The 2nd International Conference on Multidisciplinary Engineering (2nd ICOMDEN 2024)

The 2nd International Conference on Multidisciplinary Engineering: 00012(2024) https://doi.org/10.29103/icomden.v2.705
Volume 2, 2024
eISSN: xxxx-xxxx
Research Original Article

IMPLEMENTATION OF PURITY K-MEANS ALGORITHM IN ACCIDENT DATA CLUSTERING IN NORTH PADANG LAWAS DISTRICT

Khopipah Parawansah Siregar✉[1] **Bustami**[2] **Sujacka Retno**[3]
[1]Information Engineering, Universitas Malikussaleh, Aceh, Lhokseumawe, 24353, Indonesia, khopipah.200170081@mhs.unimal.ac.id
[2]Information Engineering, Universitas Malikussaleh, Aceh, Lhokseumawe, 24353, Indonesia, bustami@unimal.ac.id
[3]Information Engineering, Universitas Malikussaleh, Aceh, Lhokseumawe, 24353, Indonesia, sujacka@unimal.ac.id
✉Corresponding Author: **Email of the corresponding author | Phone: +6285262780503**

Abstract
Traffic safety is an important issue, especially in areas with high accident rates, such as North Padang Lawas Regency in North Sumatra. This study uses the K-Means Purity Algorithm to group regions based on the level of vulnerability to traffic accidents. The data analyzed includes the number of accidents, deaths, serious injuries, and minor injuries from 2019 to 2023. The results of clustering show that some sub-districts have fluctuating levels of vulnerability. Batang Onang District, for example, was categorized as "Not Vulnerable" in 2019 and 2021, but increased to "Vulnerable" in 2020, 2022, and 2023, indicating a spike in risk. In contrast, Dolok District is mostly in the "Not Vulnerable" category, except in 2023. East Halongonan sub-district is almost always in the "Vulnerable" category, indicating a consistently high risk, while Hulu Sihapas and Simangambat experience fluctuations in vulnerability levels from year to year. Ujung Batu, which is generally classified as "Not Vulnerable," indicates an increased risk in certain years. In conclusion, the K-Means algorithm successfully maps accident-prone areas, providing important insights for more effective interventions. This information can help the government in designing better road safety strategies, such as infrastructure improvements and traffic safety awareness campaigns, to reduce future accidents.

Keywords: Traffic Accidents, K-Means Purity Algorithm, Data Mining, North Padang Lawas, Accident Zoning

Introduction
Background
In modern times, traffic safety is an issue that continues to be discussed, especially due to the increasing frequency of road accidents. The WHO estimates that in the next 20 years, fatal road accidents will rise in the rankings as one of the leading causes of death. The number of traffic accidents in Indonesia in the last five years has increased with a significant trend, especially in 2023. According to reports from the National Police Corps and MTI, during 2023 there were around 134,867 accident cases with the majority involving two-wheeled vehicles, mainly due to driver error. Throughout 2023, the number of minor injuries reached the highest in the past five years, with more than 180,000 people injured, while fatal accidents caused more than 5,500 deaths.

North Padang Lawas Regency is one of the districts in North Sumatra Province, Indonesia, which is experiencing rapid population growth. Along with the increase in the population, transportation activities in this region are also increasing. Accident cases in North Padang Lawas Regency are expected to continue to increase every year. The large volume of data recorded makes it difficult to identify accident-prone areas. To overcome this, the use of data mining technology is needed, which is able to help efficiently and quickly in the clustering process to determine accident-prone areas. This technology was chosen because it has the advantage of simplifying and accelerating the analysis of large amounts of data, making it easier to map areas with high potential for traffic accidents.

Data mining is one part of the concept of Big Data, Big Data is generally described through five main characteristics: volume, speed, variation, value, and precision [1]. A data mining technique that can be used to facilitate analysis is clustering with the K-Means purity algorithm. The K-Means purity algorithm is a simple and fast method of clustering, so it is very helpful in determining areas with a high risk of traffic accidents. The use of the Purity K-Means Algorithm makes the analysis process more efficient and optimal. The lack of technology utilization in the accident victim

complaint service system at the Traffic Unit (SATLANTAS) of North Padang Lawas Regency is caused by the use of conventional methods. Due to the conventional reporting system, SATLANTAS has difficulty in collecting information about mapping areas that often experience accidents.

Based on these problems, the researcher designed a clustering model and software implementation to develop zoning of accident-prone areas. This model aims to map high-risk areas, so it can be a reference for motorists to be more careful in the area. In addition, this system will also support an increase in speed in accident information services in North Padang Lawas Regency. By utilizing the Purity K-Means Algorithm, researchers will group regions into 2 categories: Vulnerable and Unvulnerable.

Formulation of the problem

Traffic accidents that often occur are a serious problem that requires proper handling. Therefore, based on this background, a problem is formulated as follows:

((1) How to design a clustering process to map areas prone to traffic accidents in North Padang Lawas Regency District?
(2) How to apply the Purity K-Means Algorithm to identify areas prone to traffic accidents in North Padang Lawas Regency District?

Materials & Methods

Place and Time of Research

Researchers conducted this study starting in February. The location of the research used to collect data is at the North Padang Lawas Regency Traffic Unit. Data obtained from 2019 to 2023..

Research Steps

In this study, various research steps have been prepared which will later be carried out systematically. Here are the 1) Data Collection

At this stage, the researcher collected data on traffic accidents in North Padang Lawas Regency, the data collected was accident criteria data such as data on the number of accidents, the number of deaths, the number of serious injuries, and the number of minor injuries.

(2) Decision Study

At this stage, the researcher collects information carried out by previous researchers who aim to find a problem to be researched.

(3) Data Analysis

From the data that has been collected, data processing will be carried out using the k-means purity algorithm.

(4) Needs Analysis

At this stage, the researcher carries out a stage or process in finding the needs needed in designing a system.

(5) Implementation

The results of grouping accident-prone areas in North Padang Lawas Regency using the purity k-means algorithm were implemented into programming.

(6) System Testing

At this stage, the researcher conducts tests on the system that has been created, which aims to find out whether the system is running or in accordance with the desired.

Research methods

The research methods used are as follows:

(1) Data Mining

Data mining uses two main approaches, namely unsupervised learning and supervised learning. Unsupervised learning is a stage where analysis is carried out without guidance or supervision, meaning that the data does not have labels, while supervised learning involves learning accompanied by guidance from an instructor[2].

(2) Clustering

Clustering is a method in data analysis that is used to group data objects into similar groups based on their characteristics or attributes [3].

(3) Metode Clustering K-means

The K-means algorithm is to divide the dataset into groups called clusters, where the letter K represents many clusters to be formed. Next, the value of K is determined randomly as initialization. On the other hand, "Means" is a middle value that serves as the center of each cluster, also known as a centroid [4]. Here is the formula of the k-means method:

a. Determine how many Clusters (k) to the data set.
b. Initializes the centroid randomly.
c. Using the method of calculating the closest distance to the centroid, you can use the formula (2.1) below:

$$d = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad \text{..... (2.1)}$$

Information:
d = Euclidean Distance.
i = many objects.
x,y = Coordinate point of the object.
s,t = Centroid coordinate point

d. Do the way to group objects based on the closest distance to the centroid.

e. Calculate the average for each group using the formula (2.2) below:

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad \text{..... (2.2)}$$

Information:
V_ij = Centroid average of the ith cluster for the jth variable.
N_i = Number of members of the ith cluster.
i,k = Index of the cluster.
j = Index of variables.
X_kj= The data value of the k-th variable to j for the cluster.
f. Repeat from point 3 and point 4 and iterate until it reaches centroid with optimal value [5].

(4) Purity

A cluster is said to be pure if all data objects of the same class are in the same cluster as well [6]. Purity is used to measure the purity level of a cluster, which is expressed as the number of members in the cluster that best corresponds to a class. To calculate the purity of each cluster, the following formula is used [7]:

$$Purity(j) = \frac{1}{N_j} Max(n_{ij})$$

Information:

Purity(j) = purity value for the jth variable

N_j = the number of data that is a member of the Jth cluster.

i, j = index of the cluster.

(5) Phthon

Python is a programming language that is interpretive and versatile. The language focuses more on code readability, so the syntax is easier to understand, and the Python Software Foundation. Python is compatible with almost all operating systems, including Linux, where most distributions already include Python as part of their system [8].

System Scheme

A system schema is a system of several elements that are interconnected with each other, so it can be said to be a system. The scheme of the accident-prone area system using the purity k-means clustering algorithm is as follows:

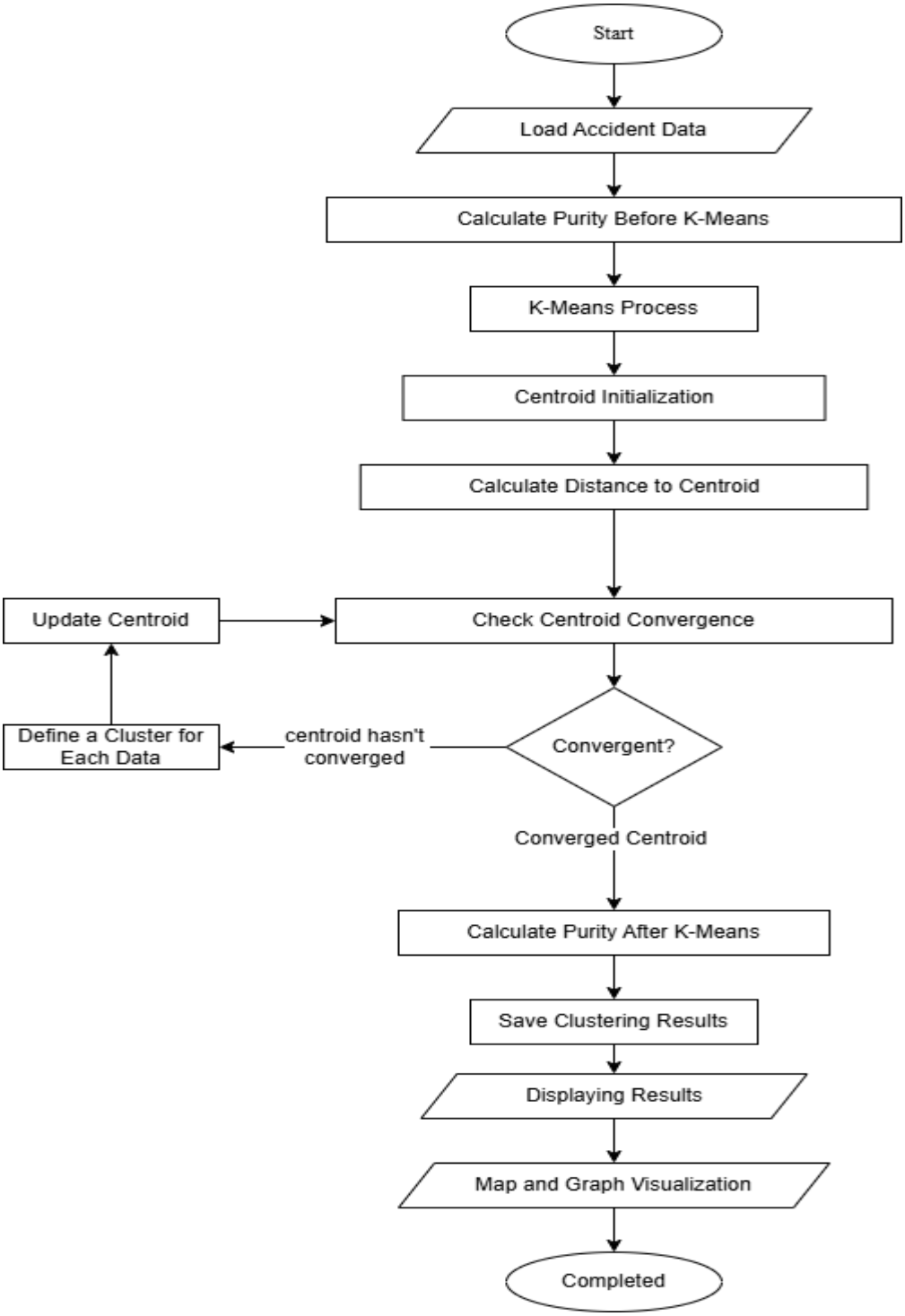

Figure 1. System Scheme

Information:

(1) Start: The process begins with a function call to process the accident data.

(2) Load Accident Data: The first step in the process is to load the accident data from the source (e.g. a file or database) for further analysis.

(3) Calculate Purity Before K-Means: Before carrying out the K-Means process, purity is calculated to assess the quality of the initial data distribution based on the level of accident susceptibility.

K-Means Process: Here begins the main process of data grouping using the K-Means algorithm.

(4) Centroid Initialization: Centroid is first initialized manually, i.e. the central point for each cluster (Prone and Invulnerable).

(5) Calculate Distance to Centroid: Each data is calculated its distance to the centroid that has been initialized, to determine which cluster the data belongs to.

(6) Update Centroid: Once the data is grouped, the centroid is updated by calculating the average position of the data in each cluster.

(7) Check Centroid Convergence: This process checks if the centroid is converged (no longer changing) or still needs to be updated.

(8) Define a Cluster for Each Data: If the centroids have not yet converged, this step is used to define a cluster for each data based on its distance to the centroid.

(9) Convergent?:

a. If the centroid is already convergent (unchanged), then the process continues to the next step.

b. If the centroid hasn't converged yet, the process goes back to the Calculate Distance to Centroid step for further centroid repairs.

(10) Converged Centroid: If the centroid has converged, then the clustering process is considered complete and we proceed to the next step.

(11) Calculate Purity After K-Means: Once the clustering process is complete, the purity is calculated again to measure how well the K-Means algorithm segments the data based on the actual level of vulnerability.

(12) Save Clustering Results: The clustering results that have been obtained are saved, for example in a file format such as Excel or a database for further analysis.

(13) Displaying Results: The clustering results are displayed to the user for further evaluation and analysis. This can be a table or a graph that shows the grouping based on the level of vulnerability.

(14) Map and Graph Visualization: Visualization is done to make it easier to understand the distribution of accidents and their vulnerability level on maps and graphs.

(15) Completed: The process is complete after the clustering results are displayed and visualized, providing a better understanding of the accident data.

Results and Discussion

The calculation test of the system program that has been tested or developed through a manual calculation test aims to obtain the same system calculation results as the calculation outside the system using the purity k-means algorithm method. For testing, data on accident-prone areas in each sub-district of North Padang Lawas Regency was used obtained from the Traffic Unit (SATLANTAS). In the grouping process, the data will be calculated based on the number of data on areas prone to traffic accidents from 2019 to 2023. Here is an example of data used for the manual system testing process taken from 2019:

Table 1. Accident Data in 2019

District	Number of Accidents	Die	Severe Injuries	Minor Injuries
Onang Trunk	93	3	17	19
Dolok	103	3	15	33
Dolok Sigompulon	56	4	12	26
Halongonan	91	7	6	11
East Halongonan	127	7	8	34
Hulu Sihapas	165	4	19	41
Padang Bolak	169	3	12	26
Padang Bolak Julu	92	8	17	17
Southeast Padang Bolak	124	8	12	49
Portibi	153	8	10	45
Simangambat	55	3	16	42
Tip Batu	70	9	10	10

Manual calculation of the Purity K-Means algorithm

Table 2. Traffic Accident Criteria

Criterion
Number of Accidents
Number of Increases
Number of Serious Injuries
Number of minor injuries

a. Number of Accidents (x1)
b. Number of Deaths (x2)
c. Number of Severe Wounds (x3)
d. Number of Minor Wounds (x4)

Determining the Number of Clusters

Table 3. Number of Potential Clusters

Cluster	Potential Label
C1	Prone
C2	Not Prone

Setting the Early Centroid
The selection of the initial center point or centroid was carried out by calculating the purity value on the 2019 accident data.
Formula for calculating purity to find the initial centroid:

$$\text{Purity}(j) = \frac{1}{N_j} \text{Max}(n_{ij})$$

Table 4. Purity Calculation Results

Name of District	Purity Calculation Value
Onang Trunk	0,7045
Dolok	0,6688
Dolok Sigompulon	0,5714
Halongonan	0,7913
East Halongonan	0,7215
Hulu Sihapas	0,7205
Padang Bolak	0,8047
Padang Bolak Julu	0,6865
Southeast Padang Bolak	0,6424
Portibi	0,7083
Simangambat	0,4741
Tip Batu	0,7070

In table 4.9, it can be seen that the highest value is in Padang Bolak District with a value of 0.8047 and the lowest value is in Simangambat District with a value of 0.4741. So Padang Bolak and Simangambat Districts are the initial centroids.

Table 5. Early Centroid

Potential Label	X1	X2	X3	X4
C1	169	3	12	26
C2	55	3	16	42

Initial Iteration Process

The data to be calculated is 2019 data by finding the Euclidean distance to the centroid with the following equation: $d = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$

1. Manual Calculation of Euclidean Distance

   C1.1=$\sqrt{((169\text{-}93+(3\text{-}3+(12\text{-}17+(26\text{-}19)=76.48)^2)^2)^2)^2}$

   C2.1=$\sqrt{((55\text{-}93+(3\text{-}3+(16\text{-}17+(42\text{-}19)=44.42)^2)^2)^2)^2}$

2. Determine the smallest value or minimum value of the entire distance that has been calculated. And then it is described in the form of the table below:

Table 6. Euclidean in 2019

It	Name of District	C1	C2	Nearby	Potential Label	Predicate
1	Onang Trunk	76,48	44,42	44,42	C2	Not Prone
2	Dolok	66,43	48,84	48,84	C2	Not Prone
3	Dolok Sigompulon	113,00	16,55	16,55	C2	Not Prone
4	Halongonan	79,75	48,71	48,71	C2	Not Prone
5	East Halongonan	43,12	72,99	43,12	C1	Prone
6	Hulu Sihapas	17,05	110,04	17,05	C1	Prone
7	Padang Bolak	0	115,18	0	C1	Prone
8	Padang Bolak Julu	77,84	44,94	44,94	C2	Not Prone
9	Southeast Padang Bolak	50,78	69,64	50,78	C1	Prone
10	Portibi	25,41	98,35	25,41	C1	Prone
11	Simangambat	115,18	0	0	C2	Not Prone
12	Tip Batu	100,48	36,34	36,34	C2	Not Prone

3. Assigning a New Centroid

   New Centroid = $\frac{\text{nilai rata−rata cluster 1}}{\text{jumlah data cluster 1}}$

   The following is a table that displays the calculated results of the most recent centroid searches:

Table 7. New Centroid Iteration-1

C1	147,6	6	12,2	39
C2	80	5,2	13,2	22,5

2nd Iteration Process

In the second iteration stage, the centroid used is derived from the previous calculation. The new centroid

values obtained after the initial Euclidean process are shown in the table below. The new centroid will be used for advanced calculations using the Euclidean distance method until similarities with the previous centroid are achieved.

Table 8. 2nd Iteration Process of 2019

It	Name of District	C1	C2	Nearby	Potential Label	Predicate
1	Onang Trunk	58,42	14,16	14,16	C2	Not Prone
2	Dolok	45,18	25,44	25,44	C2	Not Prone
3	Dolok Sigompulon	92,53	24,31	24,31	C2	Not Prone
4	Halongonan	63,45	17,55	17,55	C2	Not Prone
5	East Halongonan	21,63	48,69	21,63	C1	Prone
6	Hulu Sihapas	18,89	87,19	18,89	C1	Prone
7	Padang Bolak	25,21	89,10	25,21	C1	Prone
8	Padang Bolak Julu	60,01	14,01	14,01	C2	Not Prone
9	Southeast Padang Bolak	25,70	51,45	25,70	C1	Prone
10	Portibi	8,60	76,50	8,60	C1	Prone
11	Simangambat	92,77	31,90	31,90	C2	Not Prone
12	Tip Batu	82,92	16,76	16,76	C2	Not Prone

Based on the process of the 1st and 2nd iterations, no difference in value was found but showed the same value, so the iteration process was stopped. Where in the 1st and 2nd iterations there are 5 sub-districts that are included in the vulnerable (C1), namely East Halongonan, Hulu Sihapas, Padang Bolak, Padang Bolak Tenggara and Portibi Districts. Meanwhile, the sub-districts that are included in the non-vulnerable category (C2) are Batang Onang, Dolok, Dolok Sigompulon, Halongonan, Padang Bolak Julu, Simangambat and Ujung Batu Districts.

Visualization Results:

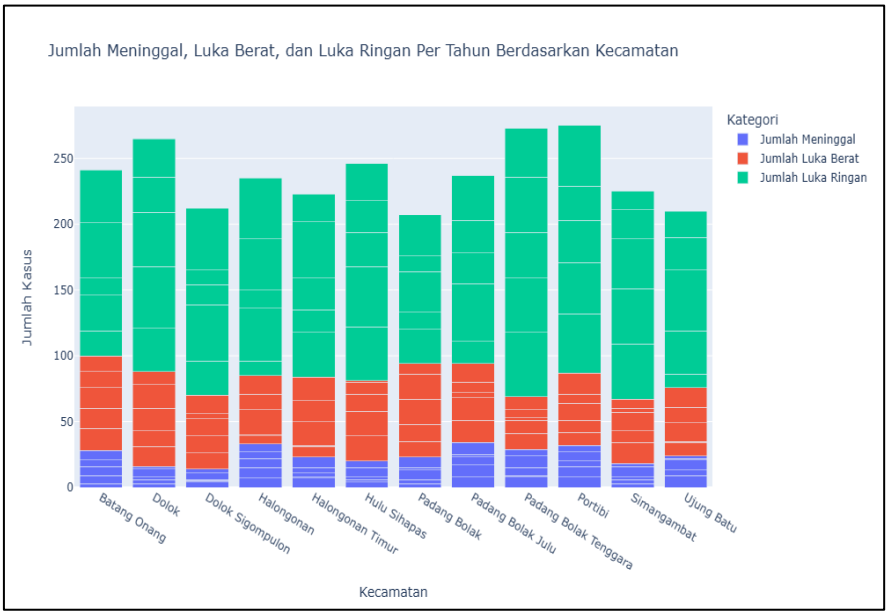

Figure 2. Visualization Results

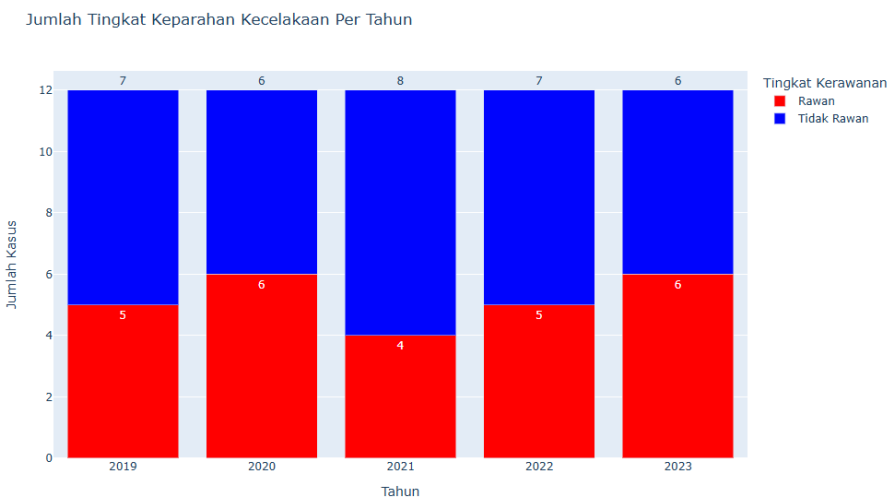

Figure 3. Visualization Results

Mapping Results:

Figure 4. Mapping Results in 2019

Figure 5. Mapping Results in 2020

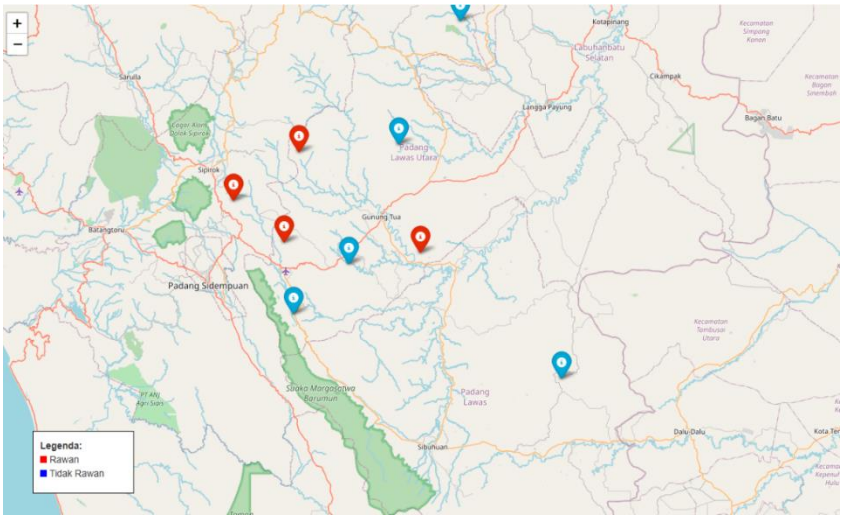

Figure 6. Mapping Results in 2021

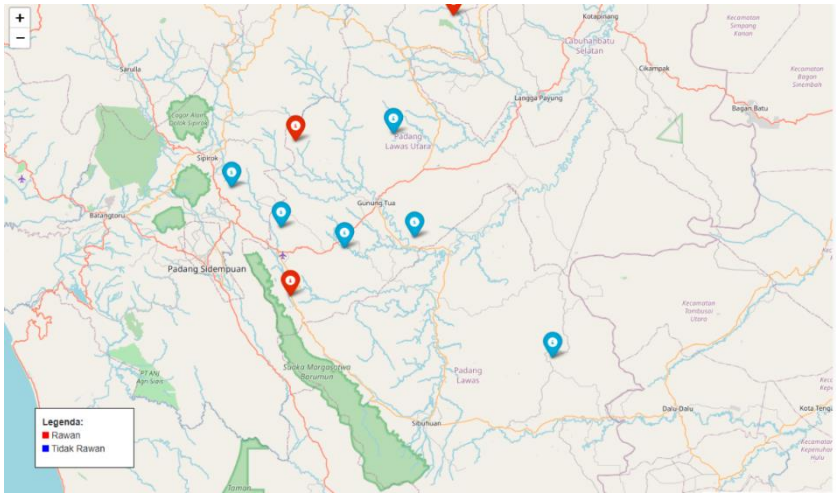

Figure 7. Mapping Results in 2022

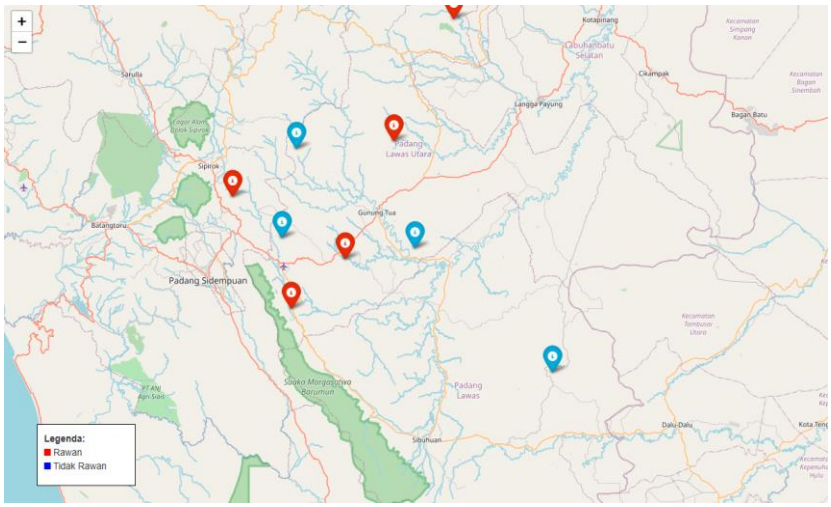

Figure 8. Mapping Results in 2022

Based on the results of clustering from 2019 to 2023, the dominant sub-district with areas prone to traffic accidents is East Halongonan District, where in 2019, 2020, 2021 and 2022 East Halongonan District is always included in

the category of accident-prone areas.

Conclusions

The conclusions obtained from the results of traffic accident data clustering in North Padang Lawas District using the K-Means Purity Algorithm are as follows:

(1) The application of the K-Means Purity Algorithm for clustering traffic accident data has successfully mapped the area in North Padang Lawas into two risk level categories: "Vulnerable" and "Not vulnerable." Using data from 2019 to 2023, this system is able to identify changes in risk patterns in various sub-districts. The results of this clustering are presented in the form of a web-based system, which makes it easier for users, such as policy makers or related parties, to understand the vulnerability conditions in each sub-district. This implementation allows for a more in-depth evaluation of how the number of accidents, fatalities, serious injuries, and minor injuries can affect the level of risk in each region. Overall, the tool offers a data-driven analytics approach that can be used for more effective traffic safety planning.

(2) Analysis of clustering results shows that the level of vulnerability in several sub-districts has fluctuated significantly during the five-year period. For example, East Halongonan and Portibi tend to be consistently in the "Vulnerable" category, indicating that these regions may have stable risk factors, such as inadequate infrastructure or high traffic volumes. In contrast, some sub-districts such as Batang Onang and Padang Bolak Julu show more dynamic patterns of change, switching between "Vulnerable" and "Not Vulnerable" in certain years, indicating external influences such as road improvements or changes in traffic patterns. The changes seen in these areas highlight the importance of flexible and sustainable interventions, where policies must be tailored to the specific conditions of each sub-district. By identifying these trends, authorities can be more proactive in developing targeted road safety strategies, such as the addition of signs, increased patrols, or traffic safety education programs.

Acknowledgments

In the process of this research, the author sometimes experienced several obstacles and difficulties, but thanks to the guidance, assistance, direction and encouragement from several parties, the author would like to express his deepest gratitude to:

1. Prof. Dr. H. Herman Fithra, S.T., M.T., IPM., ASEAN.Eng. as the Chancellor of Malikussaleh University
2. Mr. Bustami, S.Si., M.Si., M.Kom as the main supervisor and Mr. Sujacka Retno, S.T., M.Kom as the assistant supervisor who have provided a lot of guidance, provided criticism, suggestions and very useful motivation.

References

[1] J. Yang *et al.*, "Brief introduction of medical database and data mining technology in big data era," *J. Evid. Based.* Med., vol. 13, no. 1, pp. 57–69, 2020, doi: 10.1111/jebm.12373.

[2] E. E. Pratama, Helen Sastypratiwi, and Yulianti, "Analisis Kecenderungan Informasi Terkait Covid-10 Berdasarkan Big Data Sosial Media dengan Menggunakan Metode Data Mining," J. Inform. Polinema, vol. 7, no. 2, pp. 1–6, 2021, doi: 10.33795/jip.v7i2.453.

[3] R. K. Dinata, H. Novriando, N. Hasdyna, and S. Retno, "Reduksi Atribut Menggunakan Information Gain untuk Optimasi Cluster Algoritma K-Means," J. Edukasi dan Penelit. Inform., vol. 6, no. 1, p. 48, 2020, doi: 10.26418/jp.v6i1.37606.

[4] L. M. Harahap, W. Fuadi, L. Rosnita, E. Darnila, and R. Meiyanti, "Klastering Sayuran Unggulan Menggunakan Algoritma K-Means," J. Tek. Inform. dan Sist. Inf., vol. 8, no. 3, pp. 567–579, 2022, doi: 10.28932/jutisi.v8i3.5277.

[5] R. Risawandi and Y. Afrillia, "Geographic Information System Mapping Of Criminality Villed Areas In Lhokseumawe Using K-Means Method," J. Informatics Telecommun. Eng., vol. 5, no. 2, pp. 442–451, 2022, doi: 10.31289/jite.v5i2.6265.

[6] mohamad jajuli nurul rohmawati, sofi defiyanti, "Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa," Jitter 2015, vol. I, no. 2, pp. 62–68, 2015.

[7] S. Retno, "Peningkatan Akurasi Algoritma K-Means Dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid)," Tesis, no. July 2019, pp. 1–86, 2019, [Online]. Available: https://repositori.usu.ac.id/bitstream/handle/123456789/16782/177038001.pdf?sequence=1&isAllowed=y

[8] Siradjuddiin, Algoritma Pemrograman dengan Menggunakan Python, no. September. 2018.