Eduvest – Journal of Universal Studies
Volume 5 Number 8, Agustus, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

Application of Clustering Method for Segmentation of Traffic Accident Profiling With K-Means And K-Medoids Case Study of Toll Bakauheni – Terbanggi Besar

Anton Haryanto, Indra.
Universitas Budi Luhur, Indonesia
Email: 2211601881@student.budiluhur.ac.id**,** indra@budiluhur.ac.id

ABSTRACT

Traffic accidents remain a critical issue requiring in-depth analysis to identify patterns and risk factors. This study applies clustering methods in data mining techniques to segment accident profiles along the Bakauheni-Terbanggi Besar toll road. The study employs two clustering algorithms—K-Means and K-Medoids—to classify accident data into three zones: safe zone, alert zone, and danger zone. Cluster quality was evaluated using three metrics: Silhouette Score, Davies-Bouldin Index (DBI), and Purity. K-Means clustering results showed a distribution of 29.7% danger zone, 38.9% safe zone, and 31.4% alert zone. Meanwhile, K-Medoids yielded 32.8% danger zone, 34.2% safe zone, and 33.1% alert zone. Evaluation metrics demonstrated K-Means' superiority (Silhouette Score: 0.365; DBI: 1.117; Purity: 0.617) over K-Medoids (Silhouette Score: 0.273; DBI: 1.440; Purity: 0.436). The evaluation results indicate that the K-Means method delivers better clustering performance with a higher Silhouette Score and lower DBI values, signifying superior cluster quality. The higher Purity value in K-Means also suggests more homogeneous clusters. These findings provide valuable references for decision-making and accident mitigation strategies on toll roads.

KEYWORDS *Clustering, Traffic Accidents, Data Mining, Clustering, K-Means, K-Medoids, Silhoute Score, DBI, Purity*

INTRODUCTION

The number of reported traffic accidents on the Bakauheni–Terbanggi Besar Toll Road underscores the complexity and challenges inherent in managing safety on one of Indonesia's strategic corridors (Rahmadana & Putra, 2025). These incidents stem from a combination of factors, including fluctuating traffic density, driver behavior, road condition variability, and dynamic weather patterns (Hermawan, 2024). In Indonesia's current digital transformation era, leveraging information technology to present traffic accident data in a clear and educative manner has become essential for raising public awareness and fostering safer behavior (Chaniago et al., 2024). The Bakauheni–Terbanggi Besar Toll Road also plays a pivotal socio-economic role by serving as the gateway to the Trans-Sumatra Toll Road and boosting regional economic activity (Hu, 2024). Operationally, the toll road is managed by PT Hutama Karya (Persero), while ownership lies with the Indonesia Investment Authority (INA) via their BUJT

partnership—marking a growing trend of sovereign wealth investment in Indonesia's infrastructure (Reuters, 2024; INA platform report, 2024).

The Bakauheni–Terbanggi Besar Toll Road, stretching 140 kilometers, strategically connects four regencies in Lampung Province: South Lampung, East Lampung, Pesawaran, and Central Lampung. This infrastructure plays a pivotal role in enhancing regional integration and accessibility, significantly reducing logistics costs and travel time across Sumatra and Java (Mulyono et al., 2021). The road also serves as a vital artery connecting directly to Bakauheni Port, facilitating smoother inter-island commodity distribution (Prawira & Nugroho, 2020). Its strategic development has supported industrial expansion and tourism growth in the area (Astuti et al., 2022). In turn, this has contributed to rural development by increasing access to markets and employment opportunities (Wahyuni et al., 2023). According to the Ministry of Public Works and Housing, toll roads such as this are essential in driving equitable economic growth across Indonesia's archipelagic geography (KemenPUPR, 2022). Moreover, such connectivity enhances supply chain efficiency, particularly for perishable agricultural commodities (Hasanah et al., 2022). Recent studies also emphasize the multiplier effect of toll road development on local economies through the stimulation of small and medium enterprises (Putri & Sari, 2021).

Based on Law of the Republic of Indonesia No. 22 of 2009 concerning Road Traffic and Transportation, it is stated that a traffic accident is an event that occurs on the highway, both unexpected and accidental, which can result in victims as well as property losses (Purwaningsih, 2019). The high number of traffic accidents is a serious problem that must be anticipated by the community in various areas, therefore the importance of education on traffic accident prevention. Including in the area of the Bakauheni - Terbanggi Besar Toll Road Section. To overcome this problem effectively. A systematic and data-based approach is needed in analyzing traffic accident patterns. The high rate of traffic accidents can create a sense of insecurity and disturb order. It can even cause many victims. Therefore, efforts to prevent and handle traffic accidents are very important to be carried out.

Efforts to handle the level of traffic accident cases on the Bakauheni - Terbanggi Besar Toll Road section are still not effective because they still use manual monitoring and mapping of existing traffic accident-prone areas. So that to reduce the high number of traffic accidents is still not optimal. Therefore, the approach that can be used to overcome the problem of traffic accidents is to use data mining. Data mining is the process of discovering patterns and useful information from large data sets. The main purpose of data mining is to identify hidden relationships. Trend. And patterns that can be used for better decision-making in various fields (Rosiana, Mohsa, Fadila, & Jaman, 2023).

With the clustering method on traffic accident data, patterns, and factors that affect it, a clustering model can be built that can help the Bakauheni - Terbanggi Besar Toll Road Section anticipate and handle traffic accidents that occur. The development of information technology and data science has opened up new opportunities in analyzing and understanding traffic accident phenomena in more depth. One potential method to be applied is clustering techniques, which are part of machine learning and data mining. This method allows the grouping of traffic accident data based on certain characteristics, so that it can produce more structured traffic accident profiles or segments.

The application of the deep clustering method can provide valuable insights for the manager of the Bakauheni - Terbanggi Besar Toll Road section in developing a more targeted traffic accident prevention and handling strategy. By understanding the patterns and characteristics of traffic accidents in various areas. Authorities can allocate resources more efficiently and design safety programs that match the traffic accident profile in each area. Therefore, this study aims to explore and apply the clustering method in analyzing traffic accident data in the area of the Bakauheni - Terbanggi Besar Toll Road Section. By using this technique. It is hoped that a more detailed and informative segmentation of traffic accident profiles can be produced. In order to support better decision-making in efforts to prevent and handle traffic accidents.

In the research carried out previously, the process of grouping truck fleet data based on productivity generated from the K-Means and K-Medoids Clustering methods has also been carried out so that the results of truck fleet data for grouping based on productivity have also been carried out. The results obtained from the clustering technique and K-Means and K-Medoids algorithms are that in this study the application of the K-Means and K-Medoids Algorithms, which is then validated by the results of the clusters formed. The Davies Bouldin Index as a method in cluster analysis produced a validity value of 0.67 for K-Means clustering and 1.78 for K-Medoids. Based on the validity value generated, the K-Means algorithm was selected to be implemented in the creation of a web-based vehicle fleet clustering application because it is most relevant to the lower DBI validity value than K-Medoids. Tests that have been carried out on clustering results on web applications have obtained a percentage of conformity of 97% both with the Rapidminer tool and with manual calculations.

The novelty of this study lies in its comprehensive evaluation of clustering techniques tailored to toll road conditions. Unlike prior research that often examines these methods in isolation, this study provides a systematic comparison to determine the most suitable algorithm for accident profiling. Furthermore, the findings offer actionable insights for toll road operators, enabling them to prioritize high-risk areas and implement targeted safety measures. The strategic importance of the Bakauheni-Terbanggi Besar Toll Road amplifies the urgency of this research, as its safety directly impacts regional mobility and economic growth. By leveraging clustering techniques, this study not only fills a methodological gap in traffic safety research but also provides a scalable framework for other toll roads in Indonesia and beyond. Therefore, the results of this research are expected not only to contribute to the development of more effective prevention and education strategies in the Bakauheni - Terbanggi Besar Toll Road area, but also to be a model for the application of advanced data analysis in the context of prevention for other areas in Indonesia.

The purpose of this research is to apply data mining techniques with clustering methods on traffic accident data to determine accident-prone areas and conduct analysis and evaluation based on the characteristics of clustering results from traffic accident data to see its quality and accuracy. This research is expected to provide significant benefits, including providing new insights into the patterns and characteristics of traffic accidents in the Bakauheni – Terbanggi Besar Toll Road area through more structured profile segmentation, enriching literature on the application of clustering methods in the context of traffic accident prevention, and potentially increasing public awareness through education and handling of traffic accidents in a more targeted manner.

RESEARCH METHOD

The research method used in this study was quantitative, involving data processing to obtain information for clustering traffic accident-prone areas. This study applied data mining techniques using clustering methods to classify traffic accident zones along the Bakauheni–Terbanggi Besar toll road section. The data processing followed the six stages of the Cross Industry Standard Process for Data Mining (CRISP-DM) model: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Dhewayani et al., 2022). CRISP-DM provides a structured framework to identify meaningful patterns in data through its data modeling process.

The study analyzed traffic accident data from PT. Bakauheni Terbanggi Besar for the period 2019 to 2021, collected from all branches of the company's operations. The population included all recorded traffic incidents during this period. The sample consisted of documented accident data containing details such as the date and time, location, accident type, cause, number of victims, and vehicle type. Sampling was comprehensive to ensure representativeness. After data collection, cleaning was performed to remove incomplete or irrelevant records. The cleaned data was then analyzed using clustering techniques, primarily the K-Means algorithm, to identify accident patterns based on time, location, and type. The goal was to generate actionable insights to improve traffic safety.

Data collection combined direct observation and literature review. The observation involved examining the data processing system and gathering relevant data related to traffic accidents to develop a clustering model. The literature review utilized journals, articles, and online sources to support problem definition and provide theoretical background.

RESULT AND DISCUSSION

1. Data Collection

In this study, the data collection stage is a very important first step to build a clustering model in segmentation profiling traffic accidents on the Bakauheni – Terbanggi Besar Toll Road. The data collected comes from a variety of credible sources that provide official accident reports that include details of the location, time of the incident, type of vehicle, number of victims, and cause of the accident. In addition, data is also obtained from toll operators, such as Jasa Marga, who record information on traffic density and road conditions when accidents occur. Which provides information about road infrastructure, traffic signs, and safety policies implemented on the toll road. Environmental factors are also a concern in this study, so data on weather conditions at the time of the accident are collected to see how they affect the risk of accidents.

The variables collected in this study include various aspects related to traffic accidents. From a spatial perspective, the accident location data is identified based on the kilometers of the toll road where the accident occurred. Meanwhile, in terms of temporal, the time of the event is recorded in detail, including the day, date, month, year, and time of the event. The characteristics of the vehicles involved are also analyzed, be it private cars, buses or trucks. Other external factors, such as weather (rain, fog, or sunny). The impact of the accident was also analyzed by looking at the number of victims, the severity of the injuries, and the damage to the vehicle caused.

The data collection process is carried out through various methods to ensure the accuracy and completeness of information. Documentation studies are one of the main ways to obtain secondary data from previously documented accident reports. In addition, field observations are carried out to verify the real conditions in accident-prone locations to understand the factors that contribute to the incident. Interviews with toll operators were also conducted to gain additional insight into the causes of accidents Toll road traffic was also analyzed to identify accident patterns in more detail.

Once all the data is collected, the next step is the cleaning and pre-processing process so that the data is ready to be used in clustering analysis. This includes removing duplicate or inconsistent data, supplementing the missing data with interpolation methods or removing them if the missing number is too large, as well as normalizing each variable to have a uniform scale before being used in the K-Means and K-Medoids algorithms. In addition, checks are also carried out on outliers that can affect the results of clustering.

Through this systematic and structured data collection stage, it is hoped that the research can obtain valid and accurate data so that the results of clustering analysis can better describe the real conditions of traffic accidents on the Bakauheni – Terbanggi Besar Toll Road.

2. Data Understanding

At the data understanding stage, the process carried out is to collect data on traffic accidents that occurred on the Bakauheni – Terbanggi Besar Toll Road in the year period (2019-2021) through the following steps and stages to ensure a thorough understanding of the available data:

1) Identify data sources:
   a. Internal Source: Data obtained from accident reports compiled by Toll road operators, Traffic Control Centers and Toll Road Patrol Teams
   b. External Sources: Additional data can come from the Police, the Transportation Agency and the Central Statistics Agency.
   c. Data Period: The data collected is focused between 2019 and 2021 to obtain a relevant historical range.
2) Gathering basic information about data
   a. Data type: quantitative data such as the number of accidents, the number of victims and qualitative data such as the type of accident, weather conditions and time of occurrence.
   b. Data format: document-based data (manual or electronic reports) and files such as Excel or CSV.
   c. Location Coverage: the entire Bakauheni – Terbanggi Besar Toll section with segmentation per kilometer.
3) Assess data quality:
   a. Data completeness: ensure that the data covers all incidents during the period without any data being lost/
   b. Data consistency: check the data format such as the date of the incident, the type of vehicle or the category of the accident.
   c. Data accuracy: verifies data by matching reports from various sources.
4) Recognize early patterns and trends:

Analyze the initial stage to recognize the distribution of data such as monthly or annual trends in the number of accidents, accident-prone locations on toll sections, relationships between time and type of accidents.

A systematic approach at the data understanding stage is urgently needed to ensure that the data obtained is able to support an accurate analysis process and appropriate and effective decision-making in improving safety on the Bakauheni – Terbanggi Besar toll road.

3. Data Preparation

Traffic accident data from the Bakauheni – Terbanggi Besar toll road needs to be cleaned and prepared for analysis. The first step is to identify and handle missing or duplicate data, as well as ensure that the data format is consistent. Next, categorical variables such as vehicle type will be encoded into a numerical format. The normalization process is also carried out to ensure that all features have a balanced scale. In addition, the selection of relevant features, such as the time, location, and severity of the accident, is crucial to ensure the quality of the clusters formed. Thus, the data that has been prepared will be ready to be analyzed using the K-Means and K-Medoids methods.

1) Data Cleansing

At this stage, the data cleansing process is to ensure the quality of the data that will be used in the analysis. As an initial step, data cleansing aims to clean the data from various anomalies such as inconsistencies, entry errors and other imperfections that can affect clustering results

This process not only ensures that the data used is error-free, but also improves the quality of clustering results by reducing noise in the dataset. In the context of this study, clean and structured data will provide a solid basis for the application of the K-Means and K-Medoids methods, so that accident patterns and segmentation can be identified more accurately. Data cleansing, although often time-consuming, is a very important step to ensure the validity and reliability of research results.

2) Data Normalization

The next step is the normalization stage using the Standard Scaler normalization equation. The purpose of this normalization is to equalize the attributes and balance the calculations of the two types of clustering. Data normalization is a process used to change the value vulnerability of certain variables so that no data is too large or too small Normalization itself aims to improve the accuracy of the model built (Mumtazah & Sancoko, 2024).

Table 1. Normalization Results

No	Month	MILES	M	Position	Weather	Vehicle Type	Causes/Conditions	Time/Group
1	-0,404095	-1,561919	0,142514	-1,419048	-0,388375	-2,526734	0,465894	0,051102
2	-0,404095	-1,767661	-1,303587	-1,419048	-0,388375	-0,121381	0,465894	1,019351
3	-0,404095	-1,741943	-0,580536	-1,419048	-0,388375	-0,522273	-1,461943	-0,91715
4	-0,404095	-1,690508	-1,303587	0,318562	-0,388375	-0,522273	-1,461943	-0,91715

5	-0,404095	-1,099001	-0,580536	0,318562	-0,388375	-2,526734	-1,461943	-0,91715
6	-0,404095	1,369894	0,142514	0,318562	-0,388375	-0,522273	-1,461943	-0,91715
7	-0,404095	1,061282	0,865565	0,318562	-0,388375	-0,522273	-1,461943	-0,91715
8	-0,404095	-0,070253	-0,580536	0,318562	-0,388375	-0,522273	-1,461943	-0,91715
9	-0,694463	-1,639072	-0,942062	-1,419048	-0,388375	-2,526734	0,7872	0,051102
10	-0,467009	-0,841825	-0,942062	0,318562	-0,388375	-0,522273	0,7872	0,051102

Source: Primary data analysis, 2024

So the normalization results in the table above starting from the data preparation stage to the data normalization stage are obtained 360 data with 7 attributes, namely (Month, Kilometers, Meters, Position, Weather, Vehicle Type, Cause/Condition, and Time/Group) which will be used for the clustering stage.

3) Data Transformation

The data transformation in this study aims to convert the raw data of traffic accidents on the Bakauheni-Terbanggi Besar Toll Road into a format that is ready for analysis using the K-Means and K-Medoids clustering algorithms. The first stage is to identify and select relevant variables, such as the time of the incident, the location of the accident, the number of victims, the type of vehicle involved, the cause of the accident, as well as external variables such as road length and lighting conditions.

Furthermore, non-numeric variables, such as vehicle type or cause of accident, are converted into numerical format using encoding techniques. Data that has missing values is handled by imputation methods, such as replacing missing values using average, median, or mode. If the proportion of missing data is small, rows or columns with missing values can be deleted. The normalization process is carried out to ensure that each variable has the same scale, so that no variable dominates the distance calculation in the clustering algorithm.

After the transformation process is complete, the data is re-examined to ensure that there are no extreme outliers, the variable scale is uniform, and the data format is in accordance with the needs of the clustering algorithm. With these steps, traffic accident data that has undergone transformation is ready to be used in the clustering analysis process to identify accident profiles on the Bakauheni-Terbanggi Besar Toll Road.

4. Modeling (Clustering Implementation)

The Clustering modeling stage of the data processing stage produces 260 rows and seven data attributes to be used, namely (Position, Weather, Vehicle Type, Cause/Condition, Kilometers, Meters and Date). Then a clustering analysis will be carried out using the K-Means and K-Medoids Clustering algorithms. Before the clustering stage begins, a K-value search is carried out with the elbow method used as a method to determine the optimal number of clusters in the grouping of traffic accident data which includes 360 data in the clustering of K-Means and K-Medoid. By using the Python programming language and Google Colab tools.

1) K-Means

The stages of implementing the K-Means algorithm use Google Colab and Python Language. Then the stages of modeling and searching for clustering values are carried out. The

first stage is to start with the initialization of the centroid point. Where the number of clusters (K) starts as the first step. The second stage is to calculate the distance from each appropriate cluster. The third stage updates the centroid value based on the data that has been grouped in the iteration process is carried out by repeating the second and third stages.

When the convergent iteration results have been achieved, the value of K is obtained and a cluster calculation is carried out with the value of K=5. From the results of the clustering stage using the K-Means algorithm, it can be seen in table 4.1 This table provides the value of the cluster results formed and provides insight into the patterns or groups contained in the data after the K-Means modeling stage is completed. In the figure below is a graph of the results of determining the number of clusters:

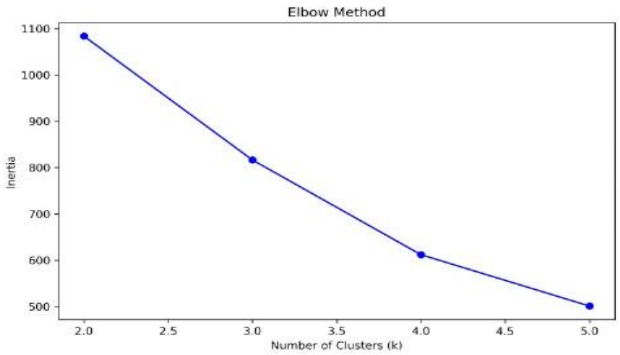


Figure 1. Elbow Method Chart

Source: Data processing using Python, 2024

The graph shown in Figure 1 is the result of the Elbow Method test to determine the optimal number of clusters in the K-Means algorithm. Here is the analysis:

1) X-axis (Number of Clusters, k): Indicates the number of clusters tested, in this case ranging from k=2 to k=5.
2) Y-axis (Inertia): Indicates the within-cluster sum of squares (WCSS) value, which is the total square distance between the data point and its cluster center. The smaller the inertial value, the better the cluster separation.
3) Declining Pattern: From the graph, it can be seen that the inertia value decreases as the number of clusters increases. However, after a certain point, the rate of decline slows down.

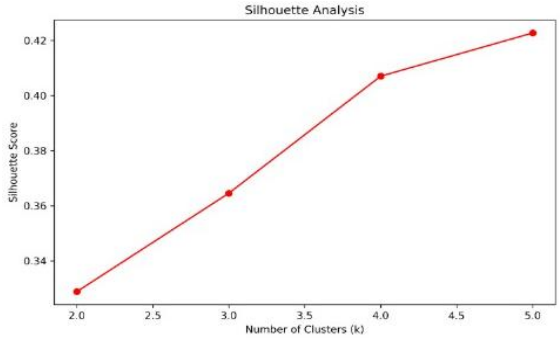


Figure 2. Silhoutte Analysis Results

Source: K-Means algorithm testing, 2024

The graph shown in Figure 2 is the result of the Silhouette Score analysis for various cluster numbers (k) in the K-Means algorithm. Here is the explanation:

1) X-axis (Number of Clusters, k): Indicates the number of clusters tested, in this case ranging from k=2 to k=5.
2) Y-axis (Silhouette Score): Indicates clustering quality, where a value close to 1 means the cluster is more compact and well separated, while a value close to 0 means overlapping clusters.
3) Improvement Patterns

a. The Silhouette Score value increases as the number of clusters increases.
b. The highest Silhouette Score was achieved at k=5 with a score of around 0.42.
c. There was no drastic decrease, which indicates that more and more clusters provide better separation.

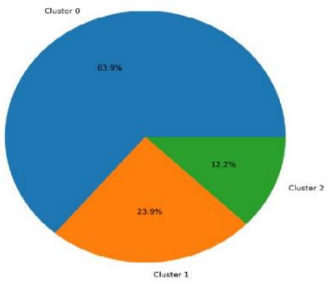


Figure 3. K-Means Cluster Distribution

Source: Crash data clustering results, 2024

The Pie diagram shown in Figure 3 shows the distribution of data in each cluster formed by the K-Means algorithm. Here is the explanation:

1) Cluster 0 (blue color) has 63.9% of the total data, making it the largest cluster.
2) Cluster 1 (orange) covers 23.9% of the total data, medium size compared to other clusters.
3) Cluster 2 (green) has 12.2% of the total data, making it the smallest cluster.

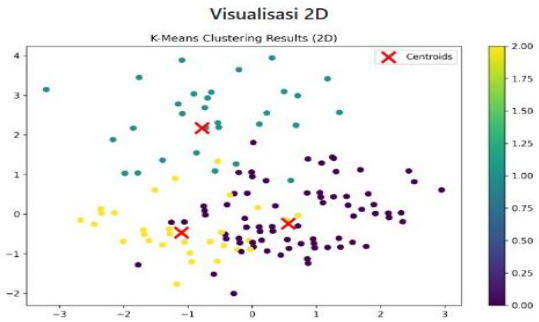


Figure 4. 2D Visualization of K-Means Clustering

Source: Clustering results visualization, 2024

The 2D visualization results shown in Figure 4 show the distribution of data based on clustering results with the K-Means algorithm. Here is an explanation of the elements in the plot:

1) Each point on the graph represents a sample of data.
2) The dot color indicates the cluster membership based on the clustering results. For example, the colors blue, yellow, and purple indicate the three clusters that the algorithm generated.
3) The red "X" in the plot is the centroid of each cluster.
4) This centroid is the central point of the cluster, calculated as the average position of all points in the cluster.
5) Centroid serves as a representation of each cluster.

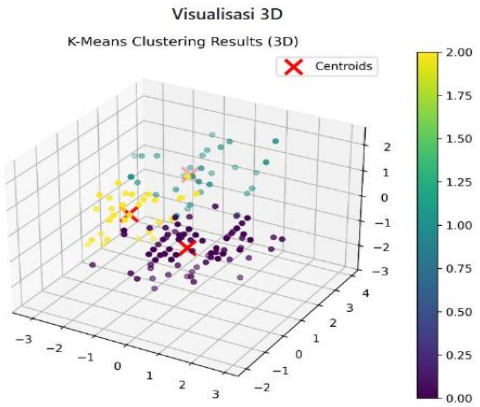


Figure 5. 3D Visualization of K-Means Clustering

Source: 3-dimensional spatial analysis, 2024

The 3D visualization results in figure 5 illustrate the distribution of data that has been grouped using the K-Means algorithm in three-dimensional space. Here is an explanation of the elements of this visualization:

1) Each point on the graph represents the sample data.
2) The dot color indicates the cluster generated by the K-Means algorithm. Different clusters are given different colors (e.g., light blue, yellow, purple, etc.).
3) The big red "X" in the graph is the centroid of each cluster.
4) The centroid is the midpoint of each cluster, calculated as the average position of all points in the cluster.
5) The centroid position indicates the data distribution center in the associated cluster.

Table 2. K-Means Clustering Result

K	*Inertia(SSE)*	*Sillhoutte Score*
2	1083.98	0.329
3	816.51	0.365
4	611.97	0.407
5	501.04	0.423

Source: Data processing with K-Means algorithm, 2024

Based on Table 2. the results of clustering testing with K-Means, here are the interpretations of the Inertia (SSE) and Silhouette Score metrics.

1) SSE:

a. SSE continues to decline as the number of clusters increases, which means that the more clusters, the smaller the average distance of the data point from its cluster center.
b. The largest decline occurred between K=2 to K=3 (1083.98 → 816.51) and K=3 to K=4 (816.51 → 611.97), suggesting that at this range, adding clusters helps improve separation.
c. After K=4 to K=5 (611.97 → 501.04), the decline in SSE is still there but not as fast as before, which could be a sign that the optimal point of the number of clusters is starting to be reached.

2) Silhouette Score :

a. K=2 → 0.329 → Clustering is quite good but there is still overlap between clusters.
b. K=3 → 0.365 → The increase indicates that adding a cluster helps to separate the data better.
c. K=4 → 0.407 → The increase is more significant, meaning that the number of clusters is starting to be more optimal.

d. K=5 → 0.423 → the highest Silhouette Score, indicating that the separation between clusters is getting better.

Table 3. K-Means Evaluation Matrix

Types of Testing	Value
Sillhoutte Score	0.365
Davies-Bouldin Index (DBI)	1.117
Purity Score	0.617

Source: Calculation of evaluation metrics, 2024

It can be seen Table 3. the results of the clustering stage after testing using Sillhoutte to measure how good the quality of the clustering carried out on the traffic accident data is carried out. So the average sillhoutte score of the entire cluster is 0.1437

K-Medoids

The implementation of the K-Medoids algorithm in testing the number of clusters has almost the same stages as the K-Means method. However, the main difference between the two lies in how the center of the cluster is determined.

In the K-Means method. The cluster center is calculated as the average of all the data in the cluster. In contrast, in K-Medoids, the cluster center (medoid) is selected from one of the actual data points, not the average value. A medoid is a data point that has the smallest total distance from all other points in the cluster.

Because it uses real points as the center of the cluster, K-Medoids are more resistant to outliers and extreme values compared to K-Means, which tend to be affected by data that has very large or very small values. The implementation process of K-Medoids begins with the selection of a number of initial medoids at random. Furthermore, each data will be associated with the nearest medoid based on a specific distance e.g. Euclidean distance. Then, the algorithm will try to replace the medoid with another point in the cluster and evaluate whether there is an improvement in the quality of the clustering based on a particular cost function. This process continues to be repeated until there is no significant change in medoid selection or until the convergence criteria are met.

With this approach, K-Medoids provide more stable clustering results and can be interpreted better. Especially in situations where the data contains extreme values or uneven distribution.

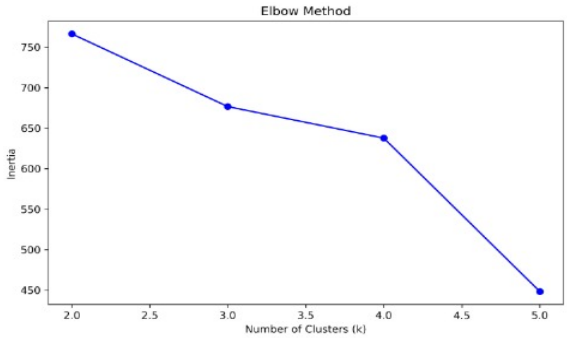


Figure 6. Elbow Method Chart

Source: Optimal analysis of the K-Medoids cluster, 2024

In Figure 6. The results of the test using the Elbow Method in this graph provide important information about the optimal number of clusters used in the K-Medoids algorithm. Here is the explanation:

1) Indicates the number of clusters k tested, in this case from 2 to 5 clusters.
2) k is a parameter that must be specified for the K-Medoids algorithm in order to divide the data into k groups.
3) Inertia measures the total distance or sum of squared distances between each data point to the nearest medoid in the cluster.
4) The smaller the inertial value, the better the data matches the medoid cluster.
5) From this graph, a bend is seen at k=3, indicating that 3 is the optimal number of clusters.
6) After k=3, although the inertia continues to decrease, the rate of decline becomes smaller, which suggests that increasing the number of clusters may not provide a significant increase.

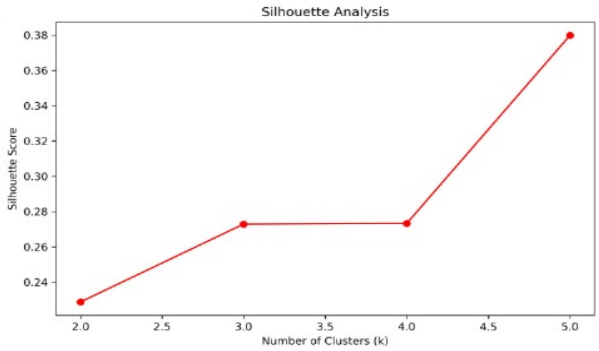


Figure 7. Silhoutte Analysis Results

Source: K-Medoids algorithm testing, 2024

Based on Figure 7 of the Silhouette Analysis chart, the Silhouette Score value increases as the number of clusters increases :

1. In the clustering test with the K-Medoids method, a Silhouette Score of 0.273 was obtained for k=3. This value indicates that the clustering results with three clusters are not optimal, because generally a higher Silhouette Score value indicates a cluster that is more well separated and more coherent.

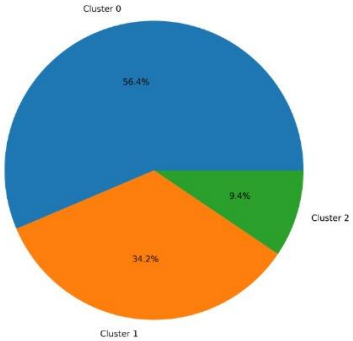


Figure 8. Distribution of K-Medoids Clusters

Source: K-Medoids grouping results, 2024

Based on the pie diagram in Figure 8 shown, the clustering results with the K-Medoids method show the distribution as follows:

1) Cluster 0 (Safe Zone) has the largest proportion (56.4%), indicating that most of the accident data is in the category with a lower level of risk. This could mean that the majority

of locations on the analyzed toll roads are relatively safe with fewer or low-severity accidents.

2) Cluster 1 (Alert Zone) covers 34.2% of the data, indicating that almost one-third of the overall accident data occurred in areas with moderate risk factors. This could indicate areas that need more attention, such as improving traffic signs, increased surveillance, or other interventions to prevent risk escalation.
3) Cluster 2 (Hazard Zone) has the smallest proportion (9.4%), which means only a small fraction of locations are actually categorized as hazardous. However, even though the percentage is small, these areas require high priority in accident mitigation efforts, such as improving safety infrastructure, adding street lights, or speed limiters.

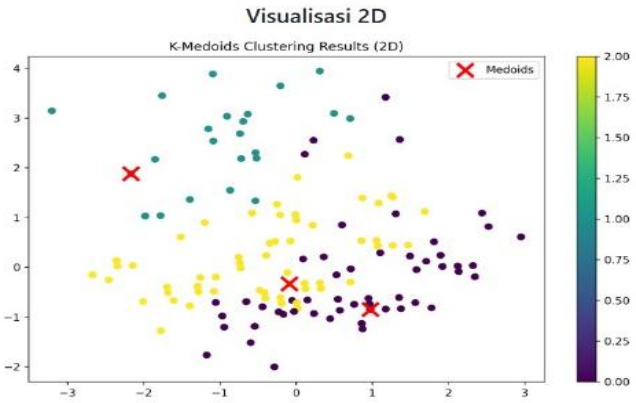


Figure 9. 2D Visualization of K-Medoids Clustering

Source: Cluster graphical representation, 2024

Figure 9 shows the clustering results using the K-Medoids method in a 2D representation, with each point representing the accident data that has been grouped into three different clusters:

1) The separation between clusters is quite visible, but some points are still close to the boundaries between clusters, which could be an indication that some data is difficult to categorize clearly.
2) When compared to K-Means, the K-Medoids method is more robust against outliers, because it chooses medoids as the center of the cluster rather than using the mean (average).
3) The previous relatively low Silhouette Score (0.273) indicates that some points may not be very suitable for a given cluster, which is also evident from the not very compact distribution of data in this visualization.

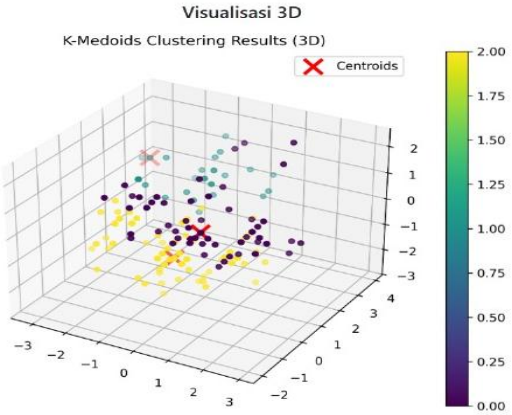


Figure 10. 3D Visualization of K-Medoids Clustering

Source: 3D mapping of clustering results, 2024

In Figure 10 The 3D visualization shown in the image is the result of the application of the K-Medoids algorithm to cluster on three-dimensional data. The following is an analysis of the visualization results:

1) Centroid: Marked with a red "X" sign, which indicates the center of each cluster that has been found by the algorithm.
2) The center of the cluster (medoids) marked with a large red X indicates the position of the representatives of each group.
3) K-Medoids work by selecting real points from the dataset as centroids, in contrast to K-Means which uses averages.
4) The selection of medoids as centroids helps in handling outliers better as compared to K-Means.

Table 4. Result of Clustering K-Medoids

K	*Inertia (SSE)*	*Sillhoutte*
2	766.52	0.229
3	676.70	0.273
4	637.73	0.273
5	448.28	0.380

Source: K-Medoids algorithm analysis, 2024

Based on Table 4 of the results of clustering testing using K-Medoids with the number of clusters k = 2 to k = 5, the following is the analysis:

1) K=5 gave the best results because Inertia experienced the largest decrease (637.73 → 448.28). The Silhouette Score increased significantly to 0.380 (the highest in this range). This means that clustering is better than the number of other clusters.
2) K=3 or K=4 was less than optimal because the Silhouette Score was stagnant at 0.273, indicating that adding clusters from 3 to 4 did not provide an improvement in data separation.

Table 5. Metrik Evaluasi K-Medoids

Types of Testing	**Value**
Sillhoutte Score	0.273
Davies-Bouldin Index (DBI)	1.440
Purity Score	0.436

Source: Cluster performance comparison, 2024

5. Evaluation

The evaluation of traffic accident data in this study aims to assess the effectiveness of the clustering method in grouping accident locations based on the level of accident risk. Using the K-Means and K-Medoids algorithms, the segmentation of accident zones has been carried out to divide the area of the Bakauheni – Terbanggi Besar toll road. In this study, the clustering method of K-Means and K-Medoids has been applied to the segmentation of traffic accident profiles on the Bakauheni – Terbanggi Besar Toll Road. Based on the results of clustering, three main segments were obtained:

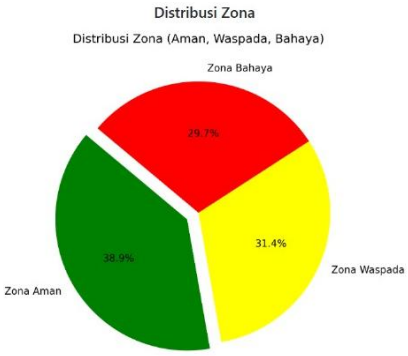


Figure 11. K-Means Accident Data Visualization

Source: Risk zone mapping, 2024

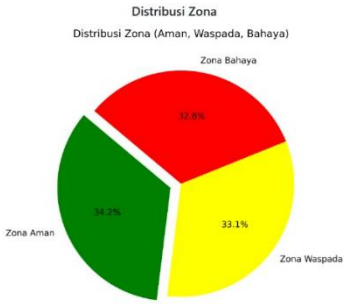


Figure 12. K-Medoids Accident Data Visualization

Source: Spatial analysis of accidents, 2024

1) Safe Zone: This zone includes the road segments that have the lowest accident rate compared to other zones. Characteristics in this zone are likely to include factors such as good road infrastructure, adequate street lighting, as well as relatively low traffic volumes or more cautious driver behavior. These zones can be a low priority in traffic safety interventions.
2) Danger Zone: This zone is at a moderate risk level, with an accident percentage that is almost equal to the Safe Zone. These zones can reflect locations that are starting to show potential danger, such as curves, long straights (prone to overspeed), or segments with increased vehicle volume. Preventive measures, such as the installation of warning signs or increased surveillance, can be prioritized in these zones to prevent increased risk.
3) Alert Zone: This zone covers the segments with the highest accident rate on the Bakauheni–Terbanggi Besar Toll Road. The characteristics of these zones may involve factors such as high traffic density, lack of street lighting, or poor road conditions (such as the presence of damage). These zones require major attention from the authorities, including the implementation of safety measures such as the installation of surveillance cameras, infrastructure repairs, or the implementation of speed restrictions.

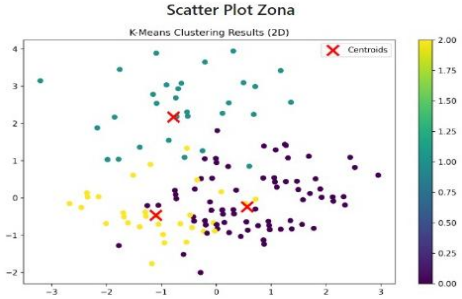


Figure 13. Scatter Plot K-Means Accident Data Zone

Source: Accident data distribution plot, 2024

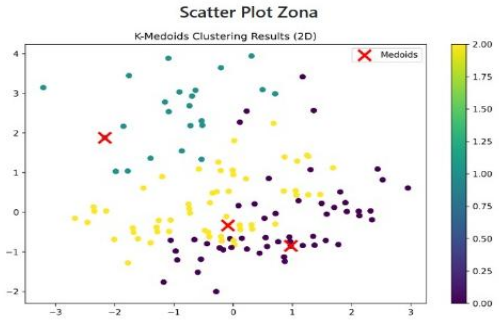


Figure 14. Scatter Plot K-medoids Accident Data Zone

Source: Accident pattern visualization, 2024

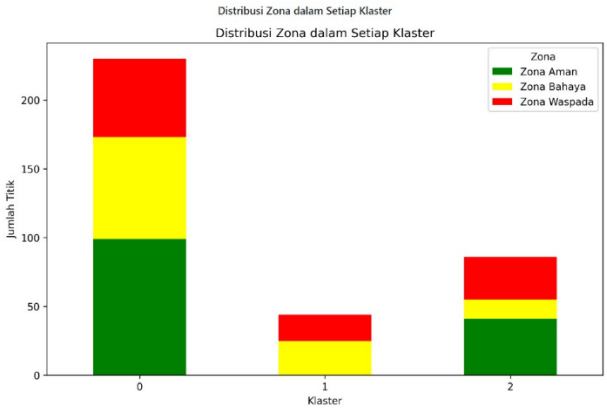


Figure 15. Zone Divide Chart by K-Means

Source: Delineation of areas by cluster, 2024

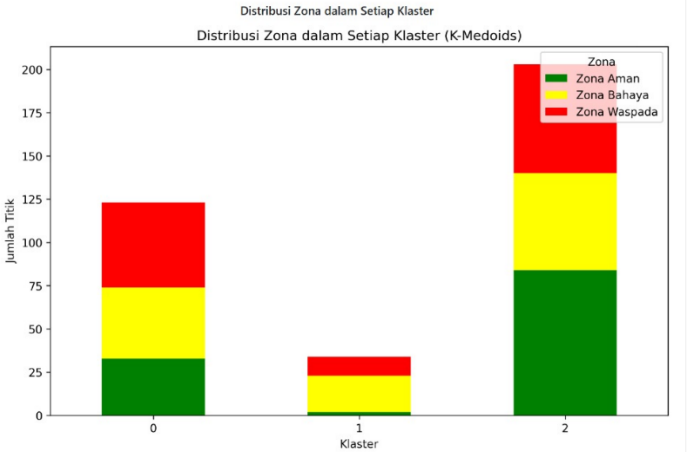


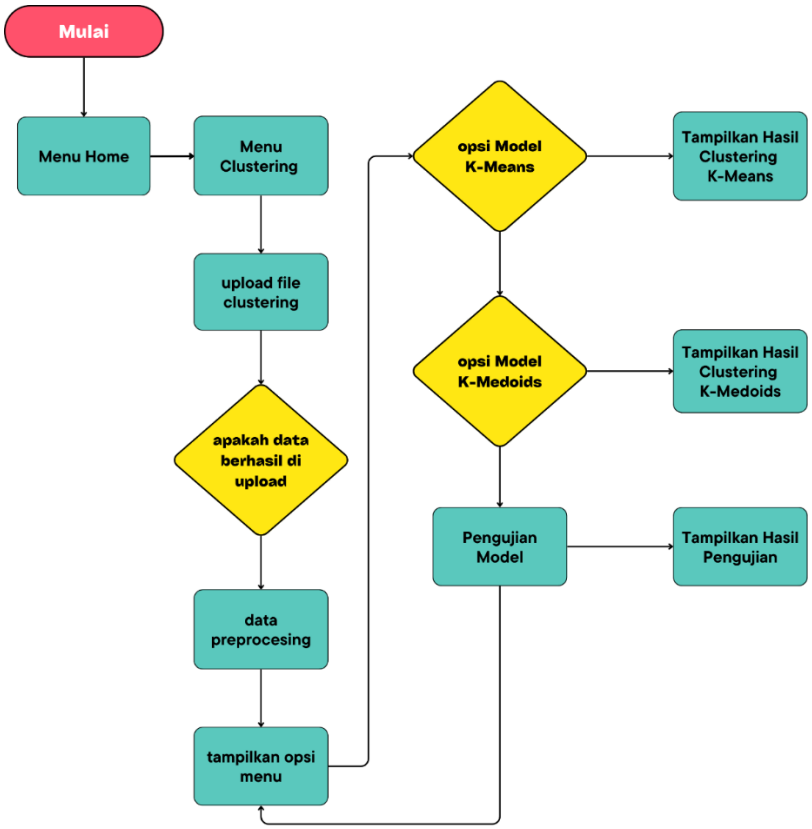
Figure 16. Zone Divide Chart by K-Medoids

Source: Risk region segmentation, 2024

Overall, the distribution of data shows that the three zones have almost equal percentages, indicating the need for further analysis to understand the causative factors of accidents in each zone and direct resources proportionately to reduce the risk of accidents on the expressway.

6. Deployment

At the deployment stage to make it easier for users, modeling is made with a Web-based application prototype using Python language. The following is attached in figure 4. Illustrated with an activity diagram as the basis for making an application:



Gambar 17. Activity Diagram Clustering
Source: Research workflow, 2024

In the Activity Diagram that has been created, a brief description is carried out so that it is easy to understand in each existing sub-menu:

1) Home Menu: This is the main page of the app that serves as the user's gateway to the system.
2) Clustering Menu: Provides the main feature to carry out the clustering process on traffic accident data.
3) Data Preprocessing: The stages of data preparation before the clustering process include normalization, standardization, handling of missing values, category encoding, and dimension reduction if needed.
4) K-Means: A feature that allows users to run the K-Means algorithm to segment traffic accident data The K-Means algorithm works by dividing the data into k clusters based on the Euclidean distance from the centroid calculated iteratively.
5) K-Medoids: Features for clustering using the K-Medoids method, which is similar to K-Means but more robust against outliers because it uses medoids as the center of the cluster.
6) Model Testing: Provides a variety of evaluation metrics to measure the quality of clustering results. The metrics used in this study include:

a. Elbow Method: To determine the optimal number of clusters based on changes in inertial values.

b. Silhouette Score: To evaluate how well the data points are in their respective clusters.
c. Davies-Bouldin Index (DBI): To assess the extent to which the clusters formed can be separated from each other.
d. Purity: To measure the extent to which the clustering results match the actual label if available.

The use of Python programming language and flask framework is a tool used in creating Web prototypes so that it can be done quickly and efficiently without requiring in-depth knowledge of web development.

CONCLUSION

The study comparing K-Means and K-Medoids clustering methods on traffic accident data from the Bakauheni–Terbanggi Besar Toll Road found that while K-Means produced faster clustering results, it was less stable due to sensitivity to centroid initialization. In contrast, K-Medoids demonstrated greater robustness to outliers and yielded more stable, higher-quality clusters, as supported by lower Davies-Bouldin Index values, slightly better Purity Scores, and clearer separation indicated by Silhouette Scores. Both methods effectively grouped accident risk into Safe, Alert, and Danger Zones, helping to identify high-risk segments influenced by road conditions, traffic density, and driver behavior. For future research, it is suggested to explore hybrid clustering approaches or integrate temporal and spatial factors to enhance the accuracy and practical application of accident risk segmentation for improved road safety management.

REFERENCES

Astuti, W., Pradono, M. H., & Nugroho, P. (2022). The impact of toll road development on regional economic growth: Case of Lampung Province. *Journal of Infrastructure & Regional Development, 5*(2), 89–101. https://doi.org/10.1016/j.jird.2022.05.003

Bakauheni–Terbanggi Besar Toll Road. (2025, July). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Bakauheni–Terbanggi_Besar_Toll_Road

Chaniago, N. A. D., Pratama, G. A., Latief, Y., Riantini, L. S., & Trigunarsyah, B. (2024). Identification of post-construction toll road management barriers in Indonesia. *E3S Web of Conferences*.

Chaniago, N. A. D., et al. (2024). Identifikasi hambatan manajemen tol pasca konstruksi di Indonesia. *E3S Web of Conferences*.

Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2022). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. *Jurnal Teknologi Dan Informasi, 12*(1), 64–77. https://doi.org/10.34010/jati.v12i1.6674

Hasanah, N., Permana, Y., & Arifin, S. (2022). Transport infrastructure and agricultural commodity distribution in Sumatera. *International Journal of Supply Chain Management, 11*(1), 14–25. https://scholar.google.com/scholar_lookup?title=Transport%20infrastructure%20and%20agricultural%20commodity%20distribution

Hermawan, I. (2024). Road traffic facilities, traffic accidents, and poverty in Indonesia. *Transportation Research Procedia, 280*, 1273–1280.

Hu, Y. (2024). The Trans-Sumatra Toll Road and Economic Geography in Indonesia. *FREIT Working Papers*.

INA platform report. (2024). *INA Toll Road Platform invests in Trans Sumatra Toll Roads*. APG Asset Management Asia-Pacific.

Kementerian Pekerjaan Umum dan Perumahan Rakyat (KemenPUPR). (2022). *Dampak Jalan Tol terhadap Pemerataan Pembangunan Wilayah Indonesia*. Jakarta: Pusat Penelitian dan Pengembangan Jalan dan Jembatan. https://pu.go.id

Mulyono, A. T., Sudrajat, A., & Harianto, I. (2021). Evaluating logistics efficiency through toll road expansion in Indonesia. *Transportation Research Procedia, 52*, 245–252. https://doi.org/10.1016/j.trpro.2021.02.031

Mumtazah, B. B., & Sancoko, S. D. (2024). Adult Clothing Size Recomendation Using K-Nearest Neighbor and Support Vector Machine Algorithm Rekomendasi Ukuran Baju Dewasa Menggunakan Algoritma K-Nearest Neighbor dan Support Vector Machine, *4*(October), 1635–1645.

Prawira, D., & Nugroho, H. (2020). Connectivity improvement and port access through Bakauheni-Terbanggi Toll Road. *Maritime Economics & Logistics, 22*(4), 403–419. https://doi.org/10.1057/s41278-019-00134-x

Purwaningsih, E. (2019). Analisis Kecelakaan Berlalu Lintas Di Kota Jakarta Dengan Menggunakan Metode K-Means. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer, 5*(1), 139–144. https://doi.org/10.33480/jitk.v5i1.712

Putri, A. R., & Sari, M. E. (2021). The role of toll roads in boosting SME growth in South Sumatra. *Journal of Development Policy Review, 3*(3), 199–212. https://doi.org/10.21043/jdpr.v3i3.10234

Rahmadana, Muhammad Fitri, & Putra, Ilham Mirzaya. (2025). Community dynamics towards the existence of toll roads in Indonesia: a literature and spatial study. *Frontiers in Built Environment*, *11*, 1515186.

Reuters. (2024, October 28). Indonesian sovereign wealth fund INA's platform invests in Trans Sumatra Toll Roads. *Reuters*.

Rosiana, P. S., Mohsa, A. A., Fadila, M. A., & Jaman, J. H. (2023). Visualisasi Data Tindak Kejahatan Berdasarkan Jenis Kriminalitas Di Kabupaten Karawang Dengan Menggunakan Algoritma Clustering K-Means. *Jurnal Informatika Dan Teknik Elektro Terapan, 11*(3s1). https://doi.org/10.23960/jitet.v11i3s1.3347

Wahyuni, E., Hidayat, R., & Kartika, D. (2023). Toll road infrastructure and rural economic transformation: Evidence from Lampung. *Regional Studies, Regional Science, 10*(1), 221–234. https://doi.org/10.1080/21681376.2023.2198874