PAPER • OPEN ACCESS

K-Means Algorithm for Clustering The Location Of Accident-Prone On The Highway

To cite this article: Diah Puspitasari et al 2020 J. Phys.: Conf. Ser. 1641 012086

View the article online for updates and enhancements.

You may also like

- Information System Based Excellent Higher Education Model Anuar Sanusi, Suhendro Y. Irianto, Herlina et al.
- <u>The technological knowledge and the</u> <u>content knowledge on acid-base concepts</u> <u>of senior high school STEM students</u> Edna B. Nabua, Jorgy O. Falcasantos and Maurine Joy Y. Jerez
- Distribution of schistosomiasis intermediate snail in Lore Lindu National Park, Central Sulawesi
 J Widjaja, Samarang, M Koraag et al.

The Electrochemical Society Advancing solid state & electrochemical science & technology



DISCOVER how sustainability intersects with electrochemistry & solid state science research



This content was downloaded from IP address 3.15.10.137 on 07/05/2024 at 13:11

K-Means Algorithm for Clustering The Location Of Accident-Prone On The Highway

Diah Puspitasari¹, Mochamad Wahyudi², Muhammad Rizaldi³, Acmad Nurhadi⁴, Kresna Ramanda⁵*, Sumanto⁶

 $^{1,6}{\rm Program}$ Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia

²Program Studi Teknologi Informasi, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia

³Program Studi Sistem Informasi, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Indonesia

⁴Program Studi Teknologi Komputer, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika, Indonesia

 $^5\mathrm{Program}$ Studi Teknik Informatika, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Indonesia

E-mail: kresna.kra@bsi.ac.id

Abstract. In Indonesia, there is a highway which connects Jakarta and Bogor, some accidents have occurred at the highway every year. This paper aims to analyze the location of the accident to classify the high and low levels of accident vulnerability in the Jakarta Bogor Highway. First, the extracted data is then grouped based on some of the same characteristics in the dataset namely cause, location, minor injuries, serious injuries and death. Second, the grouping results are visualized in the form of highway maps that can help highway managers in identifying and evaluating several accident-prone points on Jakarta Bogor Highway. The method to be used in data processing in this study is the K-Means clustering algorithm which is expected to produce useful information for Jakarta Bogor Highway managers. The results of this study indicate that accidents that often occur are in cluster 3 with a total of 80 accidents and at least there are in cluster 2 with a total of 57 accidents, the location of accidents that often occurs in cluster 1 is in KM 24, while cluster 2 is in KM 41 and cluster 3 located at KM 10.6.

1. Introduction

Accidents are things that cannot be predicted when and where they will occur, traffic accidents have increased every year due to the increasing number of motor vehicles such as motorcycles, cars and various other types of vehicles which add to the density of traffic on arterial and highway. Jakarta Bogor Highway connecting Jakarta, Cibubur, Citeureup, Bogor, and Ciawi which has experienced many traffic accidents every year that are saved. So the authors intend to analyze the location of the accident to classify the high and low levels of accident vulnerability at Jakarta Bogor Highway[1]. The World Health Organization (WHO) records nearly 3,400 people die on the road every day. Traffic accidents are ranked number 2 cause of death after a stroke. Kepolisian Republik Indonesia (POLRI) records that 80 people die every day due to traffic accidents or 3 people per hour during 2013[2]. Data mining is a computational process for finding patterns in large data sets. The process is often described as discovery knowledge

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI. Published under licence by IOP Publishing Ltd 1

in databases including: the use of algorithms, statistical tools, and machine learning to extract previously unknown patterns[3]. Clustering is an unsupervised data mining method, because there is no single attribute used to guide the learning process, so all input attributes are treated equally[4]. Discovery in Database (KDD) is a field of science that discusses a lot about the pattern of a data. A series of processes to obtain knowledge or patterns from a collection of data is called data mining[5]. The process of knowledge discovery 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 easily understood[6]. This research will be useful in assisting Jakarta Bogor Highway managers in identifying and evaluating accident-prone points so that safety monitoring can be carried out and anticipating accidents.

Flight is a thing that is often done in this very modern era. Community mobility requires faster transportation, but each transportation has its own risks. Tests carried out experienced a simplification of variables from the dataset, the variables used were only flighttype, investigation type, aircraft demange and number of engines. And take flight type as the final result[7]. Traffic accidents are unintentional road events involving vehicles with or without other road users, resulting in losses for the victims. The ordinal logistic regression model that was formed can be used to calculate the accuracy of the classification of the severity of traffic accident victims that is equal to 90.5405%. While the results of the FK-NNC analysis showed that at K = 1 the accuracy classification of the severity of traffic accident victims was obtained by 89.19%[8]. There are several algorithms that can be used for accident data clustering including K-Medoids[9] analyze the causes of maritime accidents and further ensure the safety of ship navigation, the risks of ships at sea and their impact factors are identified, Fuzzy C-Means[12] In this research will be made an analysis to accident-prone areas in Semarang City with Geographic Information System, Grid Clustering[11] Identifying road accident blackspots is an effective strategy for reducing accidents.

2. Methods

2.1. Research Stages

The steps carried out in the research stages can be explained as follows:

(i) Identification of problems.

The problem is limited by the scope of the existing problems, the problem is to find out the location of traffic accident prone in Jakarta Bogor Highway

(ii) Literature review.

Done by studying and understanding the theories used, namely data mining methods with k-means algorithm. The data is sought by collecting literature, journals and books related to topics in the form of e-books or papers.

(iii) Data collection

Data collection is a way to obtain data that will be needed in research. Collecting data in this study by researching Accident data on the Jakarta Bogor Highway from 2018 to 2019.

(iv) Method Selection The author determines the method of data processing by using data mining clustering

methods with k-means algorithm to get the results of grouping accident-prone locations.

(v) Data Processing

The data processing of this research uses clustering method, with k-means algorithm on Jakarta Bogor Highway accident data. In order to get accident-prone locations to help reduce accidents that occur on the highway.

(vi) Evaluation of Research Results

At this stage the authors test the results using the Rapid Miner software. So the information generated by the data mining process can be displayed in a form that is easily understood for research.

(vii) Conclusion of the Study Results

At this stage the research that has been completed will be given a conclusion of the existing problems that can be used by Jakarta Bogor Highway.

2.2. K-Means Algorithm

K-Means is a partition-based clustering method. This method is very simple, starting with the selection of the number of clusters of K pieces. Furthermore, K data is taken pickled from the dataset as a centroid representing a cluster. Centroid is the center or midpoint of a cluster[13]. K-Means algorithm divides data into clusters so that data that has the same characteristics are grouped into the same cluster and data that has different characteristics are grouped into other clusters[6]. To do the calculation using the Euclidian Distance formula as follows[14]:

$$d = |x - y| \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$
(1)

d = calculation of distance to the center of the cluter

 $\mathbf{x} =$ point coordinates of the object

 $\mathbf{y} = \mathbf{centroid}$ coordinate

 $\sum (i=1)^n = n$ is the amount of data to be measured the distance, while i = 1 is the clustering process starting from the first iteration

 x_i =coordinate point of the i object

 $y_i = i$ centroid coordinate point

2.3. Novelty

There are various studies conducted on the classification of traffic accident data in various regions [1][2][6]. In this study focuses on the grouping of accident-prone points that occur on the Jagorawi highway based on the criteria of cause, location, number of minor injuries, number of seriously injured and the number of victims who died.

3. Results and Discussions

3.1. Dataset

In this research, the location of accident-prone locations at Jakarta Bogor Highwayis required to attribute the location of the accident, cause, minor injuries, serious injuries and death, to be applied in the calculation of data mining using the K-Means algorithm using accident data sets from 2018 to 2019 as the table below:

Ta	able	1. Attributes Dataset
	No	Attributes
	1	Date
	2	Because
	3	Location
	4	Lr (Minor Injuries)
	5	Lb (Serious Injuries)
	6	K (Death)

The dataset used in this study is accident data that occurred in the Jakarta Bogor Highway from 2018 to 2019 with a total of 198 lines of data.

3.2. Centroid Data

The initial stage in clustering using K-means is to determine the number of clusters, in this study the number of selected clusters is 3 clusters. Then determine the cluster center point (centroid), in this study the cluster center point is chosen randomly, namely C1 from the 137th data, C2 from the 10th data and C3 from the 47th data.

Table 2. Attributes Dataset						
Cluster	Because	Location	Lr	Lb	Κ	
C1	3	25.200	0	0	0	
C2	3	40.000	0	1	0	
C3	3	12.300	2	0	0	

3.3. Implementation of K-Means on Rapidminer.

In this study using Rapidminer as supporting software for the implementation of K-Means, the following is a display of K-Means algorithm modeling using rapidminer. :



Figure 1. Design of K-Means Algorithm Model.

The picture above is a K-Means algorithm modeling to determine the location of accidentprone using Rapidminer, First the dataset to be processed is taken using the Read Excel object, then the dataset is processed using the K-Means algorithm contained in the clustering object, the results of data processing then measured modeling accuracy using performance object to measure the level of accuracy.

Figure 2. Shows that of 198 datasets that were processed and divided into 3 clusters, the data were grouped into cluster 0 (not accident-prone) as many as 61 data, cluster 1 (accidentprone) as many as 57 data and cluster 2 (very accident-prone) as many 80 data. Centroid data generated from the processing of accident-prone location datasets totaling 198 data using rapidminer is shown in Figure 3 below.

Clust	er	N	lodel
Cluster	0:	61	items
Cluster	1:	57	items
Cluster	2:	80	items
Total nu	ambe	er d	of items: 198

Figure 2. Result of K-Means Cluster Model.

Attribute	cluster_0	cluster_1	cluster_2
с	2.410	2.912	2.413
D	25285.246	40115.789	12131.250
LR	1	1.368	0.925
LB	0.410	0.439	0.637
K	0.098	0.158	0.163

Figure 3. Result of Centroid Data.

4. Conclusion

Based on the results of research, data processing and analysis that has been done, the conclusion that can be drawn is the level of traffic vulnerability that occurs in the Jakarta Bogor Highway can be known by the K-Means algorithm. Accidents that often occur are in cluster 3 with 80 accidents and the least is in cluster 2 with 57 accidents. The most common cause of accidents in each cluster is lack of anticipation in driving. Location of accidents that often occur in cluster 1 is at KM 24, while cluster 2 is at KM 41 and cluster 3 is at KM 10.6.

For the future work, it is necessary to collect data that contains more about the highway accident that occurred with a wider area coverage so that the data obtained has diversity so that the grouping process can be optimized. In addition, the use of different clustering algorithms to get a comparison of processing data generated.

References

- IW. Aprianti and J. Permadi, 2018, "K-Means Clustering Untuk Data Kecelakaan Lalu Lintas Jalan Raya Di Kecamatan Pelaihari," J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 5, pp. 613–620.
- [2] E. Wicaksono, Kusrini, and E. T. Lutfi, 2017, "Analisis Data Kerawanan Kecelakaan Lalu Lintas Menggunakan Metode K-Means (Study Kasus Polres Bantul)," Semin. Nas. Teknol. Inf. dan Multimed. 2017, vol. 5, no. 1, pp. 110–114.
- [3] D. Puspitasari, S. S. Al Khautsar, and W. P. Mustika, 2019, "Algoritma Naïve Bayes Untuk Memprediksi Kredit Macet Pada Koperasi Simpan Pinjam," J. Inform. Upgris, vol. 4, no. 2.
- [4] R. T. Wulandari, 2017, Data Mining Teori dan Aplikasi Rapidminer, 1st ed. Yogyakarta: Gava Media.
- [5] Ardiyansyah, P. A. Rahayuningsih, and R. Maulana, 2018, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," J. Khatulistiwa Inform., vol. VI, no. 1, pp. 20–28.
- [6] E. Purwaningsih, 2019, "Analisis Kecelakaan Berlalu Lintas Di Kota Jakarta Dengan Menggunakan Metode K-Means," JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer), vol. 5, no. 1, pp. 139–144.
- [7] A. Chandra, 2017, "Penerapan Data Mining Menggunakan Pohon Keputusan Dengan Algoritma C4.5 Dalam Menentukan Kecelakaan Penerbangan," Semin. Nas. Teknol. Inf. dan Multimed. 2017, vol. 5, no. 1, pp. 91–96.
- [8] C. Silvia, Y. Wilandari, and A. Hoyyi, 2015, "Ketepatan Klasifikasi Tingkat Keparahan Korban Kecelakaan

Lalu Lintas Menggunakan Metode Regresi Logistik Ordinal Dan Fuzzy K-nearest Neighbor In Every Class," vol. 4, no. Data Mining, pp. 441–451.

- [9] B. Yang, Z. Zhao, and J. Ma, 2018, "Marine accidents analysis based on data mining using K-medoids clustering and improved A priori algorithm," IOP Conf. Ser. Earth Environ. Sci., vol. 189, no. 4.
- [10] A. D. R. Pradipta, M. Awaluddin, and A. L. Nugraha, 2018, "Pemetaan Daerah Rawan Kecelakaan Di Kota Semarang Dengan Menggunakan Metode Cluster Analysis (Studi Kasus: Kecamatan Banyumanik Dan Tembalang)," J. Geod. Undip, vol. 7, no. 4, pp. 185–194.
- [11] L. Shen, J. Lu, M. Long, and T. Chen, 2019, "Identification of accident blackspots on rural roads using grid clustering and principal component clustering," Math. Probl. Eng., vol. 2019.
- [12] S. Adinugroho and Y. A. Sari, 2018, Implementasi Data Mining Menggunakan Weka, 1st ed. Malang: UB Press.
- [13] W. S. Azis and D. Atmajaya, 2016, "Pengelompokan Minat Baca Mahasiswa menggunakan Metode K-Means," Ilk. J. Ilm., vol. 8, no. 2, pp. 89–94.