Implementation of K-Modes Clustering in Determining Traffic Accident Patterns

Septyan Eka Prastya
Sari Mulia University
Department of Information Technology

Muhammad Zulfadhilah
Sari Mulia University
Department of Information Technology

Nurhaeni
Sari Mulia University
Department of Information System

ABSTRACT

The increasing number of traffic accidents in South Kalimantan continues to occur, which needs to be considered by all parties, especially the traffic police. One of the efforts to reduce it is by finding the pattern of traffic accidents through the clustering method. Data from police reports will determine the grouping of traffic accidents based on day, time, victim, type of accident, geometry, age of the perpetrator, age of the victim, weather, Profession of the perpetrator, Profession of the victim, and type of vehicle involved which are some of the factors causing traffic accidents. This study aims to find the pattern of traffic accidents that often occur using the k-modes algorithm and to find the optimal k value; this study uses the Cohesion and Separation algorithm. The application of clustering using the k-modes algorithm will produce a traffic accident pattern based on the optimal k. The results of this study by testing the K-Modes algorithm at K=2, K=3, K=4, K=5, K=6, K=7, K=8, K=9, and K=10 with each experiment. -each k 5 times produces the optimal k value, which is located at K=3 in the 1st Cohesion experiment with a value of 2641. The pattern generated from the K-Modes algorithm has 3 patterns obtained from each cluster for K=3. At the final stage of determining the pattern of traffic accidents, it is known that the first cluster is the cluster with the largest size (61), namely when the weather is sunny, there are double accidents on the road with straight geometric shapes involving motorbikes and motorbikes.

General Terms

Clustering, k-modes algorithm, pattern and traffic accidents.

Keywords

Clustering, Cohesion and Separation, K-Modes, Traffic Accident, Patterns

1. INTRODUCTION

The number of traffic accidents that occurred in South Kalimantan from 2002 to 2007 according to the Ministry of Transportation of the Directorate General of Land Transportation of South Kalimantan Province, in 2002 the number of accidents was 191, and the number of vehicles involved was 358. In 2003 the number of accidents was 232, and the number of vehicles involved was 338. In 2004 the number of accidents was 277, and the number of vehicles involved was 247. In 2005 the number of accidents was 201, and the number of vehicles involved was 290. In 2006 the number of accidents was 1140, and the number of vehicles involved was 735.

Furthermore, the number of accidents in 2007 was 1020, and the number of vehicles involved was 1512. From the above explanation, apart from 2005, the number of accidents in South Kalimantan continued to increase [1]. Based on this, it is necessary to make efforts to reduce the number of accidents. As a first step, data management is needed so that the initial variables that trigger accidents in South Kalimantan, especially in the city of Banjarmasin, can be known.

Efforts to reduce the number of traffic accidents require knowledge to find patterns of traffic accidents that often occur by grouping. Reporting data by the police and the accuracy of police reports will ensure the determination of the classification of traffic accidents based on the day, time, victim, type of accident, driver factor, road factor, age of the perpetrator, age of the victim, weather, Profession of the perpetrator, Profession of victim and type of vehicle involved. Several factors cause accidents. To classify traffic accidents, this study uses the clustering method. Clustering is a data mining method that separates/solves/segments data into a number of groups (clusters) according to specific desired characteristics; in grouping work, the label of each data is not yet known and by grouping, it is hoped that the data group can be identified and then labelled as desired [2]. There are many methods that can be used to perform clustering, one of which is the K-Modes method. This study aims to obtain the pattern of traffic accidents that often occur using the k-modes algorithm. Moreover, to find the optimal k value, this study uses the Cohesion and Separation algorithm. The application of clustering using the k-modes algorithm will produce a traffic accident pattern based on optimal k.

2. RESEARCH METHODE

The research procedure carried out in this study is as follows:

1) Data Collection

The data was obtained from data collection at the Banjarmasin City Police Resort from 2014 - 2016.

2) Data Selection

Stages of selecting relevant data needed for analysis purposes. In selecting the data, the data used is data sourced from traffic accident data. The variables used are day, time, victim, type of accident, geometric shape, age of the perpetrator, age of the victim, weather, Profession of the perpetrator, Profession of victim and type of vehicle involved.

3) Data Integration

Data integration is the stage to combine related data into a single unit. In the traffic accident data in .pdf format, there is no professional data available for the victim and the perpetrator, while the data on the age of the victim and the age of the perpetrator are incomplete, so a new table is made which will later be merged into a single unit in the form of Ms Excel.

4) Data Transformation

The stages of the data transformation process into a form that is ready to be processed in data mining.

5) K-Modes Clustering

This stage is a process for grouping (clustering) the data that has been determined at the transformation stage by using the K-Modes and Cohesion and Separation algorithms and the use of k values of 2, 3, 4, 5,...,10 so that you can find out the value of k. best to determine the accident pattern.

The following is a flow chart of the research procedure:

6) Determination of Optimal K

The stages of determining Optimal K are carried out to see the best K value obtained from the results of clustering k-modes. Determination of Optimal K is done by experimenting 5 times on the system.

7) Pattern Analysis

Pattern analysis is information from research that has been carried out which will provide results to the user from the results of clustering, the patterns generated in this study are in the form of centroid results and cluster members.

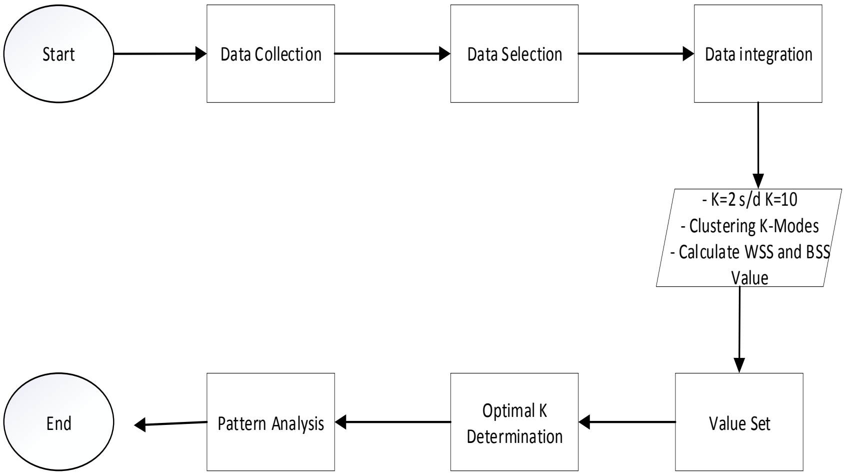


Figure 1 Research Procedure

3. LITERATURE REVIEW

3.1 Traffic Accident

According to Law Number 22 of 2009 concerning Traffic and Road Forces: Traffic Accident is an event on the road that is unexpected and unintentional involving a vehicle with or without other road users resulting in human casualties and/or property loss [3].

Traffic accidents are something that every road user wants to avoid. However, sometimes these traffic accidents happen suddenly because of poor road infrastructure or because of the negligence of the road users themselves [4].

3.2 Data Mining

Data mining is an activity that includes collecting and using historical data to find regularities, patterns or relationships in extensive data. The output of this data mining can be used to help decision making in the future. The development of KDD causes pattern recognition to decrease because it has become part of data mining [5].

Data mining can also be referred to as a series of processes to explore added value in the form of information that has not been known manually from a database by extracting patterns from data to manipulate data into more valuable information obtained by extracting patterns that have been identified. Important or interesting from the data contained in the database [6].

According to Larose, data mining is an analysis of reviewing data sets to find unexpected relationships and summarizing data in a way different from before, which is understandable and useful for data owners. Data mining is a scientific field that combines techniques from machine learning, pattern recognition, statistics, databases, and visualization for handling problems of retrieving information from large databases [7].

Meanwhile, in the journal Tampubolon, et al., it is stated that data mining is also known as Knowledge Discovery in Database (KDD), which is defined as the extraction of potential, implicit and unknown information from a set of data. Process Knowledge Discovery in The database involves the results of the data mining process (the process of extracting the tendency of a data pattern), then converting the results accurately into information that is easy to understand [8].

3.3 K-Modes Algorithm

The K-Modes Clustering Algorithm is an extension of the K-Means Clustering algorithm to cluster categorical data. In data mining, K-Means is the most widely used algorithm for clustering data because it is efficient in massive grouping data. However, the k-means clustering process cannot be applied to categorical data because of the Euclidean remote function and the use of means to represent cluster centres. To use K-Means to group categorical data, it is necessary to convert each unique category into a dummy binary attribute that uses 0 or 1 to indicate that the category value does not exist in a data record [9].

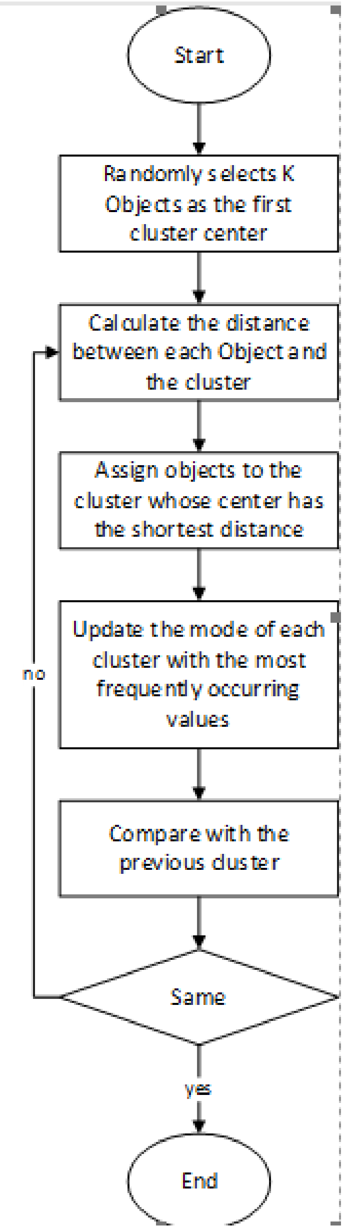

Start
Randomly selects K Objects as the first cluster center
Calculate the distance between each Object and the cluster
Assign objects to the cluster whose center has the shortest distance
Update the mode of each cluster with the most frequently occurring values
Compare with the previous cluster
Same
no
yes
End

Figure 2 K-Modes Clustering Algorithm Flowchart

To group X categorical data sets into k clusters, the k-modes clustering process consists of the following steps [10]:

1. Select the initial mode of k number.
2. Calculate the distance between each object and the cluster mode.
3. Assign objects to the cluster whose centre has the shortest distance to the object; repeat this step until all objects are assigned to a group.
4. It is updating the model (as the centroid) of each cluster with the category value that frequently appears in each cluster.
5. Then compare it with the previous model. If different, return to Step 2.
6. If the same, stop.

4. RESULT

The data collected is from traffic accident data in the Banjarmasin area from 2014 to 2016. The data source was obtained from the Banjarmasin City Police Office in the field of Traffic Accidents totalling 159 records. Traffic accident data contains a record of accident events in the form of causes and consequences of traffic accidents by the police.

The stage of selecting the relevant data needed for analysis purposes. In data selection, not all data and attributes will be used, so only data related to research interests will be used. The data used in this study is data sourced from traffic accident data from the Banjarmasin Police, which is related to the variables used.

There are many variables in traffic accident data. However, to determine the pattern of traffic accidents, only a few variables that have the most influence on determining the pattern is used, including:

1) Day, based on the Big Indonesian Dictionary, is from morning to morning again (i.e. one circle of the earth on its axis, 24 hours). Alternatively, it can be interpreted as the day on which the accident occurred. The days in this data are divided into two, namely weekdays and weekends.

2) In the Big Indonesian Dictionary, time is the entire series of moments when a process, action, or condition exists or takes place. Alternatively, it can be interpreted by when the accident occurred.

3) A victim is someone who is injured the most in an accident. The victims in this data are divided into 4, namely minor injuries, serious injuries, death and no injuries.

4) Type of Accident is the number of vehicles involved in the accident. Types of accidents are divided into two, namely single and multiple.

5) Geometry form is the shape of the road condition at the location of the accident. The geometric shapes are divided into 8 of them: straight, O (roundabout), bend, T (intersection), X or + (intersection), Y (intersection), bridge and TL (intersection 4 is not parallel).

6) The age of the perpetrator is the age of a person who is a suspect in an accident. The age of the perpetrators was categorized into 9.

7) Age of Victim is the age of a person who is a victim of an accident. Similar to the age of the perpetrator, the age of the victim is also categorized into 9.

8) Weather is the state of nature at a time that is relatively short and changing. Weather is divided into 3, namely sunny, rainy/drizzle and cloudy.

9) The profession of a Perpetrator is the occupation of a person who is a suspect in an accident. The professions of the perpetrators are divided into 11, including female students, private sector, students, drivers, labourers, housewives, civil servants, students, employees, nurses and teachers.

10) A victim's profession is the occupation of someone who is a victim in an accident. Several victim professions are different from the perpetrator's profession. The victim's profession is divided into 16, namely students, homemakers, private sector, labourers, civil servants, police, retirees, students, flats, employees, drivers,

tradespeople, parking attendants, nurses, college students and teachers.

11) Type of Vehicle Involved is the type of transportation involved in the accident. There are several types of transportation involved in this accident, namely, cars, motorbikes, pedal bikes, tricycles, 3-wheeled vehicles, etc. However, in this data, the types of transportation used are simplified into 3, including cars, motorcycles, etc. (pedal bicycles, tricycles, 3-wheeled vehicles, etc.)

The variables needed in this study will be combined into data that is ready to be processed through the data integration stage.

The data transformation stage is carried out by transforming numeric data into categorical data. The following data is transformed from numeric to categorical form:

Table 1 Age data transformation

ID Number	Victim's Age	
	Numerical	Categorical
1	15	Teenager
2	47	Elderly
3	45	Adult
4	29	Adult
5	18	Teenager
6	15	Teenager
7	50	Elderly
8	74	Seniors
9	19	Teenager
10	58	Elderly
11	32	Adult
12	57	Elderly

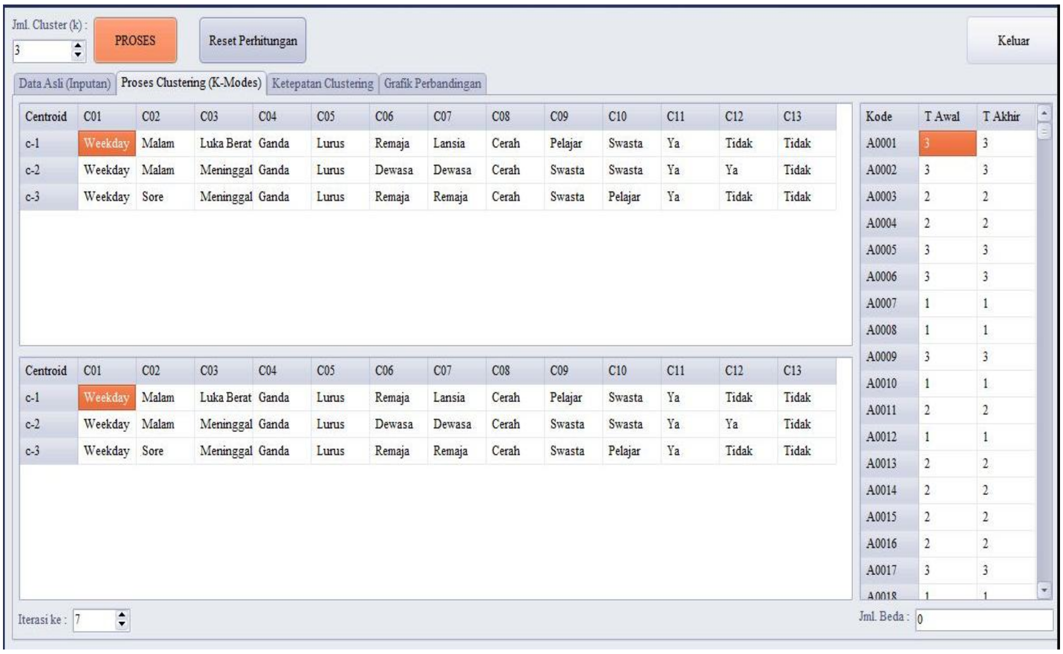
The table below is part of the time data that has been carried out in the transformation stage:

Table 2 Time Data Transform

ID Number	Time	
	Numerical	Categorical
1	13:15	Afternoon
2	02:00	Night
3	21:15	Night
4	07:00	Morning
5	07:30	Morning
6	11:00	Afternoon
7	07:40	Morning
8	07:30	Morning
9	08:30	Morning
10	12:00	Afternoon
11	10:00	Afternoon
12	16:00	Evening
13	21:15	Night
14	23:00	Night
15	13:15	Afternoon

The data that has been selected and carried out in the transformation stage will proceed to the next stage, namely data integration. Data integration is the stage to combine related data into a single unit. After the traffic accident data has been integrated, the data will be processed by data mining using the clustering method with the K-Modes algorithm and producing the best k value using the Cohesion and Separation method.

The next stage is the stage of being implemented into the system; at this stage, a random centroid is also carried out so that the results of calculations with the system will be different from the results done manually. The results obtained using the system at K = 3 are as follows:



Jml. Cluster (k) : 3

Centroid	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13
c-1	Weekday	Malam	Luka Berat	Ganda	Lurus	Remaja	Lansia	Cerah	Pelajar	Swasta	Ya	Tidak	Tidak
c-2	Weekday	Malam	Meninggal	Ganda	Lurus	Dewasa	Dewasa	Cerah	Swasta	Swasta	Ya	Ya	Tidak
c-3	Weekday	Sore	Meninggal	Ganda	Lurus	Remaja	Remaja	Cerah	Swasta	Pelajar	Ya	Tidak	Tidak

Centroid	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13
c-1	Weekday	Malam	Luka Berat	Ganda	Lurus	Remaja	Lansia	Cerah	Pelajar	Swasta	Ya	Tidak	Tidak
c-2	Weekday	Malam	Meninggal	Ganda	Lurus	Dewasa	Dewasa	Cerah	Swasta	Swasta	Ya	Ya	Tidak
c-3	Weekday	Sore	Meninggal	Ganda	Lurus	Remaja	Remaja	Cerah	Swasta	Pelajar	Ya	Tidak	Tidak

Kode	T Awal	T Akhir
A0001	3	3
A0002	3	3
A0003	2	2
A0004	2	2
A0005	3	3
A0006	3	3
A0007	1	1
A0008	1	1
A0009	3	3
A0010	1	1
A0011	2	2
A0012	1	1
A0013	2	2
A0014	2	2
A0015	2	2
A0016	2	2
A0017	3	3
A0018	1	1

Iterasi ke : 7

Jml. Beda : 0

Figure 3 Last Centroid For K=3

Determination of the optimal K value is carried out to determine the best K value. In the third step, the K-modes algorithm is doing a random centroid. The existence of this random centroid can cause the cohesion and separation values at K=2, K=3, K=4, K=5, K=6, K=7, K=8, K=9 and K=10 to change every time. Times the k-modes algorithm is repeated. The change in cohesion and separation values every time the k-modes algorithm is repeated will cause changes to the optimal K value. Thus, to determine the best K value, the k-modes algorithm experiment was carried out on the system implementation 5 times using cohesion and separation charts.

From the results obtained in this study, it can be concluded that the calculation of the distance of the K-Modes algorithm will produce values in the form of integers so that it becomes more difficult to enter these values into each cluster. In comparison, the results of the K-Means algorithm are decimal numbers so that the distance results are always different. So, the drawback of the K-Modes algorithm in this study is that it cannot determine the closest cluster with certainty at all times because the distance calculation in the K-Modes algorithm uses 0 or 1 input so that the results of the distance value may have the same value and, in this study, if the distance values are the same, then the maximum distance value is used as a cluster because, in the IF function, the cluster determination is checked from the leading one.

To overcome this deficiency, another distance calculation is needed because with the current distance formula, variables that have levels cannot be seen. When viewed in terms of the time variable, the time variable has levels; for example, morning to morning, the value is 0, morning to afternoon is 1/3, morning to evening is 1/2 and morning to night is 1, so actually, there are levels for these variables. Certain. There are 4 types of levels in a variable, namely purely categorical, ordinal attributes, numerical intervals, and ratios. There are some variables that are not purely categorical but have a hierarchy, so it is necessary to use another distance formula. At the same time, the calculation of the distance of the K-Modes algorithm, which is now more precise, is only used for purely categorical variables because it ignores the hierarchy.

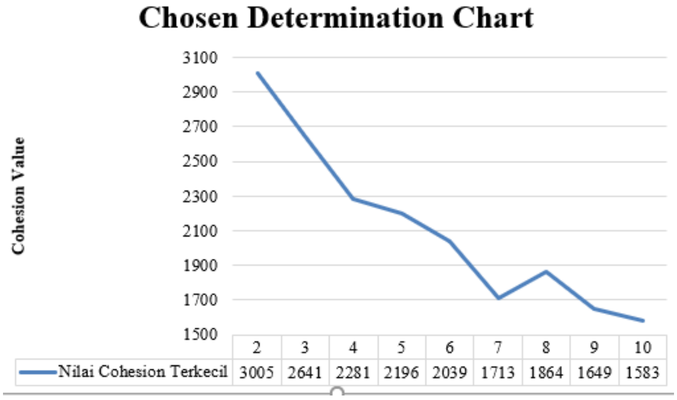


Figure 4 Chosen Determination Chart

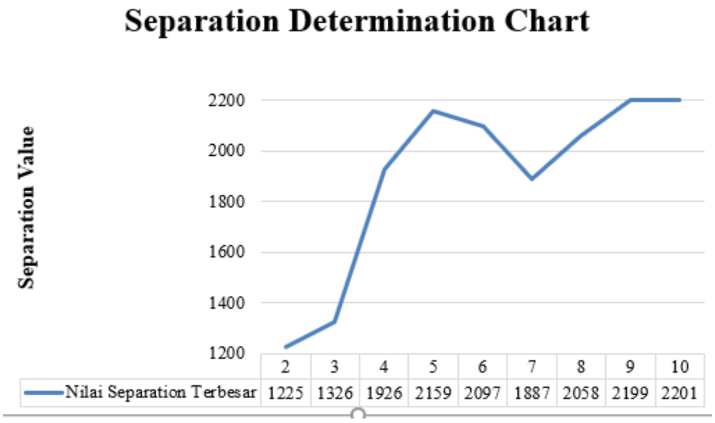


Figure 5 Separation Determination Chart

The results of the experiment using the cohesion graph show that the optimal K is at K=3 and the separation graph shows that the optimal K is at K=4, for the best cohesion and separation values are in the 1st experiment with values of 2641 and 1926, respectively, but the best value of separation is only used as support because the pattern results obtained only focus on the best cohesion value. So even though the best value and the best K from the cohesion and separation experiments are different, what is determined to be a pattern is only from the cohesion value.

Further analysis resulted from traffic accident patterns for all clusters, namely:

- Traffic accidents from the 3 patterns all occur on weekdays, which are 3 clusters.
- At the time of traffic accidents in most of the clusters, as many as 2 clusters occurred at night, and a small portion occurred in the afternoon as many as 1 cluster.
- The weather at the time of the traffic accident of the 3 patterns all occurred when the weather was sunny, namely 3 clusters.
- The types of traffic accidents from the 3 patterns all have multiple types of accidents.
- Accidents occur on roads with straight geometric shapes in all traffic accident cluster patterns.
- In all clusters, there are 3 types of motorized vehicles, and a small number of cars are 1 cluster.

- Then most of the clusters of perpetrators are teenagers, with the profession of perpetrators being students and the private sector, with different ages of victims in each cluster pattern, namely the elderly, adults and adolescents with the profession of victims being mostly private and a small part are students.
- In traffic accidents, the status of most of the victims in the cluster is dead with 2 clusters, and a few are seriously injured with 1 cluster.

5. CONCLUSION

The conclusions from the research that has been carried out are as follows:

1) The optimal K value in the K-Modes algorithm in determining the pattern of traffic accidents is obtained after 5 experiments, which is located at K = 3.
2) The pattern generated from this study consists of 3 cluster patterns, namely:
   a. The pattern in the first cluster is Day = Weekday, Time = Night, Victim = Serious Injury, Type of Accident = Double, Geometry Shape = Straight, Age of Perpetrator = Teenager, Age of Victim = Elderly, Weather = Sunny, Profession of Perpetrator = Student, Profession of Victim = Private, Motorcycle = Yes, Car = No, Other Vehicle Type = No.
   b. The pattern in the second cluster is Day = Weekday, Time = Night, Victim = Died, Type of Accident = Double, Geometry Shape = Straight, Age of Perpetrator = Adult, Age of Victim = Adult, Weather = Sunny, Profession of Perpetrator = Private, Profession of Victim = Private, Motorcycle = Yes, Car = Yes, Other Vehicle Type = No.
   c. The pattern in the third cluster is Day = Weekday, Time = Afternoon, Victim = Died, Type of Accident = Double, Geometry = Straight, Age of Perpetrator = Teen, Age of Victim = Teen, Weather = Sunny, Profession of Perpetrator = Private, Profession of Victim = Student, Motorcycle = Yes, Car = No, Other Vehicle Type = No

6. ACKNOWLEDGMENTS

Thank you to the "Ministry of Research, Technology and Higher Education" for funding this research so that it can be carried out to completion.

7. REFERENCES

[1] B. P. S. Indonesia, “Statistik Transportasi Darat 2015,” 2015.

[2] D. T. Nugrahadi and F. I. S. Rahayu, “Clustering Penentuan Potensi Kejahatan Daerah Di Kota Banjarbaru Dengan Metode K-Means,” *Kumpul. J. Ilmu Komput.*, vol. 1, no. 1, pp. 33–45, 2014.

[3] A. RI, “Undang-Undang Republik Indonesia Nomor 22 Tahun 2009 Tentang Lalu Lintas dan Angkutan Umum,” 2009.

[4] P. C.E, “Analisis Karakteristik Kecelakaan dan Faktor Penyebab Kecelakaan Pada Lokasi Blackspot di Kota Kayu Agung,” *Tek. Sipil dan Lingkung.*, vol. 2, no. 1, pp. 154–161, 2014.

[5] Fadlina, “Data Mining Untuk Analisa Tingkat Kejahatan Jalanan,” *Inf. dan Teknol. Ilm.*, vol. 3, no. 1, pp. 144–154, 2014.

[6] Hexagraha A., “Data Mining Kredit Usaha Mikro di Bank XXXX,” in *Konferensi Nasional Sistem Informasi 2014, STMIK Dipanegara Makassar*, 2014, vol. 1, pp. 1–5.

[7] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining: Second Edition*. 2014.

[8] P. . Simbolon, “Implementasi Data Mining Pada Sistem Persediaan Barang Menggunakan Algoritma Apriori (Studi Kasus : Srikandi Cash Credit Elektronic dan Furniture),” *J. Ris. Komput.*, vol. 6, no. 4, pp. 401–406, 2019.

[9] A. . Prakash, “Review on K-Mode Clustering,” *Int. J. Eng. Comput. Sci.*, vol. 5, no. 11, 2016.

[10] E. K. Nduru, E. Buulolo, and P. Pristiwanto, “IMPLEMENTASI ALGORITMA K-Modes UNTUK MENENTUKAN STRATEGI MARKETING STMIK BUDI DARMA,” *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 12–19, 2018, doi: 10.30865/komik.v2i1.903.