Classification of Traffic Accident Levels in West Java Using the K-Means Algorithm

Gusti Musyaffa Razan[1], Baenil Huda[2], Shofa Shofiah Hilabi[3]
Information Systems Study Program, Faculty of Computer Science, Universitas Buana Perjuangan Karawang

Traffic accidents are an important factor in improving public safety in West Java Province, because the population movement rate there is very high. The high number of accidents is directly related to increased deaths and material losses, but the use of historical data is still limited to administrative archiving tasks without any process of identifying regional vulnerability patterns. This study aims to classify accident-prone areas using Data Mining techniques and the K-Means Clustering algorithm, as well as applying the CRISP-DM framework approach. The analyzed dataset comes from the Indonesian National Police Traffic Accident Database (Pusiknas Polri) for the period 2020 to 2025, consisting of 552 observations with indicator variables covering the number of fatalities, serious injuries, and minor injuries. The determination of the most appropriate number of clusters was tested using the Silhouette method to ensure more accurate and objective modeling results. The analysis shows that the number of clusters (k=3) is the most appropriate, with a Silhouette metric value of 0.398. The application of the model produces three levels of risk: the red zone, which indicates high risk with 51 cases and the highest mortality rate; the yellow zone, which indicates moderate risk with 189 cases; and the green zone, which indicates low risk with 312 cases. The visualization of these mapping results is expected to be an important tool for the police and local governments in formulating mitigation policies, improving patrol efficiency, and accelerating infrastructure improvements in high-risk areas, thereby reducing the number of accidents in the future.

Keywords: Data Mining, K-Means Clustering, Traffic Accidents, CRISP-D, Risk Mapping.

Corresponding Author:
Gusti Musyaffa Razan
Information Systems Study Program, Faculty of Computer Science, Universitas Buana Perjuangan Karawang
si22.gusti@mhs.ubpkarawang.ac.id

1. Introduction

Transportation is a vital instrument in facilitating human mobility and promoting sustainable regional economic growth. Jawa Barat, one of the most densely populated provinces and a major national industrial hub in Indonesia, experiences extremely high vehicle movement across its primary road networks. However, the rapid growth in the number of vehicles has not been matched by proportional improvements in road capacity and quality. This imbalance between vehicle volume and road capacity has generated numerous roadway problems, the most serious of which is the risk of fatal traffic accidents. Traffic accidents are unforeseen events involving motor vehicles that can result in substantial losses, including tragic loss of human life. Moreover, the diverse geographical conditions of West Java ranging from steep mountainous areas to heavily trafficked coastal routes pose additional challenges to ensuring land transportation safety in the region [1].

Traffic accident rates in West Java continue to rise, requiring serious attention from authorities due to their destructive impact on social structures and economic stability. According to data collected by the Polisi Daerah Jawa Barat in 2024, there were 7,442 traffic accident cases across various regencies and municipalities. Although this represents a 21% decrease compared to the previous year, the fatality rate remains alarmingly high. The data indicate 2,778 fatalities, 627 severe injuries, and 8,313 minor injuries resulting from these accidents. This statistical phenomenon suggests that a reduction in the number of

incidents does not necessarily correspond to a reduction in danger or casualties on the ground. Consequently, traffic accidents in West Java remain a major event with severe consequences for regional productivity [2].

A major challenge faced by stakeholders is the suboptimal utilization of accident data for strategic purposes. Accident data routinely collected by the police or transportation agencies are generally used only for administrative recapitulation or institutional reporting. These large datasets have not been thoroughly processed, analyzed, or mined to uncover hidden insights, such as spatial distribution patterns by region type, objective severity classifications, or identification of high-risk areas. Without a clear, organized, and data-driven risk map, responsible agencies often struggle to determine priorities for risk reduction. Strategic decisions such as installing early warning signs, repairing damaged roads, or placing traffic monitoring posts may therefore be less effective because they are not grounded in comprehensive risk mapping analysis [3]

To bridge the gap between raw data availability and strategic information needs, the application of modern information technology through data mining is required. Data mining refers to a set of computational processes used to discover patterns, relationships, and anomalies within large datasets using machine learning algorithms. One well-established unsupervised learning method for numerical data clustering is the K-Means Clustering algorithm. This algorithm partitions datasets into groups (clusters) based on similarity by calculating Euclidean distances. K-Means is recommended in this study due to its computational efficiency, ability to handle high-dimensional data, and effectiveness in revealing patterns of accident severity that may not be detectable through conventional statistical methods [4].

Previous studies have demonstrated the importance of data mining in transportation safety. Research by Rahma and Latifah (2023) applied K-Means to identify high-risk accident areas in Samarinda, successfully determining the most vulnerable times and locations for preventive planning. Similarly, Nugroho (2022) used K-Means Clustering to analyze traffic accidents in Jawa Tengah, producing three primary clusters reflecting accident severity levels [5].

Despite numerous prior studies, this research offers novelty by expanding the geographical scope and enhancing methodological rigor. Unlike earlier studies focused on a single city or district, this study utilizes province-level secondary data for West Java obtained comprehensively from the National Criminal Information Center (Pusiknas) of the Indonesian National Police. Additionally, the research strictly applies the CRISP-DM framework, ensuring each stage is systematically validated. The optimal number of clusters is determined not arbitrarily but through the Elbow Method, reinforced by the Silhouette Coefficient to minimize mathematical errors and improve clustering accuracy [6].

The primary objective of this study is to apply the K-Means Clustering algorithm to classify jurisdictions in West Java based on accident severity variables: number of fatalities, severe injuries, and minor injuries. The clustering is designed to produce three risk categories high risk (red), medium risk (yellow), and low risk (green). The main contribution of this research is the development of a high-validity visual map of accident-prone areas based on empirical data. This cluster map is expected to serve as a decision-support tool for local governments, transportation agencies, and law enforcement authorities in Indonesia. It can assist in formulating road safety policies, infrastructure management strategies, and targeted accident prevention plans that are clearer, more responsive, and more effective in reducing traffic fatalities [7].

2. Literature Review

Traffic accidents represent a significant transportation problem because they lead to loss of life, injuries, and economic losses. High traffic density and inadequate road infrastructure capacity often increase

accident risks in rapidly developing regions. The use of historical accident data is therefore important to understand patterns of accidents and support evidence-based transportation safety policies [1].

Recent studies have widely applied data mining techniques to analyze large transportation datasets and identify hidden patterns. One commonly used method is the K-Means clustering algorithm, which g roups data based on similarity using distance calculations. This method is considered effective for identifying accident-prone areas and classifying accident severity levels objectively [5]. Several studies have successfully applied this approach in traffic accident analysis.

Although previous studies confirm the usefulness of clustering methods, several limitations remain. Many studies focus only on limited geographic areas and emphasize accident frequency rather than incorporating multiple accident severity indicators such as fatalities, serious injuries, and minor injuries. These limitations indicate the need for a more comprehensive analysis that integrates multiple accident variables and broader datasets to improve accident risk classification.

Based on the literature review, it can be observed that research on traffic accident analysis using clustering techniques still faces several limitations, particularly in terms of spatial coverage and the variables used to represent accident severity. Previous studies generally focus on specific regions and often rely only on accident frequency data, which may not fully represent the actual severity of accidents.
Therefore, this study aims to analyze traffic accident data by incorporating multiple severity indicators and applying the K-Means clustering algorithm to classify accident risk levels. The research problem can be formulated as follows:

1. How can the K-Means clustering algorithm classify traffic accident risk levels in West Java based on accident severity variables?
2. What patterns of accident risk distribution can be identified from the clustering results?

Based on this formulation, the research hypothesis is:
H1: The K-Means clustering algorithm can effectively classify traffic accident risk levels into several clusters based on accident severity indicators.

3. Methods

This research stage adopts the CRISP-DM phases adapted to the writing structure presented below.

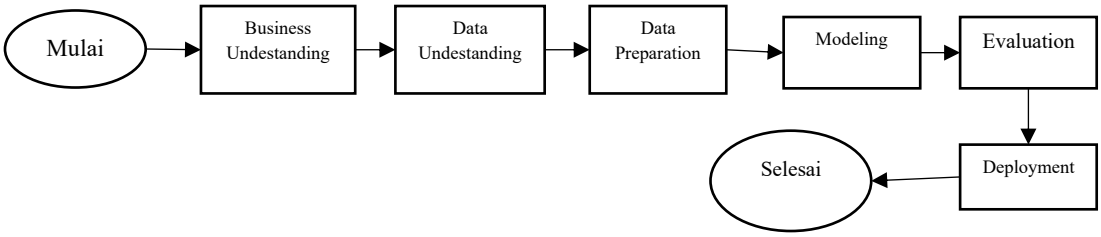

Figure 1. Research Phase Flow

Research Framework

This study is an applied quantitative research employing a computational approach through data mining methods. The fundamental problem-solving approach strictly follows an internationally recognized standard methodology, namely the Cross-Industry Standard Process for Data Mining (CRISP-DM). This approach was selected because it is systematic, well-structured, and encompasses six comprehensive stages: business understanding, data understanding, data preparation, modeling, evaluation, and

deployment into real-world practice. The object examined in this study is the mapping of regional risk levels in West Java, based on historical traffic accident data [8].

Analytical testing was conducted using secondary data obtained from the National Police Information Center (Pusiknas Polri) through the IRSMS Korlantas system, supported by reports from the Statistics Agency (BPS) of West Java. The analyzed accident period spans from early 2020 to mid-2025. Research variables used as indicators of accident severity include the number of fatalities (MD), serious injuries (LB), minor injuries (LR), and material losses. The fundamental hypothesis of this study is that the K-Means Clustering algorithm can accurately classify regions in West Java into three risk clusters: red, yellow, and green zones [9]

Research Stages

Systematically, the technical software engineering stages implemented to produce a valid clustering model are described through the following essential sub-processes:

Data Collection and Pre-Processing (Data Preparation)

Raw accident data collected in spreadsheet format (CSV/Excel) cannot be directly processed by algorithms due to potential noise and anomalous values. Therefore, data pre-processing or cleaning is mandatory. The first step involves detecting and removing duplicate records to prevent computational bias. Missing or incomplete data are then handled through logical imputation or row deletion if the proportion of corrupted data is too large.

After ensuring data cleanliness, feature transformation through normalization is conducted. This is necessary because the K-Means algorithm is highly sensitive to differences in variable scale (e.g., population counts versus material losses) (Maulana et al., 2024). Z-score standardization transforms data to have a mean of zero and a standard deviation of one, mathematically expressed as:

$$z = \frac{x-\mu}{\sigma} \quad [1]$$

where z is the standardized value, x is the original value, μ is the variable mean, and σ is the standard deviation. This standardization is crucial to ensure that each indicator variable has equal weight in distance calculations, thereby preventing domination by variables with large numerical ranges [10].

Determination of the Optimal Number of Clusters

Before modeling implementation, the optimal number of clusters (k) is determined using a multi-criteria approach to minimize researcher subjectivity. The initial method applied is the Elbow Method, which analyzes the Within-Cluster Sum of Squares (WCSS) across different values of k. The inflection point ("elbow") on the WCSS curve indicates the optimal balance between model complexity and clustering accuracy.

To strengthen validity, the Silhouette Coefficient metric is also applied. This metric evaluates clustering quality by measuring internal cohesion and inter-cluster separation on a scale from −1 to 1. Mathematically, the silhouette score is defined as:

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

Where a (i) represents the average intra-cluster distance (cohesion) and *b(i)* represents the average nearest-cluster distance (separation).

Implementation of the K-Means Clustering Algorithm

The final phase is deployment, in which clustering results are visualized into a traffic accident risk zoning map to provide strategic policy recommendations for police authorities and local governments. During core modeling, initial cluster centroids are randomly assigned according to the predetermined *K* value. The algorithm then assigns each accident data record in a region to the nearest centroid. Similarity and proximity are measured using the Euclidean distance formula, a type of spatial distance, expressed as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \quad (x_i - y_i)^2} \qquad [1]$$

where *d(x, y)* represents the absolute distance between observed data and the cluster center, x_i denotes the characteristics of the *i*-th region, and y_i denotes the coordinates of the *i*-th cluster centroid. Centroid positions are iteratively updated by calculating the mean of all cluster members until convergence is achieved—when no further data reassignment occurs. Computation was performed using Orange Data Mining software (Ratna Sari et al., 2025; [10].

Model Evaluation (Silhouette Coefficient)

The final step in the CRISP-DM cycle is model validation to demonstrate that the classification of traffic accident risk zones has been established logically and systematically. Independent evaluation is conducted using the Silhouette Coefficient metric, which simultaneously measures cohesion (how similar a data point is to its own cluster) and separation (how distant it is from neighboring clusters). Scores range from −1 to 1; values closer to 1 indicate higher clustering quality. This means the algorithm can identify accident-prone areas clearly, compactly, and without overlap (Ratna Sari et al., 2025).

4. Result and Discussion

Analysis for Determining the Optimal Number of Clusters

The implementation of the modeling phase within the CRISP-DM framework began with computational processing of the traffic accident dataset of West Java Province for the 2020–2025 period, which had previously undergone Z-score standardization. The initial and crucial step was determining the number of clusters (k) that most accurately represents the vulnerability levels across jurisdictions. This study employed a dual-validation approach namely the Elbow Method and the Silhouette Coefficient to ensure that group formation was grounded in objective mathematical calculations rather than researcher subjectivity [11].

Within the CRISP-DM modeling framework, determining the most representative number of clusters (k) for grouping accident risk levels in West Java Province constituted the starting point. This determination utilized a multi-criteria approach combining the Elbow Method and the Silhouette Coefficient to minimize subjectivity in data grouping [12]. Based on tests conducted using the Orange Data Mining software, the researcher iterated k values from 2 to 10 on a dataset that had been standardized using Z-score scaling. Observation of the Elbow Method graph revealed a sharp decline in the Within-Cluster Sum of Squares (WCSS) when moving from one to two clusters, followed by a gradual flattening at k = 3.

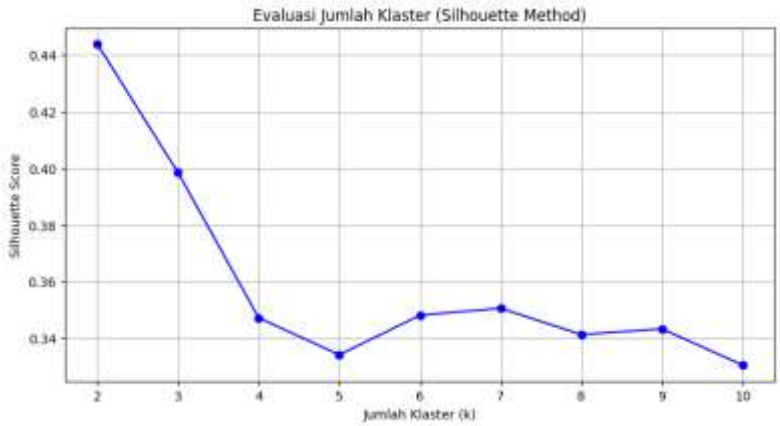

Figure 2. Silhouette Method

Validation was further strengthened using the Silhouette Coefficient, which measures the cohesion within clusters and the separation between different clusters. The results indicated that the highest silhouette score mathematically occurred at k = 2 with a value of 0.443, while k = 3 produced a score of 0.398. Although k = 2 yielded a slightly higher score, the researchers selected k = 3 as the optimal number of clusters because the difference was not substantial and k = 3 better aligned with the study's objective of segmenting risk into three zones (Red, Yellow, and Green). Establishing k = 3 provides more detailed and actionable information for the Regional Police of West Java Province in determining mitigation priorities compared to dividing the data into only two broad groups [13].

Cluster Characteristic Analysis (Centroid Analysis)

After setting k = 3, the K-Means algorithm iteratively computed Euclidean distances until convergence was achieved. The final clustering results partitioned accident data into three significantly different risk profiles based on the variables of fatalities (MD), serious injuries (LB), and minor injuries (LR).

The first cluster was identified as the Red Zone (High Risk), characterized by centroid values for fatalities (MD) in the highest positive range between 2 to 4 standard deviations above the provincial mean. This zone reflects areas with extreme fatality levels, where each accident incident carries a substantially higher probability of death compared to other regions.

The second cluster was categorized as the Yellow Zone (Moderate Risk), with data points concentrated around the average to mid-range values [14]. Areas in this category experience relatively frequent accidents dominated by minor injuries, but with a lower fatality ratio than the Red Zone.

The third cluster was defined as the Green Zone (Low Risk), where most data points are concentrated in regions with negative coordinate values, meaning below the provincial statistical average. These negative values resulting from Z-score scaling indicate jurisdictions with more stable and safer transportation performance compared to other areas in West Java Province.

Table 1. Cluster Information

Cluster	Zone
C1	Red
C2	Yellow
C3	Green

Interpretation of Graph Dimensions and Variable Weights

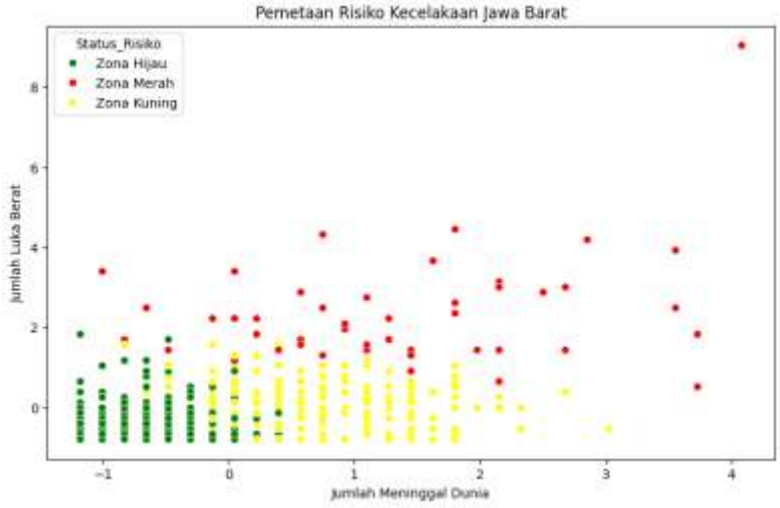

Figure 3. Scatter Plot

Although the K-Means algorithm in this study simultaneously processes three primary variables namely the number of fatalities (MD), severe injuries (LB), and minor injuries (LR)—the visualization in Figure 5 presents only two principal dimensions. This limitation arises from the constraints of two-dimensional visual media, which require the researcher to select variables with the highest variance and fatality impact to clearly represent the cluster distribution. The selection of MD and LB as the coordinate axes is based on their significance as the principal indicators of crash severity.

Although the Minor Injury (LR) variable is not displayed as a physical axis, its influence remains mathematically integrated into the cluster color determination. Each data point represented in red, yellow, or green inherently incorporates the weighted values of all three variables. The Red Zone is concentrated in the upper-right quadrant because these regions exhibit fatality and severe injury figures that substantially exceed the provincial average (positive values on the standardized scale). Conversely, the Green Zone clusters in the lower-left quadrant, reflecting fatality indicators (MD and LB) that fall below the average or yield negative standardized values.

Distribution of Jurisdictional Areas by Cluster

The frequency distribution analysis of 552 accident records in West Java reveals an uneven pattern, indicating disparities in transportation safety levels across jurisdictional areas. Based on data processing using the K-Means algorithm, the classification of jurisdictional areas by cluster is described as follows:

Table 2. Sample Clustered Data by Regency, January 2025

Location	Date of Incident	Cluster	Silhouette	Number of Fatalities	Number of Serious Injuries	Number of Minor Injuries	New Cluster	Risk Status
Banjar	January 2025	C3	0.685435	-0,828	-0,803	-1,122	1	Green Zone
Bogor	January 2025	C1	0.614233	1,627	3,66	3,336	2	Red Zone
Ciamis	January 2025	C3	0.640957	-0,828	-0,541	0,146	1	Green Zone
Cianjur	January 2025	C2	0.588931	1,101	1,166	0,261	0	Yellow Zone

Location	Date of Incident	Cluster	Silhouette	Number of Fatalities	Number of Serious Injuries	Number of Minor Injuries	New Cluster	Risk Status
Cimahi	January 2025	C3	0.582276	-0,653	-0,803	0,607	1	Green Zone
Cirebon City	January 2025	C3	0.571228	-0,127	-0,672	0,223	1	Green Zone
Garut	January 2025	C3	0.552197	0,049	0,378	-0,392	1	Green Zone
Indramayu	January 2025	C2	0.588231	0,926	0,378	2,029	0	Yellow Zone
Karawang	January 2025	C2	0.639045	1,101	-0,278	1,529	0	Yellow Zone
Bandung City	January 2025	C2	0.613461	1,101	-0,41	0,223	0	Yellow Zone
Greater Bandung	January 2025	C2	0.550411	-0,127	0,378	0,492	0	Yellow Zone
Bogor City	January 2025	C3	0.669905	-0,828	0,116	-0,969	1	Green Zone
Cirebon City (Municipality)	January 2025	C2	0.592567	0,049	-0,147	0,991	0	Yellow Zone
Kuningan	January 2025	C3	0.675051	-0,828	-0,278	-0,431	1	Green Zone
Majalengka	January 2025	C3	0.613103	0,049	-0,41	-0,431	1	Green Zone
Pangandaran	January 2025	C3	0.684952	-1,004	-0,278	-0,776	1	Green Zone
Purwakarta	January 2025	C2	0.603003	0,049	0,509	0,838	0	Yellow Zone
Subang	January 2025	C2	0.573361	0,049	1,166	1,145	0	Yellow Zone
Sukabumi	January 2025	C3	0.682982	-0,653	-0,541	-1,238	1	Green Zone
Sukabumi City	January 2025	C3	0.674703	-0,478	-0,803	-1,238	1	Green Zone
Sumedang	January 2025	C2	0.573613	1,452	-0,803	-0,277	0	Yellow Zone
Tasikmalaya	January 2025	C3	0.682674	-0,653	-0,672	-0,623	1	Green Zone
Tasikmalaya City	January 2025	C3	0.682334	-0,828	-0,803	-0,584	1	Green Zone

Analysis of the Red Zone (High Risk)

This cluster represents the most critical group, with a total of 56 accident incidents exhibiting extreme fatality levels. The jurisdiction of the Bogor Regency Police consistently dominates this zone, with the standardized mortality (MD) variable ranging from 1.627 to 2.153. The high positive values on the X-axis

coordinates indicate that the number of fatalities in this area is significantly above the provincial average of West Java. This condition is driven by Bogor's geographical characteristics as a buffer region for the capital, marked by very high vehicle density and the presence of primary arterial roads frequently traversed by heavy-tonnage vehicles [15].

Analysis of the Yellow Zone (Medium Risk)

The Yellow Zone comprises 138 accident incidents distributed across industrial areas and inter-provincial corridors. Jurisdictions such as the police regions of Karawang Regency, Indramayu Regency, Cianjur Regency, and Bandung alternately occupy this category. For example, in January 2025, Karawang recorded a standardized mortality score of 1.101 and a minor injury score of 1.529. The characteristics of this zone indicate that although mortality rates are not as extreme as in the Red Zone, the frequency of accidents resulting in injuries remains very high, reflecting significant traffic vulnerability along major economic corridors.

Analysis of the Green Zone (Low Risk)

The Green Zone is the largest cluster, encompassing 358 accident incidents. Jurisdictions included in this category comprise the police regions of Banjar, Ciamis Regency, Cimahi, Kuningan Regency, and Cirebon. Most areas in this zone exhibit negative coordinates on the fatality axis, such as Banjar, which recorded an MD value of −0.828 and a severe injury (LB) value of −0.803. These negative values indicate that safety performance in these regions remains within controllable limits and is considerably safer than the statistical average of other areas in West Java.

Visualization of the map of the West Java province according to its clusters and zones

Figure 4. Cluster Map of West Java Province

Based on this updated map, the regional classification can be explained as follows:

1. Red Zone (C1) – High Risk
   On the map, the regions highlighted in red are Bogor Regency and Subang Regency. This corresponds with the document's findings, which state that:
   a. The Bogor Police jurisdiction consistently dominates this zone due to its extremely high mortality (MD) level, significantly above the provincial average.
   b. Subang is also classified in this category because of its exceptionally high fatality rate.
2. Yellow Zone (C2) – Medium Risk

The areas shaded in yellowish-brown represent major economic and industrial corridors, including Karawang Regency, Purwakarta Regency, Indramayu Regency, Sumedang Regency, Garut Regency, Bandung Regency, and Bandung.

The defining characteristic of this zone is the high frequency of accidents resulting in numerous injuries, although the mortality rate is not as extreme as in the Red Zone.

3. Green Zone (C3) – Low Risk

The regions marked in green, now consistent with the tabulated data, include Sukabumi Regency, Cianjur Regency, Tasikmalaya Regency, Ciamis Regency, Pangandaran Regency, Kuningan Regency, and Cirebon. Mathematically, these regions exhibit negative standardized values, indicating that their safety performance is considerably more stable and safer compared with the average of other areas in West Java.

Table 3. Average of the Most Dominant Clusters in 2024–2025 in West Java Province

Location	Most Frequent Cluster (2024–2025 Period)
Banjar	C3
Bogor	C1
Ciamis	C3
Cianjur	C3
Cimahi	C3
Cirebon City	C3
Garut	C2
Indramayu	C2
Karawang	C2
Bandung City	C2
Greater Bandung	C2
Bogor City	C3
Cirebon City (Municipality)	C3
Kuningan	C3
Majalengka	C3
Pangandaran	C3
Purwakarta	C2
Subang	C1
Sukabumi	C3
Sukabumi City	C3
Sumedang	C2
Tasikmalaya	C3
Tasikmalaya City	C3

Discussion of Results and Hypothesis Validation

The experimental results presented in the previous subsection provide strong evidence supporting the fundamental hypothesis of this study. The initial hypothesis proposed that the K-Means Clustering algorithm could accurately classify regions in Jawa Barat into three risk cluster levels: red, yellow, and green zones. Through dual validation using the Elbow Method and Silhouette Coefficient, it was determined that dividing the data into three clusters (k = 3) produced the most stable structure, with an average silhouette score of 0.398. This finding confirms that the algorithm successfully identified hidden patterns within the fatality variables (fatalities, severe injuries, and minor injuries), thereby establishing clear boundaries between risk zones [4].

The scatter plot distribution analysis indicates that the number of fatalities (MD) is the most dominant variable in separating regions into the Red Zone. Areas with high positive standardized values, such as the Police District of Bogor, were distinctly isolated from the main cluster due to their extreme fatality levels. The algorithm's ability to detect such anomalies without compromising the stability of other clusters (Green Zone) demonstrates that K-Means is a reliable instrument for province-scale risk mapping.

Comparison with Related Studies

To strengthen the validity of these findings, the clustering results were compared with prior studies referenced in this research. Several essential points of comparison are as follows:

1. Algorithm Effectiveness: Consistent with the study by Rahma and Latifah (2024) in Cirebon, the use of K-Means proved effective in identifying accident characteristics based on numerical parameters. However, this study advances the methodology by applying Z-Score standardization, which was not emphasized in the Cirebon study, thereby enhancing objectivity in handling variable scale differences across West Java.
2. Cluster Optimization: While Anshori & Nuraini, 2020 in Tasikmalaya identified four clusters as optimal based on the Davies–Bouldin Index, this study found that, at the provincial scale of West Java, three clusters are more representative and easier for policymakers to interpret. This suggests that the determination of k is highly dependent on the dimensionality and local data distribution.
3. CRISP-DM Methodology: The application of the CRISP-DM framework aligns with the approach used by Afdal & Prana Disastra, 2022 at the Police District of Solok. Adherence to the systematic stages from data preparation to evaluation has proven to produce a model that is not only mathematically robust but also practically relevant for accident risk reduction recommendations.
4. Characteristics of High-Risk Areas: In line with findings by [9] in Brebes, high-risk areas are generally located along primary arterial routes connecting economic centers. This study reinforces that pattern, as industrial buffer zones such as Bogor and Karawang exhibit higher risk levels compared to predominantly administrative regions such as Banjar or Ciamis.

Synthesis of Findings

Overall, this discussion confirms that accident risk levels in West Java are heterogeneous and strongly influenced by fatality-related variables. The implementation of K-Means Clustering effectively addresses the subjectivity traditionally associated with identifying accident-prone areas, which has often relied solely on total incident frequency without considering severity levels. With consistently positive silhouette scores and clear visual separation among risk zones, the model is deemed valid as a foundation for data-driven road safety policy planning in Jawa Barat.

5. Conclusion

This study empirically demonstrates that the implementation of the K-Means Clustering algorithm using the CRISP-DM framework is capable of objectively and measurably classifying traffic accident risk levels in Jawa Barat. Based on the analysis of the Pusiknas Polri dataset for the 2020–2025 period, dual validation testing using the Elbow Method and Silhouette Coefficient determined that a three-cluster structure (k = 3) is the most optimal, with a silhouette score of 0.398. The findings indicate that the number of fatalities is the most dominant variable in determining regional fatality levels. The Red Zone (High Risk) is consistently occupied by areas with dense vehicle concentrations such as Bogor and Subang, while the Yellow Zone (Moderate Risk) is dominated by economic corridors such as Karawang and Indramayu, and the Green Zone (Low Risk) includes areas with above-average safety performance such as Banjar and Ciamis.

These findings address the primary issue of optimizing the use of historical police data, which has previously served mainly administrative purposes, into strategic information for disaster mitigation. Although the study produced a valid risk mapping, it is limited by its reliance on secondary data focusing solely on casualty variables without incorporating external causal factors such as weather conditions or vehicle roadworthiness in depth. Therefore, future research is expected to integrate more dynamic spatial data and employ hybrid algorithms to improve clustering accuracy in areas with highly dense data distributions, enabling transportation safety policies to be implemented more precisely and effectively in the future.

6. References

[1] M. Afdal and R. Prana Disastra, "Analisis Pola Kecelakaan Lalu Lintas Menggunakan Algoritma K-Means dan FP-Growth Studi Kasus: Polres Solok," *Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi*, vol. 8, no. 1, pp. 31–40, 2022.

[2] A. G. Ramadhan, "Data Mining untuk Segmentasi Pelanggan dengan Algoritma K-Means: Studi Kasus pada Data Pelanggan di Toko Retail," *Syntax Literate*, vol. 8, no. 10, 2023, doi: 10.36418/syntax-literate.v6i6.

[3] S. S. Hilabi and B. Huda, "Technoxplore Jurnal Ilmu Komputer dan Teknologi Informasi," vol. 4, 2019.

[4] F. R. Sari, Tukino, S. S. Hilabi, and B. Priyatna, "Klasifikasi Text Ulasan Pengguna Aplikasi Wondr by BNI Menggunakan Algoritma Naïve Bayes," *Jurnal Teknologi Informasi*, vol. 6, no. 2, 2025, doi: 10.46576/djtechno.

[5] S. Setyaningtyas, B. I. Nugroho, and Z. Arif, "Tinjauan Pustaka Sistematis: Penerapan Data Mining Teknik Clustering Algoritma K-Means," *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, vol. 10(2), pp. 52–61, 2022, doi: Https://Doi.Org/10.21063/Jtif.2022.V10.2.52-61.

[6] I. Ferdiansyah, B. Huda, and A. Hananto, "Analisis Clustering Menggunakan Metode K-Means Pada Kemiskinan Di Jawa Timur Tahun 2020," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, pp. 858–869, 2024.

[7] Tukino, R. A. Nanda, Gunawan, S. Wijono, S. Y. J. Prasetyo, and S. Trihandaru, "Analysis Transfer Data Image Processing and Face Recognition Using Camera ESP32CAM Web Browser IoT," *ICIC Express Letters*, vol. 17, no. 6, pp. 717–724, 2023, doi: 10.24507/icicel.17.06.717.

[8] I. F. Anshori and Y. Nuraini, "Pengelompokan Data Kecelakaan Lalu Lintas di Kota Tasikmalaya Menggunakan Algoritma K-Means," *Jurnal Responsif*, vol. 2, no. 1, pp. 118–127, 2020.

[9] T. A. Permana, O. S. Bachri, and R. M. H. Bhakti, "Pemetaan Wilayah Rawan Kecelakaan Lalu Lintas di Kabupaten Brebes Menggunakan Algoritma K-Means," *ELKOM: Jurnal Elektronika dan Komputer*, vol. 18, no. 1, 2025.

[10] D. Safitri, S. S. Hilabi, and F. Nurapriani, "Analisis Penggunaan Algoritma Klasifikasi dalam Prediksi Kelulusan Menggunakan Orange Data Mining," *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 8, no. 1, pp. 75–81, 2023, doi: 10.36341/rabit.v8i1.3009.

[11] Y. Aprianti, Tukino, A. L. Hananto, and S. S. Hilabi, "Klasifikasi Sentimen Komentar Pengguna pada Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes Ruangguru menunjukkan inovasi dalam pendekatan Naïve Bayes menjadi alat klasifikasi teks yang populer . Penerapan Penelitian lain oleh Artanti Inez TF-IDF pen," *METIK JURNAL*, pp. 101–110, 2025, doi: 10.47002/metik.v9i1.1023.

[12] F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, and Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model Crisp-Dm," *Jurnal Teknologi Dan Informasi*, vol. 12, pp. 64–77, 2022, doi: 10.34010/jati.v12i1.

[13] A. Saptiani, B. Huda, E. Novalia, and A. B. Purba, "Pengelompokkan Data Obat-Obatan Pada Pelayanan Kesehatan Menggunakan Algoritma K-Means Clustering," *Jurnal Sistem Informasi dan Manajemen*, vol. 10(3), 2022.

[14] H. Raslin, "Evaluasi Kualitas Aplikasi Digital Korlantas Polri guna Mendukung Tugas Kepolisian dalam Rangka Penguatan Pelayanan Publik," *Jurnal Litbang Polri*, vol. 27, no. 3, pp. 191–215, 2024, doi: 10.46976/litbangpolri.v27i3.245.

[15] M. D. R. P. Dio, B. P. Priyatna, A. L. Hananto, and S. S. Hilabi, "Deteksi Objek Kecelakaan pada Kendaraan Roda Empat Menggunakan Algoritma YOLOv5," *Teknologi*, vol. 12, no. 2, pp. 15–26, 2022, doi: 10.26594/teknologi.v12i2.3260.