The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4135

Implementation of Density-Based Spatial Clustering of Applications with Noise and Fuzzy C – Means for Clustering Car Sales

Sephia Nazwa Auliani[1], Mustakim[2], Rice Novita[3], M.Afdal[4]
12050321905@students.uin-suska.ac.id[1], mustakim@uin-suska.ac.id[2], ricenovita@uin-suska.ac.id[3], m.afdal@uin-suska.ac.id[4]
[1,2,3,4] Sultan Syarif Kasim State Islamic University of Riau

Article Information

Received : 14 Jun 2024
Revised : 10 Jul 2024
Accepted : 25 Jul 2024

Keywords

Data Mining, Clustering, Car Sales, Density-Based Spatial Clustering of Applications with Noise and Fuzzy C-Means

Abstract

This study compares the performance of two clustering algorithms, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Fuzzy C-Means (FCM), in clustering car sales data at PT. XYZ. The dataset, comprising sales transactions from 2020 to 2023, includes information about vehicles, customers, and transactions. Preprocessing methods such as data transformation and normalization were applied to prepare the data. The results indicate that DBSCAN produces clusters with better validity, measured using the Silhouette Score, compared to FCM. Specifically, DBSCAN achieves the highest Silhouette Score of 0.7874 in cluster 2, while FCM reaches a maximum score of 0.3666 in cluster 3. Thus, DBSCAN proves to be more optimal for clustering car sales data at PT. XYZ, highlighting its superior performance in terms of cluster validity.

A. Introduction

The geographical conditions of Indonesia, consisting of islands, necessitate transportation tools for carrying both people and goods, which are expected to boost the country's economy [1]. The development of transportation has progressed rapidly along with technological advancements in the water, land, and air transportation sectors. In Indonesia itself, due to its geographical conditions, land transportation is the most dominant [2]. Generally, land transportation consists of several types of vehicles, with the most commonly used being two-wheeled vehicles or motorcycles, and four-wheeled vehicles or cars [3]. Cars are considered to have a higher level of safety compared to two-wheeled vehicles [4].

The lifestyle of Indonesian society continues to evolve over time, and one notable aspect of this change is the growing interest in cars [5]. As a country in Southeast Asia, Indonesia has become a rapidly growing automotive market in recent years [6]. Every year, car sales keep increasing, reflecting a strong drive from the public to own personal vehicles [6]. This aligns with the development of the automotive industry, which is increasingly enlivening the market with various new car brands [7].

The number of four-wheeled vehicles in Indonesia has increased each year. According to data from Korlantas Polri as of February 2023, here is the breakdown of the four-wheeled motor vehicle population: Passenger Cars amount to 19,177,264 units, Goods Vehicles to 5,700,000 units, Buses to 213,788 units, and Special Vehicles to 85,113 units [8]. This total number of four-wheeled vehicles reflects a significant increase compared to the 2018 data from the Central Statistics Agency (BPS), which recorded a total of 26,757,713 vehicles, consisting of 16,440,987 passenger cars, 7,778,544 goods vehicles, and 2,538,182 buses [9]. This rise indicates continued growth in the use of four-wheeled vehicles in Indonesia. As the four-wheeled vehicle industry develops further, it is likely to lead to the emergence of related businesses such as dealerships, workshops, and showrooms [10].

In 2023, PT. XYZ held the top position in sales with 336,777 units, showing an increase from 331,797 units in 2019. PT. ABC ranked second with 188,000 units, up from 177,284 units in 2019. PT. DEF was third with sales of 138,967 units, a slight increase from 137,339 units in 2019. PT. GHI took fourth place with sales of 77,416 units, experiencing a decline compared to 119,011 units in 2019. Meanwhile, PT. JKL was in fifth place with 81,057 units, down from 100,383 units in 2019 [11]. In 2023, the Riau Province saw significant growth. Total car sales in Riau in 2023 reached 30,500 units, showing growth from previous years. PT. XYZ continued to dominate the market with 10,100 units sold, maintaining its top position in the province. In second place, PT. GHI managed to sell 7,200 units, showing a significant increase compared to 2019. PT. DEF took third place with 5,600 units sold, followed by PT. ABC with 4,500 units sold. PT. JKL remained in fifth place with 3,100 units sold [12]. Overall, the automotive market in Indonesia also recorded strong performance despite a slight decline compared to the previous year. National car sales reached around 1 million units in 2023, slightly down from 1.048 million units in 2022 [13].

Starting its business in 1937 as PT. XYZ, and officially founded by Kiichiro PT. XYZ in XYZ City, Japan, in 1937, the company has continued to grow and develop

into the modern era [14]. PT. XYZ operates in the automotive and industrial equipment sectors and has been globally recognized as a high-quality brand, consistently creating high-value and advanced products [14][15]. PT. XYZ is the main distributor of PT. XYZ four-wheeled vehicles (cars) in Indonesia, operating directly under the auspices of PT. XYZ [16]. PT. XYZ offers a variety of PT. XYZ car models such as the Avanza, Fortuner, Innova, Yaris, and others. PT. XYZ was established on April 12, 1971, and has a network of dealerships across Indonesia [17], including in Pekanbaru.

Based on the data obtained, PT. XYZ has achieved an average monthly sales of 200 to 250 PT. XYZ cars. Over a span of two years, sales can exceed 5,000 units. This sales data can be processed into knowledge and recommendations that can be used by management for marketing purposes. For instance, transaction data such as addresses, time, and the type or model of vehicle purchased by consumers can be processed by PT. XYZ marketing team to make marketing efforts more efficient and targeted [18]. This also allows for the grouping of consumer interests in PT. XYZ car models, which plays a crucial role in the promotion and marketing process.

As the automotive industry evolves and competition intensifies, developers are required to identify patterns that can enhance product sales and marketing [19]. An effective way to achieve this is through the utilization of transaction data. The use of information systems in a highly competitive environment is one of the external challenges faced by the company [20]. Data mining, also known as Knowledge Discovery in Database (KDD), involves activities related to data collection and the use of historical data to discover knowledge, information, patterns, or relationships in large datasets [21]. The output from data mining can be used as an alternative for decision-making or to improve future decision-making processes [22].

To address this issue, the application of data mining methods is required [20]. Some popular and effective algorithms for data clustering are the FCM algorithm and the DBSCAN algorithm [23]. The FCM algorithm is a clustering method that divides data into several groups based on membership degrees, where each data point can belong to more than one cluster with a certain level of membership [24]. On the other hand, DBSCAN is a clustering algorithm that can find clusters of arbitrary shapes and handle noise effectively [25].

The application of unsupervised learning in this context aims to identify potential customer segments to market products more effectively and serves as a reference to evaluate market segments based on time and location of marketing to maximize profits [26]. By leveraging these clustering algorithms, PT. XYZ can better understand the preferences and behaviors of their customers, allowing for more targeted and efficient marketing strategies. For example, using the FCM algorithm [27], PT. XYZ can identify overlapping customer segments that may have shared interests in multiple types of vehicles, thus tailoring marketing campaigns to address these shared interests [28]. Meanwhile, DBSCAN can help in identifying distinct clusters of customers who might be geographically dispersed or have unique purchasing patterns, enabling the company to focus their resources on high-value customer segments and regions [29]. By implementing these data mining techniques, PT. XYZ can gain valuable insights from their sales data,

improve their decision-making processes, and enhance their competitive edge in the automotive market.

Several studies have compared the FCM algorithm and the DBSCAN algorithm [30]. Both are unsupervised learning methods used for clustering data, but they have different approaches and applications [25]. This study titled "Clustering Medical Image Data Also Comparing the Performance of FCM and DBSCAN"compares the performance of two clustering algorithms, FCM and DBSCAN, in segmenting ultrasound images of fetal heads. The results indicate that FCM produces better segmentation outcomes than DBSCAN, especially when using fewer clusters. FCM demonstrates superior capability in accurately representing color distribution in grayscale images. However, this method requires a longer execution time compared to DBSCAN. Although DBSCAN is faster, its segmentation results are less optimal. Therefore, for the segmentation of fetal head ultrasound images, FCM is more effective in achieving high-quality segmentation, despite needing more processing time [31].

Based on previous issues and research, it is concluded that this study will utilize FCM(FCM) and DBSCAN (DBSCAN) algorithms, as both have relative advantages depending on the data clustering context. This research will employ transactional data from PT. XYZ Sutomo to generate insights that can be used by company leaders in making strategic decisions.

B. Research Method

This section outlines the research methodology for comparing the performance of DBSCAN and FCM algorithms in clustering car sales data at PT. XYZ.

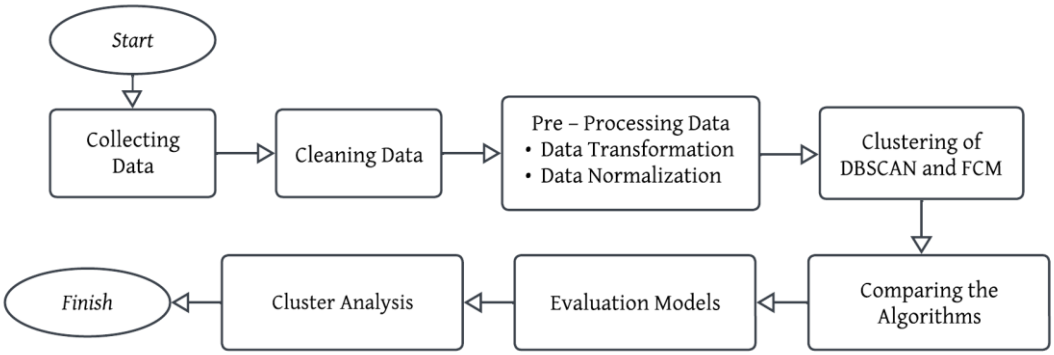

Figure1. Research Method

1) Data Collection

PT. XYZ has sales transaction data from the years 2020 to 2023. This data includes information about vehicles, customers, and transactions. The analysis of this data will use clustering methods to understand customer grouping patterns based on purchasing characteristics.

2) Pre – processing Data

Pre-Processing Data is a crucial stage in data analysis that includes several key techniques to prepare raw data into a more suitable format for analysis. One of

these techniques is Data Transformation, which modifies raw data through methods like aggregation and converting categorical data into numerical form, ensuring the data is more consistent and easier to interpret. Additionally, Data Normalization standardizes the scale of data so that each feature contributes proportionally to the analysis, typically by converting feature values into a uniform range such as between 0 and 1. This normalization process is essential for enhancing the performance of machine learning algorithms by reducing distortions caused by differing scales. By performing pre-processing, transformation, and normalization of data, we can improve the quality and accuracy of analysis results and predictive models [32].

3) Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

The DBSCAN algorithm is a method of data clustering based on density. DBSCAN divides data into clusters based on spatial proximity between data points [33]. The main steps in the DBSCAN algorithm include:

1. Parameter Determination:
   The algorithm begins by setting two main parameters epsilon (ε), which is the maximum distance between two data points that are still considered neighbors, and min_samples, which is the minimum number of data points within a radius ε required to form a cluster.
2. Core Point Identification:
   A data point p is considered a core point if there are at least 'min_samples' data points within a radius ε, including the point p itself. Mathematically, this is expressed as :

   $$|N_\varepsilon(p)| \geq \text{min\_samples} \quad (1)$$

   Where $N\varepsilon(p)$ is the set of points within a radius ε from p, and $d(p,q)$ is the distance between points p and q :

   $$N\varepsilon(p) = \{ q \in D \mid d(p,q) \leq \varepsilon \} \quad (2)$$

3. Connecting Core Points:
   Core points that are close to each other (within a radius ε) are connected to form clusters. Cluster C is formed from a set of core points that are directly or indirectly connected :

   $$C = \{p \mid p\} \quad (3)$$

4. Border Point Identification:
   Data points that are not core points but are within a radius ε from a core point are called border points. These border points are considered part of the cluster of the core point.

5. Noise Identification:
   Data points that do not meet the criteria as core points or border points are considered noise and are not included in any cluster. Data point p is noise if :

   $$|N_\varepsilon(p)| < \text{min\_samples} \quad (4)$$

The implementation of DBSCAN is done by initializing a DBSCAN object with the specified parameters (ε and min_samples), and then applying the fit_predict() method to perform data clustering.

4) Fuzzy C – Means

FCM is a clustering algorithm that divides a dataset into several clusters [34]. The steps performed in the FCM algorithm involve determining the number of clusters, which in this case is set to 11 clusters, specifying the smallest expected error epsilon (ε) as 0.005, setting the maximum iteration value, and initializing the partition matrix values for each data randomly within each cluster. Calculate the sum of each data in the normalized immunization dataset with the initial partition matrix data using the equation :

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \tag{5}$$

$$\mu_{ik} = \frac{\mu_{ik}}{Q_i} \tag{6}$$

From the membership degrees of the 11 clusters, calculate the mean value for each cluster to obtain the cluster centroids. Then, compute the average objective function value between the normalized immunization data and the cluster centroids using the equation :

$$P_t = \sum_{i=1}^{n} \sum_{k=1}^{c} \left(\left[\sum_{j=1}^{m} (X_{ij} - V_{kj})^2 \right] (\mu_{ik})^w \right) \tag{7}$$

The FCM process will stop if the objective function value minus the objective function value of the previous iteration is less than or equal to the epsilon value, or if the maximum number of iterations has been reached.

5) Silhouette Score

The Silhouette Score is a metric used to assess the quality of clustering by measuring how well data points within a cluster are grouped [35]. The Silhouette Score ranges from -1 to 1, where higher values indicate better clustering. This value considers two main factors: how close data points are to other points within the same cluster (cohesion) and how far data points are from points in other clusters (separation). For a data point i, the silhouette score s (i) is calculated using the following steps:

1. The average intra-cluster distance $(a(i))$ is calculated as follows:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i,j) \tag{8}$$

Where Ci is the cluster where point i belongs, and $d(i,j)$ is the distance between point i and point j. The average intra-cluster distance $a(i)$ is the average distance from point i to all other points within cluster Ci.

2. The average distance to the nearest cluster ($b(i)$) is calculated as follows :

$$b(i) = \min_{C_k \neq C_i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \quad (9)$$

Where Ck represents clusters other than cluster Ci. The average distance to the nearest cluster $b(i)$ is the smallest distance from point i to points in other clusters.

3. The Silhouette Score for point i ($s(i)$) is calculated as :

$$s(i) = \frac{b(i) - a\,(i)}{\max\,(a(i), b(i))} \quad (10)$$

The value of $s(i)$ ranges between -1 and 1:

a) Values approaching 1 indicate that point i is far from other clusters and close to its own cluster.
b) Values approaching 0 indicate that point i is on the boundary between two clusters.
c) Values approaching -1 indicate that point i is closer to another cluster than its own cluster.

4. The Average Silhouette Score is calculated as:

$$s = \frac{1}{N} \sum_{i=1}^{N} s(i) \quad (11)$$

Where N is the total number of data points. This average silhouette score provides a measure of the overall clustering quality for the entire dataset.

5. Interpreting the Silhouette Score:
The Silhouette Score measures how well data points are clustered by comparing intra-cluster proximity (within the same cluster) to inter-cluster distances (with other clusters).

a) Silhouette Score > 0.5:
Clustering is considered good. Data points are generally well-clustered into their respective clusters.
b) 0 < Silhouette ≤ 0.5:
Clustering is fair. Data points are close to the boundary between clusters.
c) Silhouette Score ≤ 0:
Clustering is poor. Data points may be in the wrong cluster or clusters are poorly defined.

By considering intra-cluster proximity and inter-cluster distances, the silhouette score provides a clear indication of the quality of the cluster structure within the dataset.

C. Result and Discussion

1) Data Collection Stage

In the data collection stage, the data used in this analysis is obtained from the sales records of PT. XYZ located at Pekanbaru Riau. This sales data covers

the period from 2020 to 2023, detailing the number of cars sold each year. The sales data obtained is as follows (Table I).

TABLE 1. Data Collection Stage

YEARS	SALES AMOUNT
2020	986
2021	1625
2022	1931
2023	646
TOTAL	**5188**

2) Cleaning Data

This sales data will be used for this final project. After the data is collected, the next step is data cleaning. The cleaning stage for DBSCAN and FCM clustering involves organizing the collected data in an Excel format and cleaning it from errors or missing data.

Before Data Cleaning, the Sales Data consists of 5188 entries with 27 columns, including the payment proof number (BP), payment proof date, vehicle order letter number (SPK), vehicle order letter date, vehicle model, vehicle engine capacity in cubic centimeters (CC), vehicle transmission type, vehicle fuel type, vehicle engine number, payment method used, vehicle price before discount, amount of discount given on the vehicle price. Customer's name or identity, additional variations or accessories selected by the customer, vehicle color ordered, information about whether the customer received free insurance or not, the name of the supervisor overseeing the transaction or sales process, the district where the customer lives, the sub-district where the customer lives, the village where the customer lives, a description of the road conditions in the customer's area, the customer's occupation or profession, the name of the insurance company used by the customer, the type of insurance owned by the customer, such as all risk or combination, buyer type, whether the buyer is a first buyer or an additional buyer, the duration or period of insurance taken by the customer.

TABLE 2. Cleaning Data

No	PURCHASE DATE	MODEL	...	REGENCY	...	ROAD CONDITIONS	...	DURATION OF INSURANCE
1	2020-01-09	VELOZ 1.5 M/T	...	PELALAWAN	...	The road is flooded when it rains	...	2 THN ALL RISK, 3 THN TLO
2	2020-01-10	HIACE 2.8 PREMIO	...	PEKANBARU	...	Solid but Smooth and Urban	...	2 THN ALL RISK, 3 THN TLO
3	2020-01-11	CALYA 1.2 E M/T	...	PELALAWAN	...	Flooded, badly damaged	...	2 THN ALL RISK, 3 THN TLO
4	2020-01-11	FORTUNER Grand 2.4 VRZ A/T	...	SIAK	...	Homely, Secluded, Connected,	...	2 TAHUN

		4x2 TRD Diesel				Traditional.		
5	2020-01-15	CALYA 1.2 E M/T	…	PEKANBARU	…	Solid but Smooth and Urban	…	5 TAHUN
…	…	…	…	…	…	…	…	…
3315	2023-04-30	AVANZA 1.5 G M/T	…	PEKANBARU	…	Solid but Smooth and Urban	…	9 THN ALL RISK, 4 THN TLO
3316	2023-04-30	AVANZA 1.5 G M/T	…	PELALAWAN	…	Solid, Smooth, Modern, Evolving	…	2 THN ALL RISK, 3 THN TLO
3317	2023-04-30	HILUX_DC 2.4 G 4X4 MT (E4)	…	PELALAWAN	…	Busy, Organized, Developing, Modern	…	10 THN ALL RISK, 2 THN TLO
3318	2023-04-30	RUSH 1.5 S A/T GR SPORT	…	PEKANBARU	…	Solid but Smooth and Urban	…	10 THN ALL RISK, 3 THN TLO
3319	2023-04-30	RUSH 1.5 S M/T GR SPORT	…	INDRAGIRI HULU	…	Road Conditions are Severely Damaged	…	2 THN ALL RISK, 3 THN TLO

After the data cleaning process, the resulting data is cleaner and ready for further analysis. Irrelevant columns have been removed, and all remaining rows do not contain missing values. This improves the quality of the data and facilitates further analysis. The final dataset has 21 columns containing essential information: payment proof date, vehicle model, vehicle engine capacity in cubic centimeters (CC), vehicle transmission type, vehicle fuel type, payment method used, vehicle price before discount, amount of discount given on the vehicle price. Customer's name or identity, additional variations or accessories selected by the customer, vehicle color ordered, the name of the supervisor overseeing the transaction or sales process, the district where the customer lives, the sub-district where the customer lives, the village where the customer lives, a description of the road conditions in the customer's area, the customer's occupation or profession, the name of the insurance company used by the customer, the type of insurance owned by the customer, such as all risk or combination, buyer type, whether the buyer is a first buyer or an additional buyer, the duration or period of insurance taken by the customer. The dataset consists of 3319 valid and complete rows.

3) Pre-Processing Stage

a. Data Transformation Stage

Before using clustering algorithms, data needs to be converted into a suitable format. This transformation is important to ensure accurate and reliable clustering results. After transformation, the data is ready to be used in the selected clustering algorithm. DBSCAN is suitable for numerical data with high density, while FCM is suitable for fuzzy clustering.

b. Data Normalization Stage

Data normalization is a crucial preprocessing step in many clustering algorithms, including DBSCAN and FCM. Normalization ensures that all features have the same scale, preventing features with larger magnitudes from dominating the clustering process. Normalization typically involves scaling each feature to a range between 0 and 1 or transforming it to have a mean of 0 and a standard deviation of 1. This ensures that all features contribute equally to the clustering process, regardless of their original scales. After normalization, the data is ready to be fed into the clustering algorithm, allowing for more accurate and reliable clustering results.

4) Implementation of Density-Based Spatial Clustering of Applications with Noise Algorithm

Several clustering experiments were conducted using the DBSCAN algorithm to find the optimal number of clusters. Each experiment used different values for Eps and MinPts. The Eps value used in this study was 1.5, with MinPts set to 11. The clustering results can be seen in the table below :

TABLE 3. Cluster Data DBSCAN

NO	PURCHASE DATE	MODEL	...	REGENCY	...	ROAD CONDITIONS	...	CLUSTER	EPS	MIN_SAMPLER
0	2020-01-09	194	...	6	...	12	...	0	1,5	11
1	2020-01-10	70	...	10	...	27	...	0	1,5	11
2	2020-01-11	33	...	6	...	5	...	0	1,5	11
3	2020-01-11	68	...	9	...	3	...	0	1,5	11
4	2020-01-15	33	...	10	...	27	...	0	1,5	11
...	...	...	...	...	...	...	...	...	...	11
3314	2023-04-29	15	...	10	...	27	...	0	1,5	11
3315	2023-04-29	15	...	6	...	28	...	0	1,5	11
3316	2023-04-29	82	...	6	...	50	...	0	1,5	11
3317	2023-04-29	166	...	10	...	27	...	0	1,5	11
3318	2023-04-29	169	...	36	...	19		0	1,5	11

5) Evaluation Model Density-Based Spatial Clustering of Applications with Noise Algorithm

After performing the clustering step to find the most optimal number of clusters, cluster validity was evaluated using the Silhouette Score. The cluster validity results can be seen in Figure 2 below :

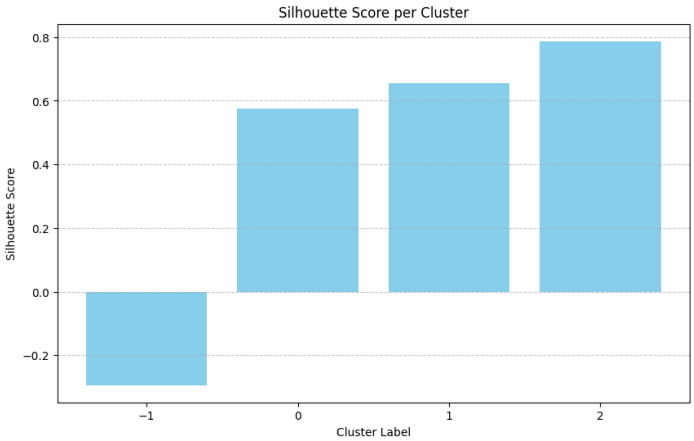

Figure 2. Silhouette Score for DBSCAN

From the Silhouette Score values, cluster 2 has the highest Silhouette Score of 0.7874. This indicates that this cluster has good quality in terms of inter-cluster separation and intra-cluster cohesion. However, Cluster -1 is not good as it contains noise with a negative Silhouette Score. This is normal and expected in DBSCAN clustering. On the other hand, Clusters 0, 1, and 2 are considered good because they have fairly high Silhouette Scores (above 0.5), indicating that the points within these clusters are well-grouped. However, the results from Cluster 0 are neutral and include general data from the overall dataset.

TABLE 4. Cluster Data Fuzzy C - Means

NO	TGL BP	MODEL	...	REGENCY	DISTRICT	SUB-DISTRICT	ROAD CONDITIONS	...	CLUSTER
134	2020-03-07	INNOVA G AT BENSIN	...	PEKANBARU	SAIL	SUKAMAJU	Dense but Smooth and Urban	...	2
135	2020-03-07	INNOVA G AT BENSIN	...	SIAK	TUALANG	MEREDEN BARAT	Rural but with good road conditions	...	2
229	2020-04-30	AVANZA 1.3 G M/T	...	PEKANBARU	SAIL	SUKAMJU	Dense but Smooth and Urban	...	2
415	2020-07-30	CALYA 1.2 G M/T	...	PELALAWAN	LANGAM	LANGGAM	Dense but Smooth and Urban	...	2
499	2020-11-24	INNOVA V AT DIESEL	...	KAMPAR	TAPUNG HILIR	TANDAN SARI	The roads are damaged and need repairs	...	2
...	...	...	...	...	...	...	...	...	2
...	...	...	...	...	...	...	...	...	2
2604	2022-10-29	FORTUNER 2.8 VRZ 4x2 AT GR Diesel (E4)	...	INDRAGIRI HILIR	TEMBILAHAN HULU	TEMBILAHAN HULU	The road conditions are good and frequented by several large vehicles	...	2
2606	2022-10-29	RAIZE 1.0T GR Sport S CVT TSS Two Tone	...	SIAK	BUNGA RAYA	TEMUSAI	Dense but Smooth and Urban	...	2
2700	2022-11-26	RUSH 1.5 S A/T GR SPORT	...	PEKANBARU	SAIL	SUKAMAJU	Dense but Smooth and Urban	...	2
297	2023-	INNOVA	...	PEKANBARU	SAIL	SUKAMAJU	Dense but	...	2

5	01-31	Zenix 2.0 Q HV Modelista CVT Prem TSS					Smooth and Urban		
3239	2023-04-18	INNOVA Zenix 2.0 Q HV CVT TSS Modelista Prem	...	SIAK	KOTO GASIB	DESA KUALA GASIB	Well-maintained roads with rare traffic jams but frequent damages	...	2

Cluster 1 generally includes travel or commuting vehicles. Models such as Innova G AT Petrol, Calya 1.2 G A/T, Avanza 1.3 G M/T, Fortuner 2.8 Vrz 4X2 AT Gr Diesel, Raize 1.0t Gr Sport S CVT TSS Two Tone, Rush 1.5 S A/T Gr Sport, and INNOVA Zenix 2.0 Q HV Modelista CVT Prem TSS are predominantly MPVs (Multi-Purpose Vehicles) and SUVs (Sport Utility Vehicles). These vehicles are known for their larger passenger capacity, comfort on long-distance trips, and versatility on various road conditions. Models like the Innova and Avanza are highly popular as family cars, while the Fortuner and Rush are recognized SUVs suitable for long journeys. Therefore, vehicles in Cluster 1 are suitable for family trips or travel, offering the comfort and efficiency needed for long-distance travel across varied road conditions from good to damaged. The cars in Cluster 1 vary in price reflecting different models and specifications. The majority of customers are from PT. Swadaya Abdi Manunggal, spread across various regions such as Pekanbaru City with Sail District, Siak Regency with Tualang, Koto Gasib, Bunga Raya, Pelalawan Regency with Langgam, Kampar with Tapung Hilir, and Indragiri Hilir with Tembilahan Hulu. Customers have diverse occupations including entrepreneurs, civil servants, and private sector employees. Insurance providers used include ABDA and Ramayana offering TLO, all-risk, and combined insurance types. Buyer types include additional and first buyers with insurance durations varying up to 11 years.

NO	**TGL BP**	**MODEL**	**...**	**REGENCY**	**DISTRICT**	**SUB-DISTRICT**	**ROAD CONDITIONS**	**...**	**CLUSTER**
283	2020-06-30	ALPHARD 2.5 G AT Prem	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban	…	2
654	2021-01-30	ALPHARD 2.5 G AT	…	SIAK	SIAK	KAMPUNG REMPAK	Scenic, Remote, Connected, Traditional	…	2
929	2021-05-31	ALPHARD 2.5 G AT	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban	…	2
1177	2021-08-31	ALPHARD VELLFIRE 2.5 G AT	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban Perkotaan	…	2
1271	2021-09-30	ALPHARD 2.5 G AT	…	SIAK	BUNGA RAYA	TEMUASI	The road conditions are good and are traversed by several large vehicles	…	2

1272	2021-09-30	ALPHARD 2.5 G AT	…	SIAK	BUNGA RAYA	TEMUSAI	Keadaan Jalan Baik dan dilalui beberapa kendaraan besar	…	2
1273	2021-09-30	ALPHARD 2.5 G AT Prem	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban	…	2
2036	2022-05-31	ALPHARD 2.5 G AT	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban	…	2
2186	2022-06-30	ALPHARD 2.5 G AT Prem	…	PEKANBARU	TAMPAN	TUAH MADANI	Busy, Connected, Modern, Developing	…	2
2611	2022-10-31	ALPHARD 2.5 G AT Prem	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban	…	2
2730	2022-11-30	ALPHARD 2.5 G AT	…	PEKANBARU	SAIL	SUKAMAJU	Dense yet Smooth and Urban	…	2

Based on PT. XYZ sales data, there is an interesting purchasing pattern observed in cluster 2. This cluster is dominated by entrepreneurs who purchase high-priced Alphard cars, both through cash and credit transactions. Customers in this cluster typically opt for All Risk or Combined insurance to protect their vehicles. They reside in areas with diverse road conditions, ranging from busy and smooth to scenic and remote. The chosen insurance durations vary, ranging from 1 year to 12 years.

This buying pattern indicates that the Alphard is a popular choice among entrepreneurs in Pekanbaru seeking high mobility and driving comfort. Their willingness to pay high prices and choose comprehensive insurance reflects their appreciation for quality and safety.

6) Implementation of Fuzzy C-Means Algorithm

This study applies the FCM Clustering algorithm to analyze patterns in the data. Using 11 clusters, the minimum error is set to 0.005, and the maximum iteration limit is set to 1000. The resulting clusters are as follows :

TABLE 5. Cluster Data Fuzzy C - Means

NO	PURCHASE DATE	MODEL	…	REGENCY	...	ROAD CONDITIONS	…	CLUSTER
1	2020-01-09	194	…	6	…	12	**…**	9
2	2020-01-10	70	…	10	…	27	**…**	7
3	2020-01-11	33	…	6	…	5	**…**	2
4	2020-01-11	68	…	9	…	3	**…**	7
5	2020-01-15	33	…	10	…	27	**…**	2
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**

3315	2023-04-30	15	…	10	…	27	**…**	9
3316	2023-04-30	15	…	6	…	28	**…**	4
3317	2023-04-30	83	…	6	…	50	**…**	7
3318	2023-04-30	166	…	10	…	27	**…**	0
3319	2023-04-30	169	…	2	**…**	19	**…**	4

7) Evaluation Model Fuzzy C – Means Algorithm

This figure illustrates cluster validity using Silhouette Score for each number of clusters and identifies the number of clusters that yield the highest Silhouette Score. A higher Silhouette Score indicates better cluster division.

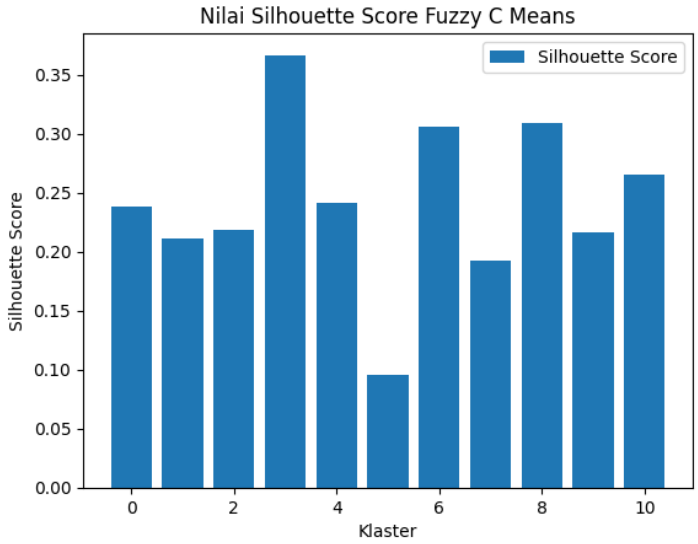

Figure 3. Silhouette Score for Fuzzy C-Means

From the above Silhouette Score values, cluster 3 has the highest Silhouette Score, which is 0.366609067535743. Therefore, cluster 3 can be considered the best cluster in terms of data separation and compactness. Hence, the optimal number of clusters in this case is 3. Looking at cluster 3, the purchase of vehicles with model 165, having an engine capacity of 1496 cc, and using manual transmission is prominent. Vehicles with model 165 are predominantly driven in districts 29 and 9, in villages 56 and 45, with good road conditions. The majority of purchases are made in cash by individual clients with formal occupations, with a price range of Rp258,600,000 - Rp289,950,000 and discounts ranging from Rp2,000,000 - Rp8,000,000. They mostly opt for comprehensive insurance with a duration of 36 - 70 months.

8) Comparison of Silhouette Score Results between Density-Based Spatial Clustering of Applications with Noise and Fuzzy C-Means Algorithms

Figure 4 illustrates the Comparison Graph of the highest Silhouette Score index values for each algorithm. The DBSCAN algorithm with a Silhouette Score index value in cluster 2 has the highest score, which is 0.7874, while the FCM algorithm has a Silhouette Score value of 0.3666 in cluster 3.

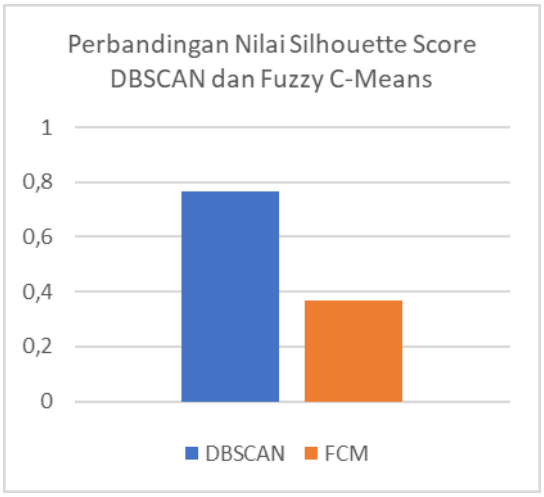

Figure 4. Silhouette Score Comparison

Figure 4 shows that the DBSCAN algorithm has the best cluster validity compared to the FCM algorithm. Therefore, the most optimal clustering algorithm is the DBSCAN with cluster 2 having the highest value.

9) Cluster Analysis

From the comparison of clustering model evaluations, it can be concluded that cluster 2 from the DBSCAN algorithm is the best. The data from cluster 2 reveals various information about customers, the vehicles they purchase, and their buying patterns. This buying pattern indicates that the Alphard car is a popular choice among entrepreneurs in Pekanbaru seeking high mobility and driving comfort. Their willingness to pay high prices and choose comprehensive insurance reflects their appreciation for quality and safety. This cluster information can be used by PT. XYZ to develop more targeted marketing and sales strategies. By understanding the characteristics of customers in cluster 2, PT. XYZ can better target its products and services, thereby improving the company's profitability.

D. Conclusion

This study compares the DBSCAN and FCM algorithms for clustering car sales data in Pekanbaru. For DBSCAN, achieved the highest Silhouette Score of 0.7874 for cluster 2, indicating more valid and well-separated clusters and more effective in handling data with irregular distributions and noise. Also for Fuzzy C-Means, achieved the highest Silhouette Score of 0.3666 in cluster 3 and identified patterns in purchasing specific car models, albeit with lower cluster validity compared to DBSCAN. DBSCAN outperforms FCM in producing high-quality clusters compared to FCM, making it more suitable for data with irregular distributions and noise. Choosing the right algorithm is crucial for effective data analysis in the context of car sales.

E. References

[1]. S. Fatimah, *Pengantar transportasi*. Myria Publisher, 2019.
[2]. S. Gusty *et al.*, *Dasar-Dasar Transportasi*. Tohar Media, 2023.

[3]. H. A. Karim *et al.*, *Manajemen transportasi*. Cendikia Mulia Mandiri, 2023
[4]. K. Wada, *The Evolution of the Toyota Production System*. Springer, 2020.
[5]. K. Amasaka, *Examining a New Automobile Global Manufacturing System*. IGI Global, 2022.
[6]. S. Grushetsky, I. Brylev, S. Evtukov, and A. Pushkarev, "Road accident prevention model involving two-wheeled vehicles," *Transp. Res. procedia*, vol. 50, pp. 201–210, 2020.
[7]. C. M. Zellatifanny, "Tren Diseminasi Konten Audio on Demand melalui Podcast: Sebuah Peluang dan Tantangan di Indonesia Trends in Disseminating Audio on Demand Content through Podcast: An Opportunity and Challenge in Indonesia," *J. Pekommas*, vol. 5, no. 2, pp. 117–132, 2020.
[8]. L. A. Adha, "Digitalisasi industri dan pengaruhnya terhadap ketenagakerjaan dan hubungan kerja di Indonesia," *J. Kompil. Huk.*, vol. 5, no. 2, pp. 267–298, 2020.
[9]. P. Intarakumnerd, "Technological upgrading and challenges in the Thai automotive industry," *J. Southeast Asian Econ.*, vol. 38, no. 2, pp. 207–222, 2021.
[10]. M. D. Shidqi, "PENGARUH CUSTOMER EXPERIENCE, BRAND TRUST, TERHADAP REPURCHASE INTENTION MELALUI CUSTOMER SATISFACTION SEBAGAI VARIABEL INTERVENING (Studi pada konsumen Suzuki Pick UP di Kota Cilacap) 59adbis2022." FAKULTAS ILMU SOSIAL DAN ILMU POLITIK UNIVERSITAS DIPONEGORO, 2022.
[11]. R. K. Priambono, "Pengaruh Konflik Kerja Terhadap Kinerja Karyawan Pada PT. Agung Automall Cabang Ujung Batu Kabupaten Rokan Hulu Riau." Universitas Islam Riau, 2020.
[12]. E. Rahayu, "Analisis Sistem Informasi Akuntansi Penjualan Dengan Aplikasi Sap (System Application And Product In Data Processing) Terhadap Efektifitas Proses Penjualan Pada Auto 2000 Bogor".
[13]. H. HASRINA, "Pengaruh Bauran Pemasaran Terhadap Keputusan Pembelian Mobil Toyota Yaris Pada PT. Hadji Kalla Cabang Urip Sumoharjo di Kota Makassar." FE, 2017.
[14]. A. Navi, "Analisis Penetapan Harga Jual Beli Mobil Bekas Untuk Meningkatkan Penjualan Showroom Japanese Motor Bandar Lampung 2019-2020," 2022.
[15]. S. P. Tamba and F. T. Kesuma, "Penerapan Data Mining Untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering: data mining; k-means-clustering," *J. Sist. Inf. dan Ilmu Komput. Prima (JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 67–72, 2019.
[16]. D. S. O. Panggabean, E. Buulolo, and N. Silalahi, "Penerapan Data Mining Untuk Memprediksi Pemesanan Bibit Pohon Dengan Regresi Linear Berganda," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, pp. 56–62, 2020.
[17]. O. R. Sirait, "Penerapan Data Mining Dalam Mengelompokkan Keberhasilan Kurir Sicepat Ekspres Menggunakan Algoritmak-Means," *SkripsiKu-2022*, vol. 1, no. 1, 2022.
[18]. R. F. Putra *et al.*, *DATA MINING: Algoritma dan Penerapannya*. PT. Sonpedia Publishing Indonesia, 2023.
[19]. H. Syukron, M. F. Fayyad, F. J. Fauzan, Y. Ikhsani, and U. R. Gurning,

"Perbandingan K-Means K-Medoids dan Fuzzy C-Means untuk Pengelompokan Data Pelanggan dengan Model LRFM: Comparison K-Means K-Medoids and Fuzzy C-Means for Clustering Customer Data with LRFM Model," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 76–83, 2022.

[20]. R. F. Putra *et al.*, *Algoritma Pembelajaran Mesin: Dasar, Teknik, dan Aplikasi*. PT. Sonpedia Publishing Indonesia, 2024.

[21]. L. G. A. Putri, "Upaya Peningkatan Kinerja Pemasaran Berbasis Strategi Digital Marketing Studi Kasus Pada Pt Astra International Bmw Sales Operation Semarang." Universitas Islam Sultan Agung, 2023.

[22]. F. Yoseph, N. H. A. H. Malim, and M. AlMalaily, "New behavioral segmentation methods to understand consumers in retail industry," *AIRCC's Int. J. Comput. Sci. Inf. Technol.*, pp. 43–61, 2019.

[23]. L. Priyadi and Y. Takahashi, "The Dynamics of the Toyota-Astra Hybrid Structure Partnership," *Institutions Econ.*, pp. 85–122, 2019.

[24]. N. Trianasari and T. A. Permadi, "Analysis Of Product Recommendation Models at Each Fixed Broadband Sales Location Using K-Means, DBSCAN, Hierarchical Clustering, SVM, RF, and ANN," *J. Appl. Data Sci.*, vol. 5, no. 2, pp. 636–652, 2024.

[25]. S. SAMSUL, "Perbandingan Kinerja Fuzzy C-Means Dengan Dbscan Untuk Menentukan Segmentasi Pelanggan Berdasarkan Rfm Studi Kasus Printo Digital Printing." Universitas Mercu Buana, 2022.

[26]. P. D. W. Ayu, "Perbandingan Kinerja Fuzzy C-Means dan DBSCAN Dalam Segmentasi Citra USG Kepala Janin," *J. Sist. dan Inform.*, vol. 9, no. 2, pp. 79–85, 2015.

[27]. A. Subasi, *Practical machine learning for data analysis using python*. Academic Press, 2020.

[28]. A. K. Wicaksana and D. E. Cahyani, "Modification of a density-based spatial clustering algorithm for applications with noise for data reduction in intrusion detection systems," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 21, no. 2, pp. 189–203, 2021.

[29]. K. V. Rajkumar, A. Yesubabu, and K. Subrahmanyam, "Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 4, p. 2760, 2019.

[30]. L. R. Nair, "A NOVEL STUDY OF SILHOUETTE METHOD TO SOLVE THE ISSUES OF OUTLIER AND IMPROVE THE QUALITY OF CLUSTER," *J. Data Acquis. Process.*, vol. 38, no. 2, p. 3099, 2023.

[31]. KOMPAS, "Jumlah Kendaraan di Indonesia 147 Juta Unit, 87 Persen Motor," Jakarta, Feb. 10, 2023. [Online]. Available: https://otomotif.kompas.com/read/2023/02/10/070200315/jumlah-kendaraan-di-indonesia-147-juta-unit-87-persen-motor

[32]. Kepolisian Republik Indonesia, "Perkembangan Jumlah Kendaraan Bermotor Menurut Jenis (Unit), 2021-2022," *BADAN PUSAT STATISTIK*, p. 1, 2024. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NTcjMg==/perkembangan-jumlah-kendaraan-bermotor-menurut-jenis--unit-.html

[33]. Gabungan Industri Kendaraan Bermotor Indonesia, "Penjualan Mobil 2023

Lampaui Sejuta Unit, Merek-merek Astra Terbanyak," *https://www.gaikindo.or.id/*, 2024. https://www.gaikindo.or.id/penjualan-mobil-2023-lampaui-sejuta-unit-merek-merek-astra-terbanyak/

[34]. B. Satria, "Terjual 3.290 Unit, Penjualan Mobil Bulan Agustus di Riau Naik 309 Unit," *Hallo Riau*, Pekanbaru, Sep. 08, 2022. [Online]. Available: https://www.halloriau.com/read-otomotif-1427383-2022-09-08-terjual-3290-unit-penjualan-mobil-bulan-agustus-di-riau-naik-309-unit.html

[35]. Gabungan Industri Kendaraan Bermotor Indonesia, "Industri Otomotif Punya Prospek Positif mulai Kuartal Keempat 2024," *Gaikindo*, 2024. https://www.gaikindo.or.id/20299/