Journal of Information Systems and Informatics
Vol. 7, No. 1, March 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i1.1002 Published By DRPM-UBD

Traffic Violation Clustering Using K-Medoids and Word Cloud Visualization

Muhammad Sabri S[1], Ema Utami[2]

[1,2]Magister of Informatics Engineering, Amikom University, Yogyakarta, Indonesia
Email: [1]muhammadsabry41@students.amikom.ac.id, [2]ema.u@amikom.ac.id

Abstract

Traffic is the space for people to move around, including both drivers and pedestrians. According to data from the Central Statistics Agency in 2020, the number of motor vehicles in Makassar City was recorded by type: 248,682 passenger cars, 17,501 buses, 85,968 trucks, and 1,338,306 motorcycles, with a tendency for an increase in the following year. The high number of vehicle users can certainly affect the rising traffic violation rates on the road. This study aims to classify traffic violation types in Makassar City by utilizing the K-Medoids algorithm and to visualize the clustering results using Word Cloud, which is expected to provide information related to patterns of traffic violation clusters. This study uses a case study from the Traffic Police Department of Makassar City in 2021, with a total of 5,893 traffic violation cases. The data used is ticket data consisting of article and vehicle type features. The clustering results show that motorcycles and minibuses are the most frequently involved in traffic violations. Motorcycles (R2) are not only dominated by violations related to the use of standard SNI helmets but also significantly involved in violations related to incomplete requirements and the possession of SIM/STNK (Driver's License/Vehicle Registration) and failing to meet roadworthiness standards such as mirrors, headlights, horns, etc. Passenger vehicles, especially minibuses and cars, also dominate traffic violations. The violations involve not only the use of seat belts for R4 vehicles but also violations such as not having complete STNK, not being able to show SIM, failing to display the Vehicle Registration Mark (TKB), and others. The results of this study demonstrate that the clustering obtained is very strong, as evidenced by the high Silhouette Score of 0.867 at k = 9.

Keywords: Violations, Clustering, Kmedoids, PCA, Elbow Method, Silhouette Score, Word Cloud.

1. INTRODUCTION

Traffic is a means for people to carry out driving activities, both as drivers and pedestrians. The use of motor vehicles in traffic spaces continues to increase every year, including public transportation and private vehicles. Based on data from the Central Statistics Agency in 2020, Makassar City recorded the following number of motor vehicles: 248,682 units of passenger cars, 17,501 units of buses, 85,968 units of trucks, and 1,338,306 units of motorcycles. In the following year, there was a significant increase to 257,015 units of passenger cars, 17,582 units of buses,

88,359 units of trucks, and 1,377,837 units of motorcycles. This congestion triggers an increase in traffic violations on the highway [1].

Various violations were committed by motorists, such as violating traffic lights, speeding, not wearing seat belts, and violating road signs and markings. Factors that contribute to traffic violations include weather conditions such as rain and heat, traffic density, and the dominance of private vehicles [2]. In addition, the limitations of the police force, the lack of public awareness of traffic rules, and the lack of supporting infrastructure are also the main causes [3]. As a result, these violations can threaten the safety of other motorists.

Observation results show that in 2020, the number of traffic violations in Makassar City reached around 15,000 cases and has the potential to continue to increase. The Makassar City Police Station handles the violation by recording manual tickets, which are then input into the electronic ticket system. However, this system only stores descriptive information from the offender without further analysis. Therefore, a data mining analysis approach is needed to explore new patterns and knowledge of various types of violations.

Several previous studies have applied the K-Means algorithm to cluster areas prone to traffic violations. For example, the research of [4] identified three clusters of violation-prone areas. Research by [5] shows that traffic violations are generally committed by teenagers and young adults with the profession of students, students, and private employees. On the other hand, Atmaja et al.'s research used the K-Medoids algorithm to analyze crime patterns, resulting in three clusters of crime rates.

Other studies, such as those conducted by Marlina et al., compared K-Medoids and K-Means algorithms in the case of distribution of children with special needs. The results show that the K-Medoids algorithm is superior based on the Silhouette value for clustering with three groups. The research of [6] also compared the two algorithms in analyzing the spatial pattern of earthquakes in Indonesia. The results show that the K-Medoids algorithm provides better performance than K-Means with the highest Silhouette value of 0.4674067 for six clusters.

Based on this review, this study aims to group ticket data from the Makassar City Police Traffic Unit using the K-Medoids algorithm. This algorithm is expected to be able to identify patterns and group various types of traffic violations that occur in Makassar City. K-Medoids was chosen because it is more resistant to outliers compared to K-Means, as it uses the medoid as the cluster center instead of the average. This method is also more suitable for categorical data, such as vehicle types or violation articles. Based on previous research, K-Medoids provides better

clustering results with higher Silhouette scores, making it the ideal choice for analyzing traffic violations in Makassar City.

2. METHODS

2. 1 K-Medoids Clustering Technique

In applying K-Medoids, it is necessary to have an optimal k value (number of clusters) to be established in the early stages. The selection of the optimal k value was carried out by looking at the results of the SSE graph plot in the form of an elbow with the Elbow Method. In addition to using elbows, the Silhouette Score value is also a consideration in determining the value of k. Kaufman and Rousseuw put forward the interpretation of the Silhouette Score value which is in the value range of 0.51 < SC < 0.70 clusters which are quite good have been formed, and in the value range of 0.70 SC < < 1 very good and strong cluster has been formed. illustrates the flowchart of the K-Medoids algorithm.

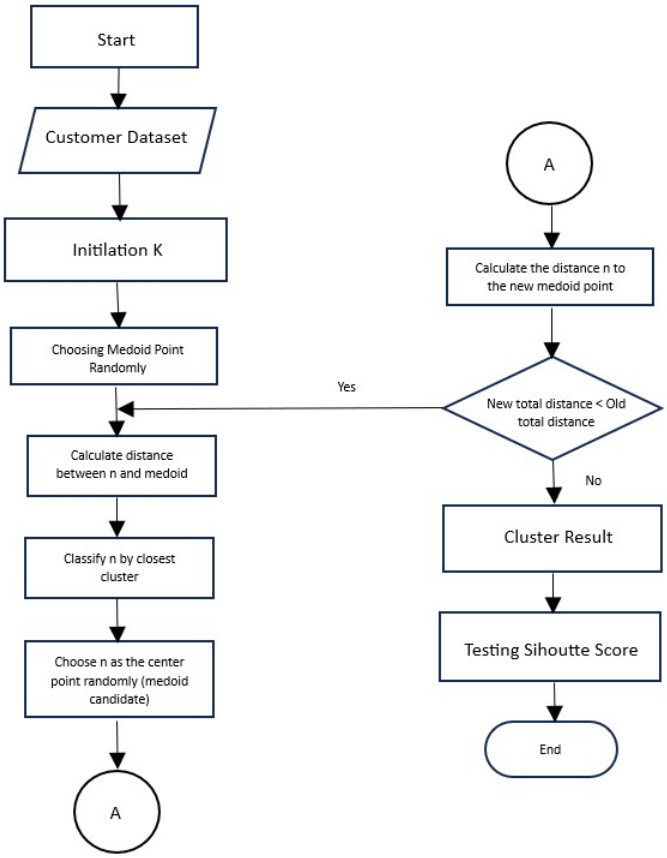

Figure 1. PCA groove

The process begins with preparing the dataset and the algorithm used for clustering. At this stage, all necessary preparations are made, such as selecting the dataset and setting the initial parameters. The dataset used contains traffic violation data, which may include variables such as the type of violation, the vehicles involved, the time of occurrence, location, and others. Next, the value of K (the number of desired clusters) is determined. K represents the number of groups that will be formed in the clustering process, based on the needs of the analysis.

In the next step, a random medoid is chosen from the dataset. A medoid is an object from the data that will serve as the center of each cluster. The random selection of the medoid is the starting point for the clustering process [7]. Then, the distance between each data point and the chosen medoid is calculated, typically using methods such as Euclidean distance, which helps determine how close each data point is to the medoid. Based on this distance, the data is grouped into clusters, with each data point being assigned to the nearest medoid.

Following the initial grouping, a new medoid is randomly chosen from the data within the cluster. This new medoid is then used to evaluate the cluster more accurately, aiming to find a more representative medoid. The distance between each data point and the newly selected medoid is recalculated to assess whether the new medoid is a better fit for the cluster. If the total distance to the new medoid is smaller than the previous total distance, the new medoid is accepted, and the algorithm continues by recalculating the distances, grouping the data again, and randomly selecting a new medoid.

This process repeats, with data points being regrouped and new medoids being selected, until no significant changes occur, meaning the medoid is considered optimal. The clustering process is then terminated. Once clustering is completed, the results are evaluated using the Silhouette Score, which measures the quality of the clustering. A higher Silhouette Score indicates that the clusters are well-separated and that the clustering has been successful. If the score is favorable, it indicates that the clustering was done properly. Finally, the clustering process is completed, and the results can be used for further analysis or decision-making [5]. K-Medoids is a clustering method that groups data using medoids as the center of the cluster. Medoid is the most representative raw data in a cluster, in contrast to K-Means which uses centroid as the center of the cluster.

Determining the value of k (number of clusters) is an important step in the application of K-Medoids. To determine the optimal k, the Elbow Method approach is used which plots the SSE (Sum of Squared Errors) value against various k values. The "elbow" point on the chart shows the optimal k value, where the SSE decline begins to slow down significantly.[2]

In addition to the Elbow Method, the Silhouette Score is also used as an additional indicator. The Silhouette Score measures the cohesiveness within the cluster and the separation between clusters, with a range of values:

a) $0.51 < SC < 0.70$: Clusters are quite good.
b) $0.70 < SC \leq 1$: Very good and strong cluster.

With these two approaches, the optimal k value can be objectively selected for use in clustering.

PCA (Principal Component Analysis) plays an important role in dimensionality reduction before performing clustering. PCA is used to reduce the number of variables or features in a dataset by transforming the data into a lower-dimensional space, while still preserving as much of the original information as possible. In the context of clustering, using PCA helps to address the "curse of dimensionality," where having more features makes it more difficult to find significant patterns. By reducing the data's dimensions, PCA makes clustering algorithms more efficient and faster, and can improve clustering accuracy by reducing noise and irrelevant features.

After applying PCA for dimensionality reduction, the clustering process becomes more focused on the important and relevant variables in determining the patterns present in the data. For example, if traffic violation data contains many features, such as vehicle type, time, and location of incidents, PCA can identify the main components that most influence the violation patterns. This way, clustering can produce more accurate and meaningful results.

Regarding the selection of the optimal number of clusters (K), this step is one of the key challenges in clustering. The choice of the optimal number of clusters significantly affects the quality of the clustering results. Too few clusters can group different data into one cluster, while too many clusters can cause irrelevant data separation. One method for choosing the optimal K is by using the Elbow method or Silhouette Score. The Elbow method identifies the point at which the reduction in variance between clusters begins to slow down, while Silhouette Score measures how well each data point fits its cluster [8]. A high Silhouette Score indicates that the data in the cluster are well-separated and more relevant to the chosen cluster, which suggests that the selected number of clusters is optimal. In this way, PCA and the proper selection of the number of clusters work together to produce more efficient and meaningful clustering results in the context of traffic violation data analysis [9]-[13].

3. RESULTS AND DISCUSSION

3.2. Experimental Results

same penalty article but refer (junction to) to a different article of demand and vice versa the data of the same article of claim but the article of punishment is different, the data set of violation cases to be clustered, there are several data that have the so that the data is in an overlapped condition. In addition, there are cases of data having violation data that is subject to more than 1 article so that in this condition it is not known for sure whether the data will be included in the dominant cluster of the first violation article or the second violation article. So, with these data conditions, a grouping process is needed by implementing data mining, namely the clustering method. This study carried out the process of grouping the types of traffic violations using the K-Medoids Clustering method. From the experiments carried out previously, data reduction will be carried out by maintaining 80% data variance. PCA reduces or reduces the dimensions of 52 features and has resulted in the 7 components of PCA presented in Figure 2.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
0	-0.464602	-0.244079	0.022307	0.371302	1.013775	-0.798495	-0.019860
1	0.901546	0.016990	-0.006602	-0.001536	0.002250	0.007710	0.003151
2	-0.429307	-0.200064	0.074312	-0.035935	0.429132	0.688389	1.105595
3	-0.429307	-0.200064	0.074312	-0.035935	0.429132	0.688389	1.105595
4	0.895206	0.066016	0.005777	-0.006732	-0.005270	0.010115	-0.000920

Figure 2. Seven components of PCA

Data contained in the 7 components of the PCA will be used in the clustering process. This 80% data variance is divided into each component of the PCA from the reduction. Tabel 1 shows the cumulative value of the variance of the data loaded on each component.

Tabel 1. Cumulative variance of 7 PCA components data

PCA Components	Cumulative data variance
PC1	0.29219445
PC2	0.13429436
PC3	0.0974807
PC4	0.08850176
PC5	0.07248716
PC6	0.06119376
PC7	0.05519946

3.2. Calculation of How the K-Medoids Algorithm Works

K-Medoids is one of the clustering algorithms that is almost similar to K-Means. The difference between these two algorithms lies in the process of updating the centroid/medoid value in each iteration. Here are the manual calculation steps of the K-Medoids algorithm. Tabel 2 presents a sample of 10 data that will be used in the clustering process. In this case, the number of k = 2 will be randomly selected. The medoid points to be selected are A(2,6) and D(4,7).

Tabel 2. Sample data

Object	X1	X2
A	2	6
B	3	4
C	3	8
D	4	7
E	6	2
F	6	4
G	7	3
H	7	4
I	8	5
J	7	6

Then the distance between medoid points and non-medoid points will be calculated using the Euclidean Distance method. Tabel 3 presents the process of calculating the distance between data points.

Tabel 3. Calculation of the first iteration distance

Object	X1	X2	Distance to medoid 1	Distance to medoid 2
A	2	6	0	$d_{A,M2} = \sqrt{((2-4)^2 + (6-7)^2} = 2.23$
B	3	4	$d_{B,M1} = \sqrt{((3-2)^2 + (4-6)^2} = 2.23$	$d_{B,M2} = \sqrt{((3-4)^2 + (4-7)^2} = 3.16$
C	3	8	$d_{C,M1} = \sqrt{((3-2)^2 + (8-6)^2} = 2.23$	$d_{C,M2} = \sqrt{((3-4)^2 + (8-7)^2} = 1.414$
D	4	7	$d_{D,M1} = \sqrt{((4-2)^2 + (7-6)^2} = 2.23$	0
E	6	2	$d_{E,M1} = \sqrt{((6-2)^2 + (2-6)^2} = 5.65$	$d_{E,M2} = \sqrt{((6-4)2 + (2-7)^2} = 5.38$
F	6	4	$d_{F,M1} = \sqrt{((6-2)^2 + (4-6)^2} = 4.47$	$d_{F,M2} = \sqrt{((6-4)^2 + (4-7)^2} = 3.6$
G	7	3	$d_{G,M1} = \sqrt{((7-2)^2 + (3-6)^2}$	$d_{G,M2} = \sqrt{((7-4)^2 + (3-7)^2}$

Object	X1	X2	Distance to medoid 1	Distance to medoid 2
			= 5.83	= 5
H	7	4	$dH,M1 = \sqrt{((7-2)2 + (4-6)2}$ = 5.38	$dH,M2 = \sqrt{((7-4)2 + (4-7)2}$ = 4.242
I	8	5	$dI,M1 = \sqrt{((8-2)^2 + (5-6)2}$ = 6.08	$dI,M2 = \sqrt{((8-4)^2 + (5-7)2}$ = 4.47
J	7	6	$_{dJ,M1} = \sqrt{(7-2)2 + (6-6)^2} = 5$	$dJ,M2 = \sqrt{((7-4)^2 + (6-7)2}$ = 1.414

Closest Distance =
Cluster members 1 = A and B
Cluster members 2 = C, D, E, F, G, H, I, and J
Total closest distance to medoid point per cluster member =

0 + 2.23 + 1.414 + 0 + 5.38 + 3.6 + 5 + 4.242 + 4.47 + 1.414 = **27.75**

Furthermore, the selection of medoid points between data points with k=2 will be carried out randomly, namely points B(3,4) and H(7,4). Then it will be calculated again to carry out the distance calculation process. Tabel 4 presents the process of calculating the distance between data points in iteration 2.

Tabel 4. Calculation of the second iteration distance

Object	X1	X2	Distance to medoid 1	Distance to medoid 2
A	2	6	$dA,M1 = \sqrt{((2-3)^2 + (6-4)2}$ = 2.23	$dA,M2 = \sqrt{((2-7)2 + (6-4)2}$ = 5.38
B	3	4	0	$dB,M2 = \sqrt{((3-7)2 + (4-4)2}$ = 4
C	3	8	$dC,M1 = \sqrt{((3-3)^2 + (8-4)^2} = 4$	$dC,M2 = \sqrt{((3-7)2 + (8-4)^2}$ = 5.65
D	4	7	$dD,M1 = \sqrt{((4-3)2 + (7-4)2}$ = 3.16	$dD,M2 = \sqrt{((4-7)2 + (7-4)2}$ = 4.24
E	6	2	$dE,M1 = \sqrt{((6-3)^2 + (2-4)^2}$ = 3.6	$dE,M2 = \sqrt{((6-7)2 + (2-4)^2}$ = 2.23
F	6	4	$dF,M1 = \sqrt{((6-3)2 + (4-4)^2}$ = 3	$dF,M2 = \sqrt{((6-7)^2 + (4-4)2}$ = 1
G	7	3	$dG,M1 = \sqrt{((7-3)2 + (3-4)2}$ = 4.12	$_{dG,M2} = \sqrt{((7-7)2 + (3-4)^2}$ = 1
H	7	4	$dH,M1 = \sqrt{((7-3)2 + (4-4)^2}$ = 4	0

Object	X1	X2	Distance to medoid 1	Distance to medoid 2
I	8	5	$d_{I,M1} = \sqrt{((8-3)^2 + (5-4)^2} = 5.1$	$d_{I,M2} = \sqrt{((8-7)^{2+(5-4)2}} = 1.414$
J	7	6	$d_{J,M1} = \sqrt{(7-3)2 + (6-4)2} = 4.47$	$d_{J,M2} = \sqrt{((7-7)^{2+(6-4)2}} = 2$

Information:
Closest Distance =
Cluster members 1 = A, B, C, and D
Cluster members 2 = E, F, G, H, I, and J
Total closest distance to medoid point per cluster member =

2.23 + 0 + 4 + 3.16 + 2.23 + 1 + 1 + 0 + 1.414 + 2 = **17.034**

Then a comparison process is carried out on the total cost of the old distance (iteration 1) with the total cost of the new distance (iteration 2). Then it was known:

Total length = **27.75**
Total new distance = **17,034**

From the results of the calculation, it is known that the total cost of the old > the total cost of the new one, then a swap is carried out at the medoid point that has a smaller total cost, so that the medoid point in the second iteration will become the new medoid point and continue the iteration process to the next stage. Then re-select the medoid points between the data points with k=2 at random, namely points B(3,4) and H(7,3) and calculate the distance. Tabel 5 Presenting the process of calculating the distance between iterated data points 3.

Tabel 5. Third iteration distance calculation

Object	X1	X2	Distance to medoid 1	Distance to medoid 2
A	2	6	$d_{A,M1} = \sqrt{((2-3)^2 + (6-4)^2} = 2.23$	$d_{A,M2} = \sqrt{((2-7)^2 + (6-3)^2} = 5.83$
B	3	4	0	$d_{B,M2} = \sqrt{((3-7)^2 + (4-3)^2} = 4.12$
C	3	8	$d_{C,M1} = \sqrt{((3-3)^2 + (8-4)^2} = 4$	$d_{C,M2} = \sqrt{((3-7)^2 + (8-3)^2} = 6.4$
D	4	7	$d_{D,M1} = \sqrt{((4-3)^2 + (7-4)^2} = 3.16$	$d_{D,M2} = \sqrt{(4-7)^2 + (7-3)^2} = 5$
E	6	2	$d_{E,M1} = \sqrt{((6-3)^2 + (2-4)^2} = 3.6$	$d_{E,M2} = \sqrt{((6-7)^2 + (2-3)^2} = 1.414$

Object	X1	X2	Distance to medoid 1	Distance to medoid 2
F	6	4	$dF,M1 = \sqrt{((6-3)^2 + (4-4)^2} = 3$	$dF,M2 = \sqrt{((6-7)^2 + (4-3)^2} = 1.414$
G	7	3	$dG,M1 = \sqrt{((7-3)^2 + (3-4)^2} = 4.12$	0
H	7	4	$dH,M1 = \sqrt{((7-3)^2 + (4-4)^2} = 4$	$dH,M2 = \sqrt{((7-7)^2 + (4-3)^2} = 1$
I	8	5	$dI,M1 = \sqrt{((8-3)^2 + (5-4)^2} = 5.1$	$dI,M2 = \sqrt{((8-7)^2 + (5-3)^2} = 2.23$
J	7	6	$dJ,M1 = \sqrt{((7-3)^2 + (6-4)^2} = 4.47$	$dJ,M2 = \sqrt{((7-7)^2 + (6-3)^2} = 3$ 3

Information:
Closest Distance =
Cluster members 1 = A, B, C, and D
Cluster members 2 = E, F, G, H, I, and J
Total closest distance to medoid point per cluster member =

2.23 + 0 + 4 + 3.16 + 1.414 + 1.414 + 0 + 1 + 2.23 + 3 = **18.448**

Then a comparison process is carried out on the total cost of the old distance (iteration 1) with the total cost of the new distance (iteration 2).
Known:
Total length = **17,034**
Total new distance = **18,448**

From the results of the calculation, it is known that the total cost of the old < the total cost of the new one, so the calculation of distance and iteration is stopped.

3.3. Implementation of K-Medoids Clustering

In this study, before entering the clustering stage, it is necessary to have an optimal hyperparameter value to find out how many clusters will be formed from the existing data set. One of the other factors that affects the need to use PCA in this study is the results of elbow data plotting using data that has not gone through the data reduction stage and the results are obtained that the data graph does not increase or decrease or it can be said that it is in a stagnant condition, so that with this condition it is not possible to determine the optimal k point for the clustering process. Figure 3 shows a comparison of plotting data using the Elbow Method before and after the application of the PCA method.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

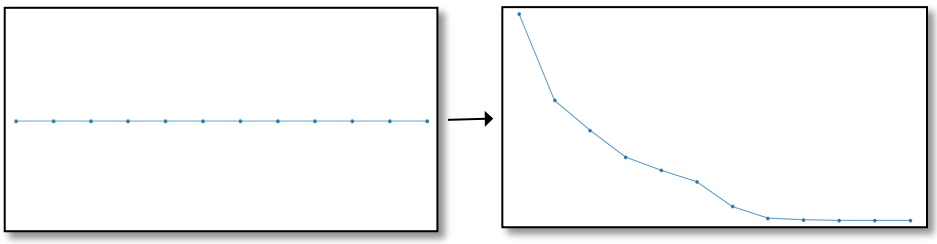

Figure 3. Comparison of Elbow Method before and after PCA application

Figure 4 shows the results of the elbow method plot using 7 main components of the feature reduction results. The results of the elbow plot show that the elbow point is not so visible, so the Silhouette Score value is needed to help determine the k value.

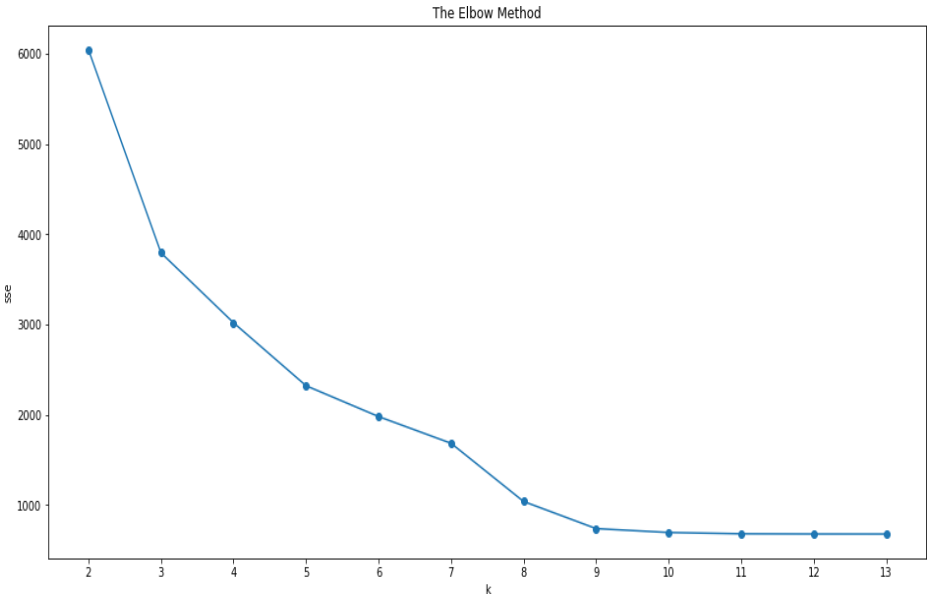

Figure 4. Data plot results using Elbow Method

Figure 5 shows the results of the Silhouette Score plot. Based on the results, the highest Silhouette value and close to 1 is 9, so it is set as the optimal k value.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

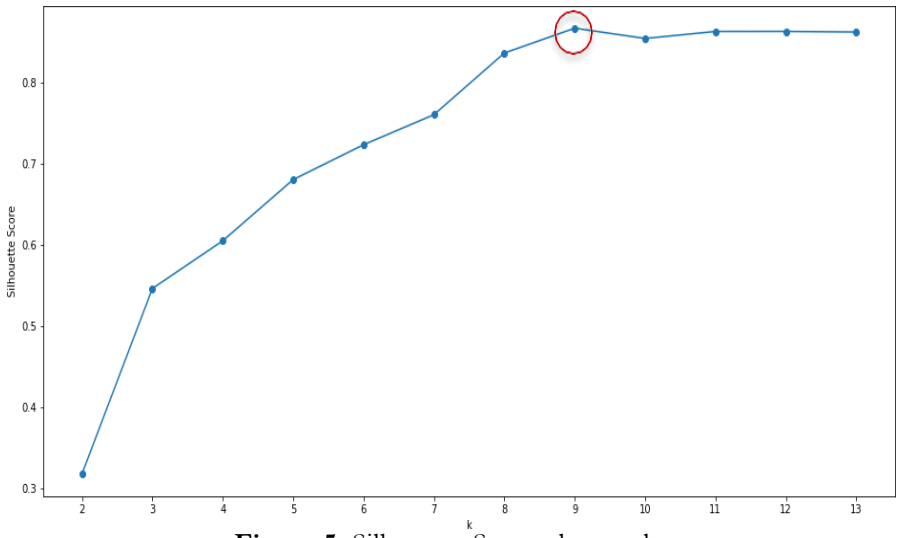

Figure 5. Silhouette Score plot results

Tabel 6 presents the Silhouette Score values for each k value in the plot. For the value of k=9, it results in a Silhouette Score value of 0.867 which means that a strong structure in the cluster has been found.

Tabel 6. Silhoutte score for each k value

K	Silhouette Score
2	0.137
3	0.546
4	0.605
5	0.680
6	0.723
7	0.760
8	0.836
9	0.867
10	0.854
11	0.863
12	0.863
13	0.862

The clustering process begins by determining the initialization of k as many as 9 as the medoid (center) point. Perform initialization for the initial stage. Next, calculate the distance between medoids and non-medoid objects using the Euclidean Distance measurement metric, then group non-medoid objects based on the nearest medoids. After that, the total cost is calculated by calculating the total distance from each member to their own cluster. At this stage, a total cost of

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3277.85 is generated. The next stage is to randomly select data points in each cluster as medoid candidates, then calculate the total cost value of the medoid candidate points obtained from the calculation of euclidean distances and obtain 2809.60.

The comparison will be made on the initial total cost with the new total cost, if the new total cost is smaller than the initial (old) total cost, then a swap will be carried out and the medoid candidate point will be set as a new medoid point. Based on the results, the total cost of the old is greater than the total cost of the new one, so a swap is carried out so that the total cost of the old = total cost of the new Iteration will continue until the medoid does not change. The final medoid point was obtained with the lowest total cost of 737.59. Tabel 7 presents the medoid points and the total cost generated by each iteration

Tabel 7. Results of medoid points and total cost per iteration

Iteration	Medoid Points	Cost
0	[[-0.52934046-0.40612733 1.08835567-0.48803499 -0.27982165-0.17980421-0.08838738] [-0.42930726 -0.20006394 0.07431186 -0.03593538 0.42913226 0.68838877 1.10559529] [-0.50638957 -0.41540787 -0.08328781 1.05540233 -0.65814652 0.01742966 0.07132291] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693] [0.90154648 0.01698963 -0.00660184 -0.00153591 0.00225016 0.00771021 0.00315091] [-0.39646172 -0.1729963 -0.01296945 0.04040416 0.24273666 0.19190408 -0.03347113] [-0.39646172 -0.1729963 -0.01296945 0.04040416 0.24273666 0.19190408 -0.03347113] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693]]	3277.85
1	[[-0.52934046 -0.40612733 1.08835567 -0.48803499 -0.27982165 -0.17980421 -0.08838738] [-0.42930726 -0.20006394 0.07431186 -0.03593538 0.42913226 0.68838877 1.10559529] [-0.50638957 -0.41540787 -0.08328781 1.05540233 -0.65814652 0.01742966 0.07132291] [-0.436736 -0.16934038 0.01003252 0.08869591 0.46195873 0.88485919 -0.8626044] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693] [0.90154648 0.01698963 -0.00660184 -0.00153591 0.00225016 0.00771021 0.00315091] [-0.37194123 0.01410228 0.02788344 0.01220014 0.10776238 0.09419249 -0.02477353] [-0.39646172 -0.1729963 -0.01296945 0.04040416 0.24273666 0.19190408 -0.03347113] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693]]	2809.60
2	[[-0.52934046 -0.40612733 1.08835567 -0.48803499 -0.27982165 -0.17980421 -0.08838738] [-0.42930726 -0.20006394 0.07431186 -0.03593538 0.42913226 0.68838877 1.10559529] [-0.50638957 -0.41540787 -0.08328781 1.05540233 -0.65814652 0.01742966 0.07132291] [-0.436736 -0.16934038 0.01003252 0.08869591 0.46195873 0.88485919 -0.8626044] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693] [0.90154648 0.01698963 -0.00660184 -0.00153591 0.00225016 0.00771021 0.00315091] [-0.68304537 1.18547402 -0.04467222 -0.02432686 -0.12356872 -0.07139693 0.01457543] [-0.52607463 -0.44605106 -0.72145784 -0.53750083 -0.17214235 -0.12125367 -0.02077002] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693]]	1190.64

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Iteration	Medoid Points	Cost
3	[[-0.52934046 -0.40612733 1.08835567 -0.48803499 -0.27982165 -0.17980421 -0.08838738] [-0.42930726 -0.20006394 0.07431186 -0.03593538 0.42913226 0.68838877 1.10559529] [-0.50638957 -0.41540787 -0.08328781 1.05540233 -0.65814652 0.01742966 0.07132291] [-0.436736 -0.16934038 0.01003252 0.08869591 0.46195873 0.88485919 -0.8626044] [-0.39646172 -0.1729963 -0.01296945 0.04040416 0.24273666 0.19190408 -0.03347113] [0.90154648 0.01698963 -0.00660184 -0.00153591 0.00225016 0.00771021 0.00315091] [-0.68304537 1.18547402 -0.04467222 -0.02432686 -0.12356872 -0.07139693 0.01457543] [-0.54291354 -0.4804519 -0.81120519 -0.63834354 -0.19236029 -0.16037677 -0.04079594] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693]]	1125.58
4	[[-0.52934046 -0.40612733 1.08835567 -0.48803499 -0.27982165 -0.17980421 -0.08838738] [-0.42930726 -0.20006394 0.07431186 -0.03593538 0.42913226 0.68838877 1.10559529] [-0.50638957 -0.41540787 -0.08328781 1.05540233 -0.65814652 0.01742966 0.07132291] [-0.436736 -0.16934038 0.01003252 0.08869591 0.46195873 0.88485919 -0.8626044] [-0.46460249 -0.24407915 0.02230682 0.3713019 1.01377474 -0.798495 -0.01986024] [0.90154648 0.01698963 -0.00660184 -0.00153591 0.00225016 0.00771021 0.00315091] [-0.68304537 1.18547402 -0.04467222 -0.02432686 -0.12356872 -0.07139693 0.01457543] [-0.54291354 -0.4804519 -0.81120519 -0.63834354 -0.19236029 -0.16037677 -0.04079594] [-0.43418123 -0.20962602 0.02226348 0.20116676 0.22861861 0.23158176 -0.11150693]]	742.18
5	[[-0.52934046 -0.40612733 1.08835567 -0.48803499 -0.27982165 -0.17980421 -0.08838738] [-0.42930726 -0.20006394 0.07431186 -0.03593538 0.42913226 0.68838877 1.10559529] [-0.50638957 -0.41540787 -0.08328781 1.05540233 -0.65814652 0.01742966 0.07132291] [-0.436736 -0.16934038 0.01003252 0.08869591 0.46195873 0.88485919 -0.8626044] [-0.46460249 -0.24407915 0.02230682 0.3713019 1.01377474 -0.798495 -0.01986024] [0.90154648 0.01698963 -0.00660184 -0.00153591 0.00225016 0.00771021 0.00315091] [-0.68304537 1.18547402 -0.04467222 -0.02432686 -0.12356872 -0.07139693 0.01457543] [-0.54291354 -0.4804519 -0.81120519 -0.63834354 -0.19236029 -0.16037677 -0.04079594] [-0.39646172 -0.1729963 -0.01296945 0.04040416 0.24273666 0.19190408 -0.03347113]]	737.60

In Figure 6, data samples identified as part of cluster 2 are displayed, illustrating the specific characteristics or patterns within the group. This figure provides a clearer picture of the distribution, relationships, and differences between the data in cluster 2, allowing for a deeper understanding of how the data are interrelated and how the group is formed based on the analyzed criteria or attributes.

Figure 6. Sample cluster 2 data

(501, 5)

	pasal	jenis_kendaraan	pasal_hukuman	pasal_tuntutan	cluster
76	pasal 281 jo pasal 77 ayat (1)	SEPEDA MOTOR	[pasal 281]	[pasal 77 ayat (1)]	2
93	pasal 281 jo pasal 77 ayat (1)	SEPEDA MOTOR	[pasal 281]	[pasal 77 ayat (1)]	2
94	pasal 281 jo pasal 77 ayat (1)	SEPEDA MOTOR	[pasal 281]	[pasal 77 ayat (1)]	2
103	pasal 281 jo pasal 77 ayat (1),pasal 287 ayat...	SEPEDA MOTOR	[pasal 281, pasal 287 ayat (1)]	[pasal 77 ayat (1), pasal 106 ayat (4) huruf a]	2
121	pasal 281 jo pasal 77 ayat (1)	SEPEDA MOTOR	[pasal 281]	[pasal 77 ayat (1)]	2

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.4. Cluster Results

Based on the results of the grouping carried out with K-Medoids, 9 types of clusters were produced. Tabel 8 showing the results of observations on each cluster of traffic violation types. By Tabe 8 Further explanation has been presented in 4.2 which discusses the results of cluster observations.

Tabel 8. Observation of clustering results

Cluster	Sum	Penalty Article	Article of Claim	Vehicle Type
2	501	PS281	PS77(1)	Motorcycle Mini Bus Goods Car/MKL Large Truck Pick Up /Passenger Car
		Ps280, Ps281	Ps68(1), Ps77(1)	
		Ps281, Ps291(1)	Ps77(1), Ps106(8)	
		Ps281, Ps285(1)	Ps77(1), Ps106(3)	
		Ps281, Ps291(2)	Ps77(1), Ps106(8)	
		Ps283, Ps281	Ps106(1), Ps77(1)	
		Ps281, Ps288(1)	Ps77(1), Ps106(5)a	
		Ps281, Ps297	Ps77(1), Ps115b	
		Ps281, Ps287(2)	Ps77(1), Ps106(4)c	
		Ps281, Ps288(2)	Ps77(1), Ps106(5)b	
		Ps291(1), Ps281	Ps106(8), Ps77(1)	
		Ps281, Ps287(1)	Ps77(1), Ps106(4)a	
		Ps281, Ps287(5)	Ps77(1), Ps106(4)g or Ps115a	
		Ps287(2), Ps281	Ps106(4)c, Ps77(1)	
		Ps285(1), Ps281	Ps106(3), Ps77(1)	
		Ps291(2), Ps281	Ps106(8), Ps77(1)	
3	295	Ps287(2)	PS106(4)C	Mini Bus Motorcycle MKL/Passenger Car Goods Car/Pick Up Sedan
4	339	PS280	PS68(1)	Motorcycle Mini Bus
		Ps280, Ps285(1)	Ps68(1), Ps106(3)	
		Ps280, Ps288(1)	Ps68(1), Ps70(2)	

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Cluster	Sum	Penalty Article	Article of Claim	Vehicle Type
		Ps280, Ps287(2)	Ps68(1), Ps106(4)c	Goods Car/MKL Pick Up /Passenger Car
		Ps280, Ps287(5)	Ps68(1), Ps106(4)g or Ps115a	
		Ps280, Ps287(1)	Ps68(1), Ps106(4)a	
5	1964	PS287(1)	PS106(4)A	Motorcycle Car Goods/Pick Up Mini Bus MKL/Passenger Car Small Truck Large Truck Miscellaneous
		PS287(1)	Ps106(4)a, Ps106(4)b	
		Ps285(1), Ps287(1)	Ps106(3), Ps106(4)a, Ps106(5)b	
		Ps287(1), Ps288(2)	Ps106(4)a, Ps106(4)b, PS106(5)B	
		Ps287(1), Ps291(1)	Ps106(4)a, Ps106(8)	
		Ps287(1), Ps291(2)	Ps106(4)a, Ps106(8)	
		Ps287(1), Ps288(1)	Ps106(4)a, Ps70(2)	
6	768	PS289	PS106(6)	Mini Bus Freight Car/Pick Up MKL Motorcycle/Jeep Passenger Car Others Sedan
7	665	Ps291(1)	PS106(8)	Motorcycle Mini Bus Miscellaneous MKL/Passenger Car
		Ps285(1), Ps291(1)	Ps106(3), Ps106(8)	
		Ps291(1), Ps297	Ps106(8), Ps115b	
		Ps291(1), Ps291(2)	Ps106(8), Ps106(8)	
		PS291(2)	PS106(8)	
		Ps285(1), Ps291(2)	Ps106(3), Ps106(8)	
		Ps291(2), Ps292	Ps106(8), Ps106(9)	
		Ps287(3), Ps291(2)	Ps106(4)e, Ps106(8)	

3.5. Cluster visualization using Word Cloud

The visualizations for each cluster are presented in 3 different types of Word Clouds. Figure 7 shows a Word Cloud that illustrates words containing data on

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

penalty articles in cluster 8. Based on the illustration of Word Cloud, this cluster is dominated by 2 types of penalty articles, namely article 288 paragraph (3) with a data frequency of 194 data and article 285 paragraph (1) as much as 221 data.

Figure 7. Word Cloud abbreviations for penal articles

Figure 8 shows a Word Cloud that illustrates a word containing the data of the claim article included in cluster 8. Based on the illustration of Word Cloud, this cluster is dominated by 2 types of demand articles, namely article 288 paragraph (3) as many as 194 data and article 285 paragraph (1) with a total of 221 data.

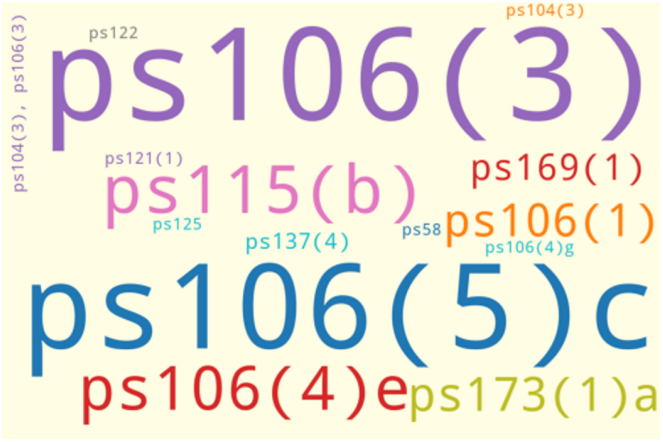

Figure 8. Word Cloud abbreviation of the claim article

Figure 9 shows a Word Cloud that illustrates words that contain vehicle type data in cluster 8. Based on the illustration of Word Cloud, this cluster is dominated by the types of vehicles that often violate in the cluster are motorcycles with 283 data and large trucks with 178 data.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

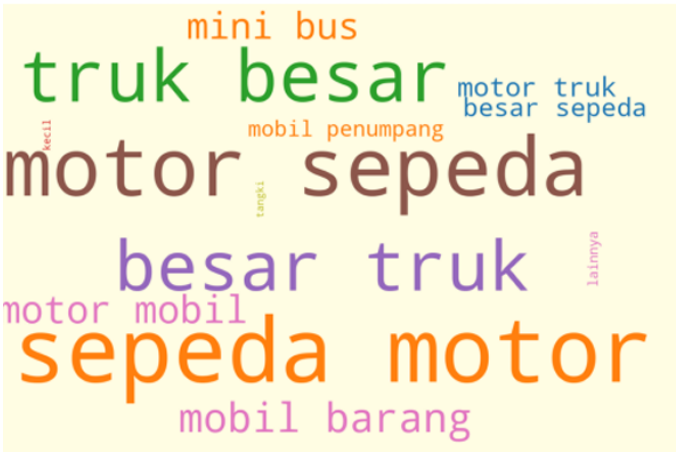

Figure 9. Word Cloud vehicle type

3.6. Discussion

The clustering analysis of traffic violations using the K-Medoids method provides a structured way to classify different types of infractions based on legal articles and vehicle types. This approach is crucial for understanding patterns in traffic violations, helping authorities develop more effective enforcement strategies. The study revealed that certain violations tend to cluster around specific vehicle categories, allowing for a data-driven approach to law enforcement. The results of the clustering process indicate a strong relationship between traffic offenses, legal regulations, and vehicle types, providing valuable insights into traffic behavior patterns.

A significant observation in the clustering process is the overlapping nature of violation data, where the same penalty article can refer to different demand articles and vice versa. Additionally, some cases involve multiple legal articles, making it difficult to assign them definitively to a single cluster. This complexity necessitated the implementation of data mining techniques, particularly clustering algorithms, to identify dominant patterns among traffic offenses. K-Medoids Clustering was chosen due to its robustness in handling non-uniform data distributions and its ability to minimize the influence of outliers compared to K-Means. To further refine the clustering process, Principal Component Analysis (PCA) was applied to reduce data dimensionality while preserving 80% of the original variance. This dimensionality reduction allowed for more efficient processing and better-defined clusters.

The clustering process identified nine distinct clusters, each representing different categories of traffic violations. Cluster 0 predominantly includes motorists who fail to provide complete vehicle registration documents (STNK), violating Article

288(1) jo Article 106(5)a. The majority of offenders in this cluster are motorcyclists and R4 vehicle drivers, including mini-buses and passenger cars. Similarly, Cluster 1 contains motorists who cannot present a valid driver's license, violating Article 288(2) jo Article 106(5)b. This group is composed mainly of motorcyclists and R4 vehicle operators, such as pick-up trucks, passenger cars, and mini-buses.

Cluster 2 is primarily dominated by motorcyclists who drive without a valid license, violating Article 281 jo Article 77(1). This category is particularly concerning as unlicensed drivers pose significant risks to road safety. Meanwhile, Cluster 3 includes violations related to traffic signaling devices, as defined by Article 287(2) jo Article 106(4)c. The most frequent offenders in this category are motorcyclists, followed by R4 vehicles, including mini-buses, passenger cars, pick-up trucks, and sedans. These violations contribute significantly to road congestion and accidents, as failure to obey traffic signals disrupts the flow of vehicles and endangers pedestrians.

A distinct pattern emerges in Cluster 4, which comprises violations related to the absence of a Motor Vehicle Sign (TKB) as stipulated in Article 280 jo Article 68(1). This cluster includes both motorcycles and R4 vehicles, such as pick-up trucks, mini-buses, and passenger cars. The absence of a vehicle identification sign can lead to difficulties in law enforcement, as vehicles without proper identification may be involved in illegal activities without being easily traced.

The most significant cluster, Cluster 5, consists of 1,964 recorded cases and includes violations related to order signs, prohibition signs, and road markings under Article 287(1) jo Article 106(4)a and b. This cluster is largely composed of motorcycles (1,525 cases) and R4 vehicles (438 cases), including mini-buses, passenger cars, pick-up trucks, large trucks, and small trucks. The high number of violations in this category indicates a widespread disregard for road signage and markings, which are essential for maintaining organized traffic flow and preventing accidents.

Another key finding is in Cluster 6, which focuses on seatbelt violations under Article 289 jo Article 106(6). This cluster is dominated by R4 vehicle drivers, including 431 passenger cars, 265 mini-buses, 14 pick-up trucks, 11 sedans, and 10 jeeps. Seatbelt usage is a critical factor in road safety, and non-compliance significantly increases the risk of fatal injuries in accidents. The concentration of seatbelt violations in this cluster suggests that awareness campaigns and stricter enforcement are necessary to promote compliance among four-wheeled vehicle drivers and passengers.

In Cluster 7, violations related to helmet usage are the most prevalent. This includes Article 291(1) jo Article 106(8), which penalizes motorcyclists for failing

to wear a Standard National Indonesia (SNI) certified helmet, and Article 291(2) jo Article 106(8), which applies to riders who allow passengers to ride without helmets. This cluster contains 659 recorded cases, all of which involve motorcycles. The high frequency of helmet-related violations indicates a serious risk to motorcyclists' safety, as helmets are proven to reduce the severity of head injuries in accidents.

Cluster 8 captures two specific types of violations. The first concerns car drivers who do not possess periodic permit certificates, violating Article 288(3) jo Article 106(5)c. The second involves motorcycles that do not meet roadworthiness standards, violating Article 285(1) jo Article 106(3). These roadworthiness violations include the absence of essential safety components such as rearview mirrors, horns, headlights, brake lights, turn signals, reflectors, speedometers, exhausts, and adequate tire groove depth. The dominant vehicle types in this cluster are motorcycles and large trucks, both of which require strict technical inspections to ensure safety on the road.

To determine the optimal number of clusters, the Elbow Method and Silhouette Score were employed. The results indicated that the best clustering structure was achieved with k=9, yielding a Silhouette Score of 0.867, which signifies well-defined and strongly cohesive clusters. The clustering process started with initial medoid selection, followed by iterative calculations of Euclidean distances between data points. The total cost of clustering was progressively minimized, starting from 3,277.85 in the first iteration and ultimately converging to 737.59, which confirmed the effectiveness of the medoid selection process.

To further enhance the interpretation of clustering results, Word Cloud visualizations were generated for each cluster. These visualizations provided a graphical representation of the most frequently occurring penalty articles, demand articles, and vehicle types. For example, the Word Cloud for Cluster 8 highlighted Article 288(3) and Article 285(1) as the most dominant violations, with motorcycles and large trucks being the most frequently involved vehicle types. This method of visualization proved to be an effective tool in summarizing the key trends within each cluster, offering a more intuitive understanding of traffic violation patterns.

The K-Medoids clustering method proved to be an effective tool in classifying traffic violations based on legal articles and vehicle types. The findings from this study provide significant insights for traffic law enforcement agencies, enabling them to implement targeted interventions and policy adjustments. By identifying the most prevalent traffic violations and their associated vehicle types, authorities can allocate resources more efficiently, enforce regulations more effectively, and ultimately work toward improving road safety and traffic management.

4. CONCLUSION

Based on the results of this study, the K-Medoids algorithm has proven to be effective in clustering traffic violation types in Makassar City. By using the K-Medoids method, 9 clusters were successfully formed, and these clusters have good quality with a Silhouette Score of 0.867. This value indicates that the formed clusters are compact and well-separated, meaning the clustering process can be considered successful. The use of K-Medoids was prioritized in this study due to its robustness against outliers and its suitability for categorical data, such as vehicle type and violation type. In this study, the second objective, which is to visualize the results of clustering traffic violations, was also successfully achieved. The visualization process was carried out using three different types of Word Clouds, each representing important aspects of the violation data, including the abbreviation of the penal code, the junction to the charge, and the type of vehicle. This visualization approach makes complex information easier to understand and engaging, providing a clear overview of the distribution of traffic violations in Makassar City.

The visualization results show that motorcycles and minibuses are the types of vehicles most frequently involved in traffic violations. This is important for authorities to understand so that they can determine appropriate policies and measures to improve safety and address the violations. This visualization also helps identify areas with higher violation rates, allowing preventive actions to be focused on the types of vehicles most frequently involved in these violations. Overall, this study successfully achieved its objectives, which were to cluster the types of traffic violations using the K-Medoids algorithm. With clear clustering results and informative visualizations, this research provides valuable contributions to traffic violation analysis in Makassar City, which can be used by authorities to design more effective policies and strategies to enhance traffic safety.

REFERENCES

[1] T. Mussweiler, "Focus of Comparison as a Determinant of Assimilation Versus," *Personal. Soc. Psychol. Bull.*, vol. 27, pp. 38–47, 1997, doi: 10.1145/3054925.

[2] S. Tufféry, "Statistical and Data Mining Software," *Data Min. Stat. Decis. Mak.*, pp. 111–166, 2011, doi: 10.1002/9780470979174.ch5.

[3] F. R. Senduk, I. Indwiarti, and F. Nhita, "Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm," *Indones. J. Comput.*, vol. 4, no. 3, pp. 65–76, 2019, doi: 10.21108/indojc.2019.4.3.359.

[4] J. Ha, M. Kambe, and J. Pe, "Data Mining: Concepts and Techniques," *Data Min. Concepts Tech.*, pp. 1–703, 2011, doi: 10.1016/C2009-0-61819-5.

[5] T. B. Ambo, J. Ma, and C. Fu, "Investigating influence factors of traffic violation using multinomial logit method," *Int. J. Inj. Contr. Saf. Promot.*, vol. 28, no. 1, pp. 78–85, 2020, doi: 10.1080/17457300.2020.1843499.
[6] E. H. S. Atmaja, "Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta," *Int. J. Appl. Sci. Smart Technol.*, vol. 1, no. 1, pp. 33–44, 2019, doi: 10.24071/ijasst.v1i1.1859.
[7] P. Dangeti, Statistics for Machine Learning, Packt Publishing Ltd., 2017.
[8] S. Dua and X. Du, Data Mining and Machine Learning in Cybersecurity, *CRC Press*, 2016.
[9] B. Johnston, A. Jones, and C. Kruger, Applied Unsupervised Learning with Python: Discover Hidden Patterns and Relationships in Unstructured Data with Python, *Packt Publishing Ltd.*, 2019.
[10] O. Maimon and L. Rokach, Eds., Data Mining and Knowledge Discovery Handbook, vol. 2, *Springer*, New York, 2005.
[11] A. Malik and B. Tuckfield, Applied Unsupervised Learning with R: Uncover Hidden Relationships and Patterns with K-Means Clustering, Hierarchical Clustering, and PCA, *Packt Publishing Ltd.*, 2019.
[12] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-Medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.
[13] T. Thinsungnoen, N. Kaoungkub, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, "The clustering validity with silhouette and sum of squared errors," *Learn.*, vol. 3, no. 7, pp. 44–51, 2015.