] sign. If there are figures and tables, they should be presented with the names of tables and figures that are equipped with sequential numbers.

4. Result and Discussion

4.1. Data Preprocessing Results

The preprocessing stage was carried out through three steps. First, the CustomerID attribute was removed because it only serves as a unique identifier with no informative value. Second, the Gender attribute was encoded using Label Encoding (Female = 0, Male = 1). Third, the two attributes used for the clustering process, namely Annual Income and Spending Score, were standardized using StandardScaler, resulting in a distribution with a mean $\mu = 0$ and standard deviation $\sigma = 1$. Standardization was necessary due to the significant differences in units and value ranges between the two attributes (Annual Income: 15-137 k$; Spending Score: 1-99) so that no attribute would dominate the Euclidean distance calculation in DBSCAN.

4.2. DBSCAN Clustering Results

The DBSCAN model was run on the standardized data using the optimal parameters $\varepsilon = 0.3$ and minPts = 4. The clustering results formed 8 clusters with 23 noise data points (11.5%) and a Silhouette Coefficient value of 0.520. This value falls within the range of 0.51–0.70, which indicates a reasonable cluster structure. Visualization of the clustering results in the two-dimensional space of Annual Income vs. Spending Score is presented in Figure 2. From the visualization, it can be seen that the clusters formed have irregular shapes and varied distributions across the feature space, reflecting DBSCAN's advantage in detecting complex data distribution patterns.

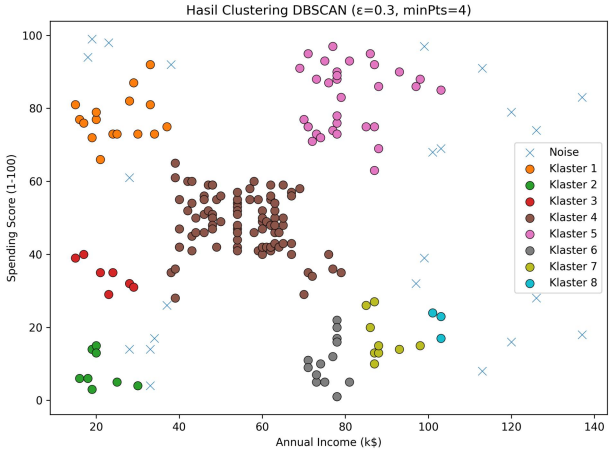

Figure 2. Visualization of DBSCAN Clustering Results (ε=0.3, minPts=3)

4.3. Parameter Optimization (Grid Search)

The determination of the ε parameter value was carried out through k-distance plot analysis with k = 4 (synchronized with the planned minPts = 4). The results of the k-distance plot are presented in Figure 3. Based on the graph, the elbow point was identified at a distance of about 0.5, but considering the wide range of distance variation (0.06-1.0), three candidate ε values, namely 0.1; 0.3; and 0.5, were set for further testing through grid search experimentation.

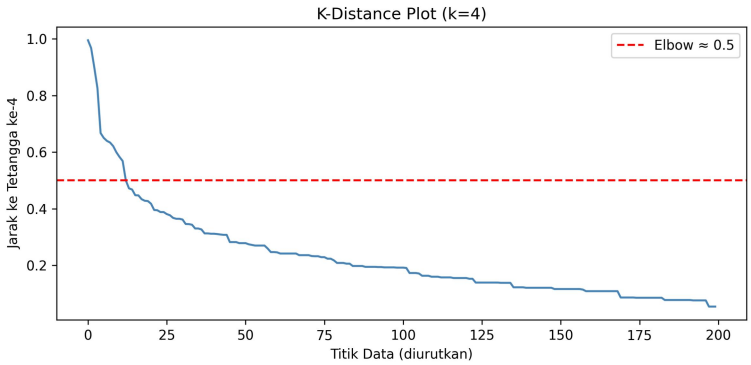

Figure 3. K-Distance Plot (k=4)

The grid search experimentation was carried out on combinations of ε ∈ {0.1-0.5} and minPts ∈ {3-5}, resulting in 15 parameter combinations. Each combination was evaluated based on the number of clusters formed, the percentage of noise data, and the Silhouette Coefficient value. The complete results are presented in Table 2.

Table 2. Results of DBSCAN Parameter Grid Search Experimentation

ε	minPts	Number of Clusters	Noise (%)	Silhouette
0.1	3	14	60.5	0.561
0.1	4	7	75.5	0.528
0.1	5	4	83.0	0.563
0.2	3	13	22.0	0.470
0.2	4	5	36.5	0.617
0.2	5	7	38.5	0.586
0.3	3	9	7.0	0.472
0.3*	**4**	**8**	**11.5**	**0.520**
0.3	5	7	17.5	0.524
0.4	3	4	5.0	0.395
0.4	4	3	7.0	0.458
0.4	5	4	7.5	0.478
0.5	3	2	3.5	0.389
0.5	4	2	4.0	0.388
0.5	5	2	4.0	0.388

Note: (*) Selected parameter

Based on Table 2, the combination of ε = 0.3 and minPts = 4 (experiment number 5) was selected as the optimal parameter with three main considerations. (1) the Silhouette Coefficient value of 0.520 falls into the reasonable structure category (0.51–0.70). Although the Silhouette value at ε = 0.2 with minPts = 4 is higher (0.617), the noise percentage reaches 36.5% (73 out of 200 data points), which is impractical for business segmentation applications because almost four out of ten customers are ungrouped. (2) the noise percentage of 11.5% (23 data points) is still within practically acceptable limits. This figure is much better compared to the ε = 0.1 group which produced noise up to 83.0% or ε = 0.2 with a minimum noise of 22.0%. On the other hand, ε = 0.4 and ε = 0.5 did produce lower noise (5–7.5%), but only formed 2–4 clusters with a Silhouette below 0.5 (weak structure category), thus less meaningful for detailed segmentation. (3) the eight clusters formed are sufficiently representative to produce meaningful segmentation for mall managers, not too many to be difficult to implement, and not too few to lose important information about variations in customer behavior.

As a comparison, the combination of ε = 0.3 and minPts = 3 produced a lower Silhouette (0.472/weak structure), while ε = 0.3 and minPts = 5 produced higher noise (17.5%) with an insignificant increase in Silhouette (0.524). Therefore, ε = 0.3 and minPts = 4 is the most balanced choice between statistical quality (Silhouette), business practicality (noise percentage), and segmentation depth (number of clusters).

4.4. Customer Segment Profile and Interpretation

The profile of each cluster was analyzed based on the average values of Age, Annual Income, Spending Score, and Gender distribution. All values presented use the original data before standardization to make business interpretation easier. The complete profile of each cluster is presented in Table 3.

Table 3. Profile of Clusters Resulting from DBSCAN Clustering

Cluster	n	Female	Male	Average Age	Average Annual Income (k$)	Average Spending Score
0 (K1)	16	9	7	23.75	25.06	77.31
1 (K2)	8	4	4	48.38	20.88	8.25
2 (K3)	7	4	3	36.71	22.43	34.43
3 (K4)	88	53	35	42.88	55.23	48.58
4 (K5)	32	18	14	32.81	80.50	82.56
5 (K6)	14	5	9	37.79	75.93	10.07
6 (K7)	9	2	7	46.56	88.78	17.00
7 (K8)	3	3	0	43.00	102.33	21.33
Outlier	*23*	*14*	*9*	*36.78*	*77.48*	*53.09*

Note: n = number of cluster members; average values using original data (before standardization)

Based on Table 3, the interpretation of each cluster is as follows.

- **Cluster 1 (n=16)** are young customers (average 23.75 years) with low income (25.06 k$) and high spending score (77.31), characterizing impulsive spending behavior dominant among young people.
- **Cluster 2 (n=8)** are middle-aged customers (48.38 years) with very low income (20.88 k$) and very low spending score (8.25), indicating a very thrifty customer group that rarely transacts.
- **Cluster 3 (n=7)** are adult customers (36.71 years) with low income (22.43 k$) and low-medium spending score (34.43), depicting a group that shops but is financially limited.
- **Cluster 4 (n=88)** is the largest cluster covering 44.0% of total customers, consisting of adult customers (42.88 years) with medium income (55.23 k$) and medium spending score (48.58). This cluster represents the mall's core customers with stable and moderate spending behavior.
- **Cluster 5 (n=32)** consists of young-adult customers (32.81 years) with high income (80.50 k$) and high spending score (82.56), making it the most business-valuable segment because it combines high purchasing power with a high willingness to spend.
- **Cluster 6 (n=14)** are adult customers (37.79 years) with high income (75.93 k$) but a very low spending score (10.07), depicting a group that has purchasing power but is very selective in spending.
- **Cluster 7 (n=9)** are middle-aged customers (46.56 years) with very high income (88.78 k$) and low spending score (17.00), reflecting a conservative elite segment; majority male (7 out of 9 people).
- **Cluster 8 (n=3)** is the smallest but most elite segment, consisting of middle-aged customers (43.00 years) with the highest average income of all clusters (102.33 k$), all female, with a low spending score (21.33) indicating very planned and selected spending behavior.

5. Conclusion

This study successfully implemented and optimized the DBSCAN algorithm for mall customer segmentation using the Mall Customer Segmentation dataset consisting of 200 data points. DBSCAN parameter optimization using a combination of k-distance plot analysis and grid search experimentation on 15 parameter combinations successfully identified the optimal parameters ε = 0.3 and minPts = 4, which produced a Silhouette Coefficient value of 0.520 (reasonable structure category), a noise percentage of 11.5%, and a business-meaningful number of clusters, namely 8 clusters. DBSCAN successfully formed 8 customer clusters with diverse and business-meaningful profiles, where Cluster 4 (n=88, 44% of total data) is the largest segment representing the mall's core customers with medium income and moderate spending behavior, Cluster 5 (n=32) is the most business-valuable segment with a combination of high income (80.50 k) andhigh spending score(82.56), and Cluster8(n=3) is the ultra−elite segment with the high estaverage income(102.33k) *and high spending score*(82.56), *and Cluster*8(*n*=3) *is the ultra−elite segment with the highes taverageincome*(102.33*k*). Furthermore, analysis of the 23 outlier data points (11.5%) revealed a group of customers with relatively high income (average 77.48 k$) and highly heterogeneous spending behavior that could not be grouped into any cluster, and unlike previous studies that discarded outliers, this study interprets this group as a high-value potential segment requiring a highly personalized marketing approach.

References

[1] R. M. Hussein, K. W. Khaw, A. Gaber, and X. Chew, “JSCDM Examining the Behaviors and Preferences of Online Shopping Customers Using Clustering Techniques,” *JSCDM*, vol. 5, no. 1, pp. 104–121, 2024, doi: https://doi.org/10.30880/jscdm.2024.05.01.009.

[2] S. Dyah, K. Wardani, A. S. Ariyanto, M. Umroh, and D. Rolliawati, “Perbandingan hasil metode clustering k-means, db scanner & hierarchical untuk analisa segmentasi pasar,” *JIKO*, vol. 7, no. 2, pp. 191–201, 2023, doi: 10.26798/jiko.v7i2.796.

[3] H. ŞENTÜRK, E. GEÇİCİ, and S. ALP, "Customer Segmentation With Clustering Methods In The Retail Industry," *İstanbul Aydın Üniversitesi Sos. Bilim. Derg.*, vol. 16, no. 4, pp. 551–573, 2021, doi: 10.17932/IAU.IAUSBD.2021.021/iausbd_v16i4004.

[4] R. Chavhan, P. Dutta, N. Samant, and S. Kar, "Data-driven strategic customer segmentation considering cart abandonment behavior: Insights from e-grocery delivery platforms," *Inf. Sci. (Ny).*, vol. 718, pp. 1–27, Nov. 2025, doi: 10.1016/j.ins.2025.122327.

[5] P. K. Rintonga and M. S. Hasibuan, "Analisis Perbandingan Silhouette dengan Elbow pada Algoritma," *METIK J.*, vol. 9 no.1, pp. 64–71, 2025, doi: 10.47002/metik.v9i1.1027.

[6] N. Diyabi, D. Çakır, Ö. M. Gül, T. Aytekin, and S. Kadry, "Evaluating Customer Segmentation Techniques in the Retail Sector," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 9, no. 3, pp. 175–190, 2025, doi: 10.9781/ijimai.2025.05.001.

[7] P. N. P. Astina, W. Supriana, and M. S. Bimantara, "Developing a Patient-Centric Healthcare IoT Platform with Blockchain and Smart Contract Data Management," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, pp. 295–306, 2024, doi: 10.14569/IJACSA.2024.01504115.

[8] X. Yan, Y. Li, F. Nie, and R. Li, "Bank Customer Segmentation and Marketing Strategies Based on Improved DBSCAN Algorithm," *Appl. Sci.*, vol. 15, no. 6, pp. 1–25, Mar. 2025, doi: 10.3390/app15063138.

[9] "Mall Customer Segmentation Dataset." Accessed: Jun. 02, 2026. [Online]. Available: https://www.kaggle.com/datasets/ineubytes/mall-customer-segmentation-dataset

[10] R. Gustriansyah, J. Alie, and N. Suhandi, "A Hybrid Machine Learning Model for Market Clustering," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 6, pp. 18824–18828, 2024, doi: 10.48084/etasr.9259.

[11] H. Gunawan, A. Chusyairi, and M. I. Saputra, "Penerapan K-Nearest Neighbor Dengan Metode Euclidean Distance Untuk Klasifikasi Tingkat Ketebalan Cat Di PT XYZ," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 2, pp. 59–72, 2025, doi: 10.65258/jutekom.v1.i2.12.

[12] S. Ernawati and R. Wati, "Evaluasi Performa Kernel SVM dalam Analisis Sentimen Review Aplikasi ChatGPT Menggunakan Hyperparameter dan VADER Lexicon," *J. Buana Inform.*, vol. 15, no. 01, pp. 40–49, 2024, doi: 10.24002/jbi.v15i1.7925.

[13] T. Akbar, G. M. Tinungki, and Siswanto, "Performance Comparison of K-Medoids and Density Based Spatial Clustering of Application With Noise Using Silhouette Coefficient Test," *Barekeng*, vol. 17, no. 3, pp. 1605–1616, 2023, doi: 10.30598/barekengvol17iss3pp1605-1616.

[14] R. A. Saputra, J. Nangi, I. P. Ningrum, M. F. Almaliki, and L. O. R. A. Pratama, "Deteksi Uang Palsu Rupiah dengan Menggunakan Metode Deteksi Tepi Laplacian of Gaussian (LoG) dan Algoritma K-Means Clustering," *J. Buana Inform.*, vol. 13, no. 02, pp. 85–92, Oct. 2022, doi: 10.24002/jbi.v13i02.5448.

[15] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, Sep. 2017, doi: 10.1145/3068335.

[16] N. A. Yolandari, L. E. Butarbutar, G. H. Rajagukguk, M. F. Zulfi, and F. Ramadhani, "Analisis Perbandingan K-Means Dan Dbscan Dalam Pengelompokan Data Travel Review Ratings Menggunakan Evaluasi Silhouette Index Dan Davies-Bouldin Index," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3, pp. 470–481, Jul. 2025, doi: 10.23960/jitet.v13i3.6884.

[17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Klasterisasi Dan Segmentasipelanggan Dengan Menggunakan Algoritma Ordering Points To Identify The Clustering Structure (Optics) Dan Dbscan," *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, Jun. 1999, doi: 10.1145/304181.304187.

[18] N. Kumar, "Intelligent customer segmentation: unveiling consumer patterns with machine learning," *J. Umm Al-Qura Univ. Eng. Archit.*, vol. 16, no. 3, pp. 774–783, Sep. 2025, doi: 10.1007/s43995-025-00180-7.

[19] J. Liu, B. Liang, and W. Ji, "An anomaly detection approach based on hybrid differential evolution and K-means clustering in crowd intelligence," *Int. J. Crowd Sci.*, vol. 5, no. 2, pp. 129–142, Aug. 2021, doi: 10.1108/IJCS-07-2020-0013.

[20] D. Lee and J. Kim, "Development of HTC-DBSCAN: A Hierarchical Trajectory Clustering Algorithm with Automated Parameter Tuning," *Appl. Sci.*, vol. 14, no. 23, pp. 1–24,

Nov.2024, doi: 10.3390/app142310995.