Journal of Applied and Research Computer Science and Information Systems
Vol. 2, No. 1, June 2024, pp. 159~170
ISSN: 2988-294X, DOI: 10.61098/jarcis.v2i1.182

K-Means Clustering for Identifying Traffic Accident Hotspots in Depok City

Herry Wahyono[1], Hari Setiaji[2], Tri Hartati[3], Ninuk Wiliani[4](✉)
[1] Krisnadwipayana University, Indonesia
[2] Pancasila University, Indonesia
wahyonos2000@unkris.ac.id, C70101180012@aeu.edu.my, ninuk.wiliani@univpancasila.ac.id

Article Info

Article history:

Received May 20, 2024
Revised May 25, 2024
Accepted June 20, 2024

Keywords:

K-Means Clustering
Traffic Accidents
Depok City
Decision Support
Road Safety

ABSTRACT

This study applies the K-Means clustering algorithm to support decision-making processes related to identifying traffic accident-prone areas in Depok City over a three-year period (2020-2022). Secondary data was obtained from the Traffic Accident Unit of the Depok Metro Police, encompassing monthly traffic accident recapitulations for each district. The data underwent preprocessing steps, including integration and selection of relevant attributes. Using RapidMiner, the data was clustered into three distinct groups, with the optimal number of clusters determined by the Davies-Bouldin Index (DBI), which yielded a score of 0.896, indicating a satisfactory clustering result. The findings reveal that four districts—Beji, Cimanggis, Pancoran Mas, and Sukmajaya—are identified as high-risk areas for traffic accidents. These results are expected to assist local authorities in implementing targeted safety measures. The study demonstrates that the K-Means clustering method is a viable tool for analyzing traffic accident data and can significantly contribute to improving road safety in urban areas.

Corresponding Author:

Ninuk Wiliani

Pancasila University
Email: ninuk.wiliani@univpancasila.ac.id

1. INTRODUCTION

Depok City is one of the cities in Indonesia experiencing rapid population growth[1]. According to the Central Statistics Agency of Depok City, the population in 2021 reached 2,085,935 people, an increase of 29,535 people (1%) from 2,056,400 in the previous year[2]. This growth is primarily due to the city's proximity to the capital, Jakarta, and its status as a residential supporting city[3].

As population density increases, so does the mobility of residents, particularly in terms of transportation needs[3]. This increased mobility contributes to a higher number of vehicles on the roads, which in turn is a significant factor in the occurrence of traffic accidents[4]. The more vehicles on the road, the less space there is between them, reducing safe distances and increasing the risk of accidents[5]. According to Law No. 22 of 2009, Article 1 Paragraph 24, a traffic accident is defined as an unforeseen and unintended event on the road, caused by a vehicle with or without other road users, resulting in human casualties and/or property damage[6]. In the past three years, from 2020 to 2022, 1,584 traffic accidents have occurred in Depok City[7]. This figure is derived from a summary of accidents across all 13 districts in Depok City, including Bojong Gede, Beji, Cimanggis, Tapos, Cinere, Limo, Cipayung, Pancoran Mas, Bojongsari, Sawangan, Cilodong, Sukmajaya, and Tajur Halang[8].

Given the high number of traffic accidents, there is a pressing need for information on areas within Depok City that are prone to traffic accidents[9][10]. Such information is crucial for minimizing the number

of accidents in these high-risk areas in subsequent years[11]. However, the Traffic Accident Unit of the Depok Metro Police has not yet provided information on these accident-prone areas[12]. To address this issue, data mining can play a vital role in generating clear information to support decision-making processes[13]. This study employs data mining to cluster accident data using the K-Means algorithm, a smart clustering method[14]. K-Means is a data mining technique that falls under the category of unsupervised learning, allowing automatic data clustering based on patterns, features, or data distribution[15].

Thus, this study aims to analyze traffic accident data in Depok City using the K-Means Clustering method[15]. The traffic accident data, obtained from the Traffic Accident Unit of the Depok Metro Police, includes monthly recapitulation per district with details such as the district name, month, year, number of incidents, fatalities, serious injuries, minor injuries, and the total number of victims (fatalities, serious injuries, and minor injuries). The goal is to cluster this traffic accident data at the district level, providing valuable information to support decision-making processes related to accident-prone areas in Depok City.

2. METHOD

In this section, the research describe the methodological approach employed in this study to analyze traffic accident data in Depok City. The overall process is depicted in Figure 1, which outlines the sequence of steps taken, from data collection to visualization of the results. Each step is designed to ensure that the data is accurately processed and analyzed, leading to meaningful insights that can support decision-making for traffic safety improvements. The following subsections detail the procedures involved, including data collection, preprocessing, data mining using the K-Means clustering algorithm, evaluation of cluster quality, and the visualization of the identified accident-prone areas[15].

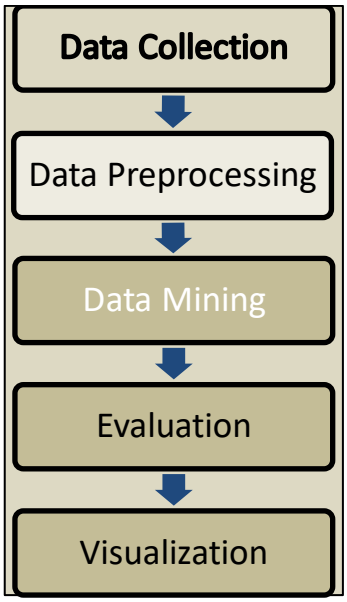

Figure 1. Research Design

1. Data Collection

This study uses secondary data. Secondary data refers to information obtained indirectly through intermediaries, such as files, documents, or records already archived. The secondary data utilized in this research comprises traffic accident data per district from 2020 to 2022, obtained from the Traffic Accident Unit of the Depok Metro Police. The data is in the form of monthly recapitulations for each year.

2. Data Preprocessing

In this study, data preprocessing involves two main steps: data integration and data selection[16]. Data integration combines data from different sources into a coherent dataset, ensuring consistency and completeness. Data selection involves choosing relevant attributes from the dataset[17], specifically district name, month, year, number of incidents, fatalities, serious injuries, minor injuries, and the total number of victims.

3. Data Mining

The study applies unsupervised data mining using the K-Means clustering algorithm[18]. The clustering technique[19] is employed to group the data into three distinct clusters, representing varying levels of accident risk across different districts. The tool used for this process is Rapid Miner, which facilitates the application of the K-Means algorithm and aids in generating the clusters.

4. Evaluation

The quality of the clusters formed by the K-Means algorithm is evaluated using the Davies-Bouldin Index (DBI)[20][21]. This index measures the average similarity ratio of each cluster, providing insight into the compactness and separation of the clusters. Following the evaluation, the results of the clusters are interpreted to understand the patterns and trends within the accident data.

5. Visualization

The final step involves visualizing the clusters by plotting them on maps using GIS tools[22]. This visualization helps in creating a clear, visual representation of the accident-prone areas across the districts in Depok City, making the information more accessible and actionable for decision-making purposes.

3. RESULTS AND DISCUSSION

3.1. Data Collection

Equations should be placed at the center of the line and provided consecutively with equation numbers in parentheses flushed to the right margin, as in (1). The use of Microsoft Equation Editor or MathType is preferred.

$$E_v - E = \frac{h}{2.m}\ (k_x^2 + k_y^2) \quad (1)$$

All symbols that have been used in the equations should be defined in the following text.

3.2. Data Preprocessing

3.2.1. Data Integration

In this study, the Data Integration process was carried out to combine accident data per district from 2020 to 2022 into a single dataset. Figure 2 shows the results of the Data Integration process (the merging) of accident data per district from 2020 to 2022.

Table 1. Results of the Data Integration Process (Merging) Accident Data by Sub-District from 2020 to 2022

No	Area	Mounth	Year	Number of Accident	Passed Away	Serious Injury	Minor Injury	Total
1	Beji	January	2020	5	0	3	5	8
2	Bojonggede	January	2020	3	0	2	2	4
3	Bojongsari	January	2020	2	0	3	2	5
4	Cilodong	January	2020	5	0	4	1	5
5	Cimanggis	January	2020	4	0	4	1	4
6	Cinere	January	2020	4	0	3	3	6
7	Cipayung	January	2020	3	0	1	2	3
8	Limo	January	2020	4	0	3	2	5
9	Pancoran Mas	January	2020	5	0	4	2	6
10	Sawangan	January	2020	4	0	1	3	4
11	Sukmajaya	January	2020	4	0	1	3	4
12	Tajur Halang	January	2020	1	0	1	0	1
…	…	…	…	…	…	…	…	…
457	Bojonggede	December	2022	3	0	0	3	3
458	Bojongsari	December	2022	3	0	2	1	3
459	Cilodong	December	2022	6	0	4	8	12
460	Cimanggis	December	2022	10	0	5	4	9
461	Cinere	December	2022	4	0	2	2	4
462	Cipayung	December	2022	4	0	1	3	6
463	Limo	December	2022	3	0	1	3	4
464	Pancoran Mas	December	2022	10	0	7	5	12
465	Sawangan	December	2022	4	0	6	0	6
466	Sukmajaya	December	2022	6	1	1	5	8
467	Tajur Halang	December	2022	0	0	0	0	0
468	Tapos	December	2022	3	0	2	2	4

3.2.2. Data Selection

Data Selection is the process of reducing the amount of data by decreasing the number of data points, variables, or attributes present in a dataset. The goal of this process is to select relevant and significant variables or attributes that support the data analysis process. The variables or attributes used in the data mining process for accident data per district include four: the number of incidents, fatalities, serious injuries, and minor injuries.

Figure 3 shows the results of the Data Selection process conducted on the previously integrated accident data per district.

Tabel 2. results of the Data Selection process conducted on the previously integrated accident data per district.

No	Area	Number of Accident	Passed Away	Serious Injury	Minor Injury
1	Beji	5	0	3	5
2	Bojonggede	3	0	2	2
3	Bojongsari	2	0	3	2
4	Cilodong	5	0	4	1
5	Cimanggis	4	0	4	1
6	Cinere	4	0	3	3
7	Cipayung	3	0	1	2
8	Limo	4	0	3	2
9	Pancoran Mas	5	0	4	2
10	Sawangan	4	0	1	3
11	Sukmajaya	4	0	1	3
12	Tajur Halang	1	0	1	0
13	Tapos	9	0	5	3
…	…	…	…	…	…
456	Beji	7	0	4	4
457	Bojonggede	3	0	0	3
458	Bojongsari	3	0	2	1
459	Cilodong	6	0	4	8
460	Cimanggis	10	0	5	4
461	Cinere	4	0	2	2
462	Cipayung	4	0	1	3
463	Limo	3	0	1	3
464	Pancoran Mas	10	0	7	5
465	Sawangan	4	0	6	0
466	Sukmajaya	6	1	1	5
467	Tajur Halang	0	0	0	0
468	Tapos	3	0	2	2

3.3. Data Mining

In this research, the K-Means method, an unsupervised learning approach, is used to automatically cluster data based on patterns and characteristics[23]. The process involves importing accident data per district into RapidMiner software, where relevant attributes, such as the number of incidents, fatalities, serious injuries, and minor injuries, are selected for analysis. Data selection reduces the dataset to focus on significant variables that support the clustering process[24].

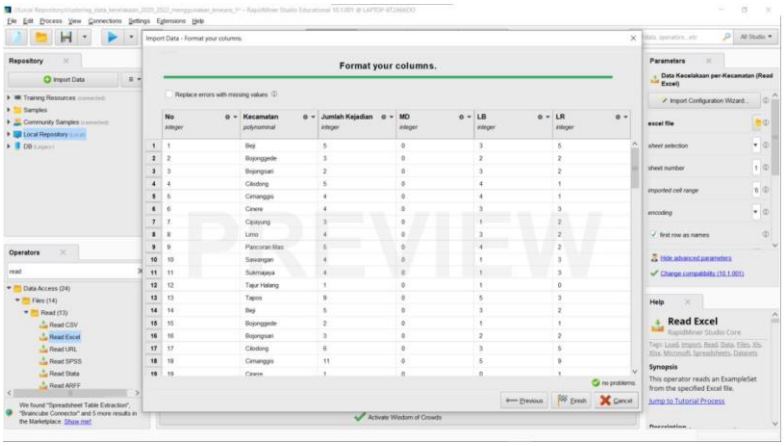

Figure 2 Importing Accident Data

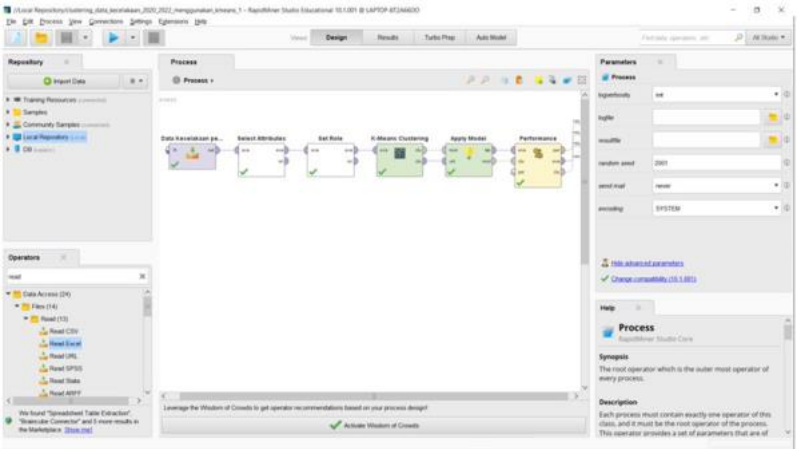

Figure 3. Clustering model

After the data import process has been completed, the next step is to build a clustering model using the K-Means algorithm, evaluate it, and display the results. In Figure 3 Clustering Model Using the K-Means Algorithm Based on the clustering model that has been developed, the author uses several operators as shown in Figure 3. Here is an explanation of each of those operators. Select Attributes is an operator for selecting a subset of attributes that are used and removing attributes that are not used[25]. In this process, the attribute that will be removed is the No attribute, which is included in the accident dataset per sub-district that is imported into the RapidMiner software.

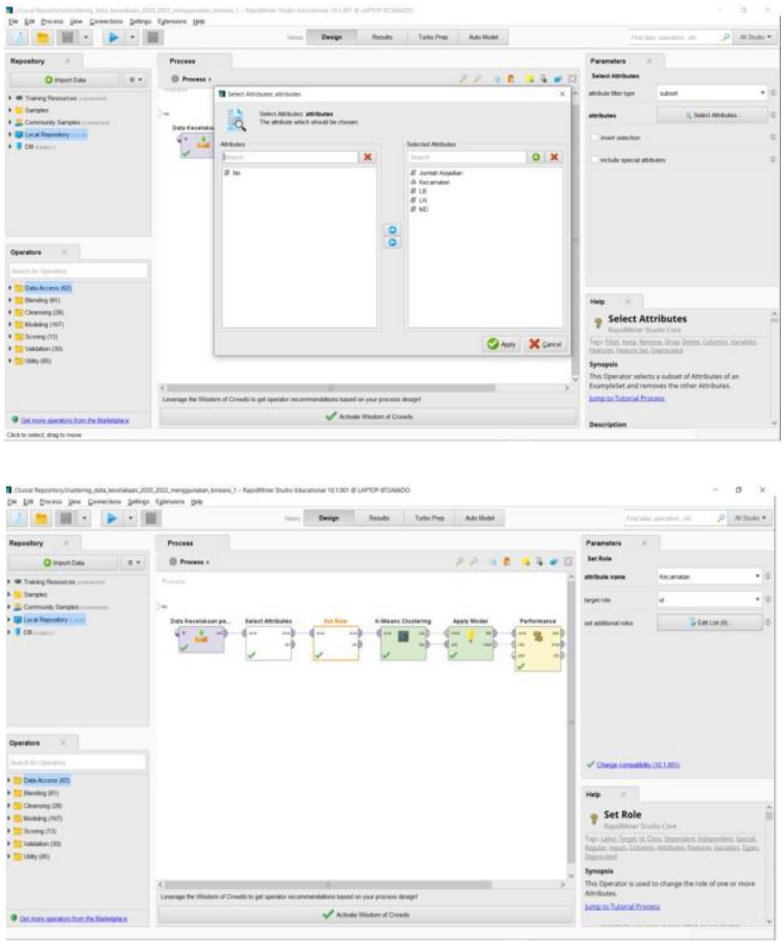

Figure 4. Operator *Set Role*

The Set Role operator changes the role of attributes, specifically transforming the sub-district attribute into an ID to exclude it from the clustering analysis[26]. The K-Means Clustering operator then groups the data into three categories—high, medium, or low accident frequency—based on patterns and distribution within the data.

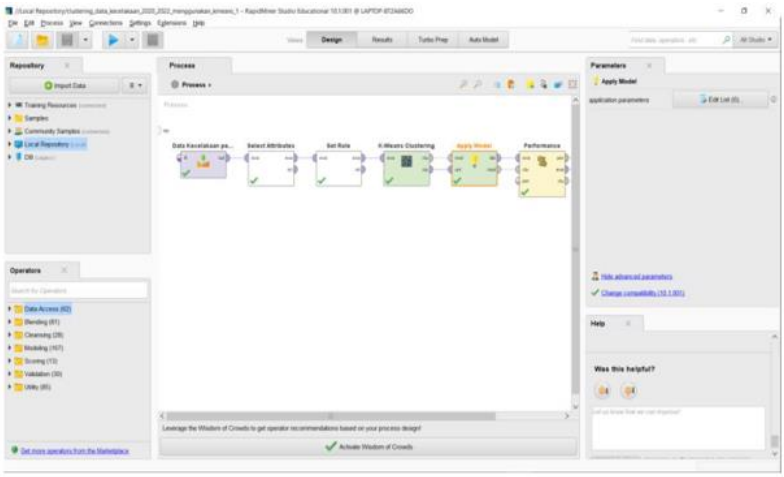

Figure 5. Operator *Clustering*

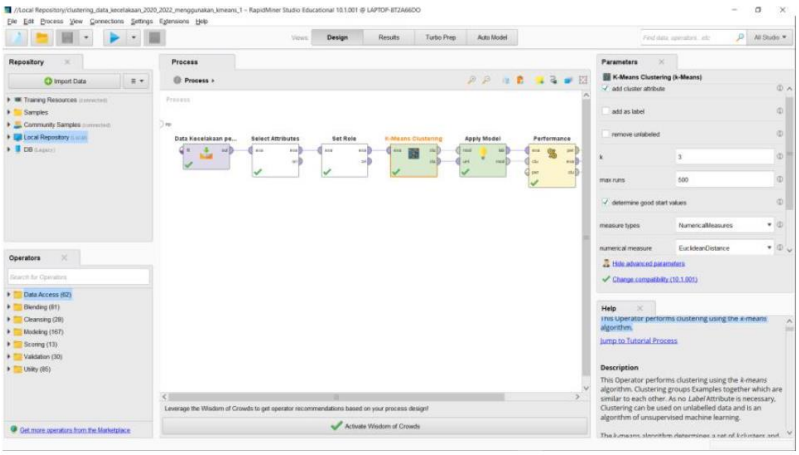

Figure 6. Operator *Apply Model*

The Apply Model is an operator for applying the model that has been created to the data being used. In this process, the clustering model that has been developed will be applied to the data in order to obtain information from the results of the clustering that has been performed. Performance is an operator used to evaluate the effectiveness of the clustering results that have been carried out.

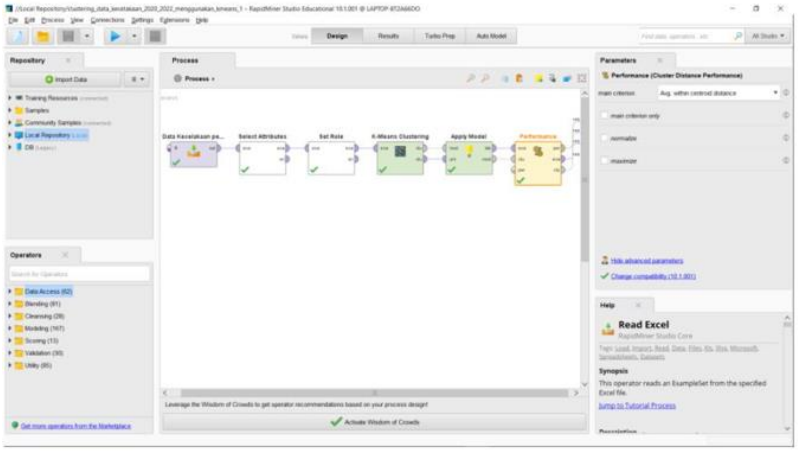

Figure 7. Operator *Performance*

Figure 8. Result of *Cluster Model*

After the clustering model development process has been completed, the next step is to run the model to see the results of the clustering that has been built. Figure 9 shows the output of the clustering based on the model that has been developed. Figure 10 Results of the Cluster Model Based on the results of the Cluster Model, a total of 468 items are divided into 3 groups or clusters: cluster_0 consists of 184 items, cluster_1 consists of 209 items, and cluster_2 consists of 75 items. Figure 4.10 shows the display of the centroid results for each cluster generated from the Cluster Model.

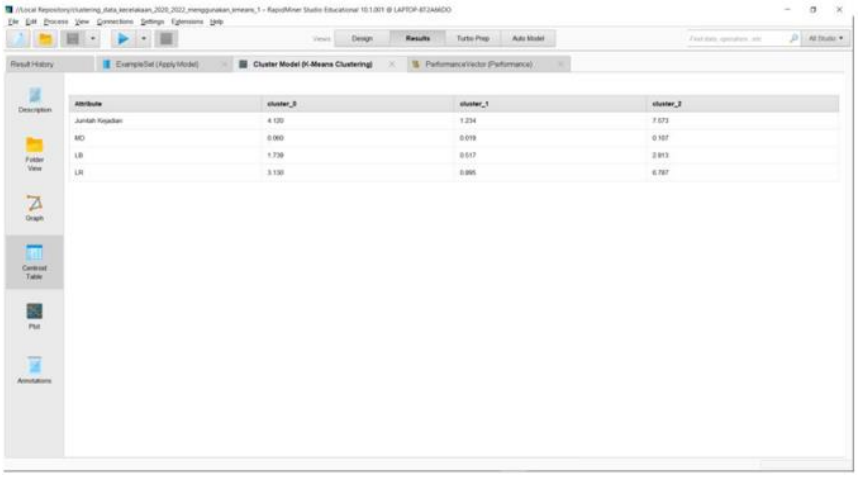

Figure 11. Results of the Centroid Values

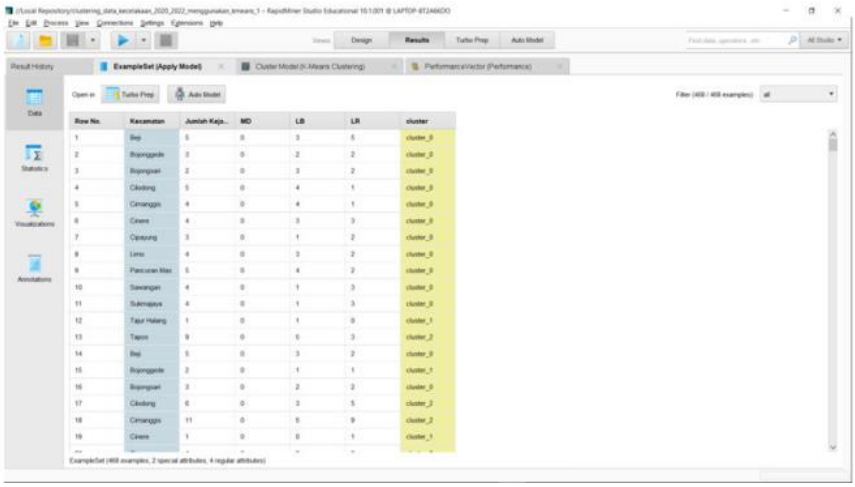

Figure 12. Results of Accident Data

Figure 11 shows the display of accident data per sub-district that has been clustered into each of the established clusters. Figure 12 Results of Accident Data per Sub-District that have been Clustered. Based on the author's observations of the data distribution in each cluster, the author concludes that cluster_0 is a cluster with characteristics or patterns of medium frequency accident data, cluster_1 is a cluster with characteristics or patterns of low frequency accident data, and cluster_2 is a cluster with characteristics or patterns of high frequency accident data. In figure 12, it shows the distribution of accident data per sub-district using scatter diagrams in each cluster.

Figure 13 Distribution of Data by Subdistrict

Figure 14 Based on Accident Frequency

Figure 13 Scatter Diagram of Data Distribution by Subdistrict in Each Cluster As an effort to facilitate the understanding of the clustering results, the author created a new attribute called Accident Frequency, which represents the characteristics of each cluster using the Generate Attributes operator. Generate Attributes is an operator used to create new attributes defined by the user using mathematical expressions. In figure 14 it shows the distribution of accident data per sub-district using a scatter diagram based on the frequency of accidents.

3.4. Evaluation

After the clustering process is complete, the next step is to evaluate the results of the clustering that has been conducted. This process aims to evaluate the performance or validity of the clustering results that have been conducted. The evaluation testing method used in this research is the internal evaluation testing method, specifically the Davies Bouldin Index. The Davies Bouldin Index (DBI) is one of the evaluation methods used to measure the validity of clusters in a clustering method. This method falls into the category of intrinsic methods, where the quality of the cluster can be assessed based on cohesion and separation. Cohesion is defined as the degree of closeness among data within a cluster. Meanwhile, separation is defined as the distance between data in different clusters. In image 15 the results of the DBI evaluation produced by the RapidMiner software are shown.

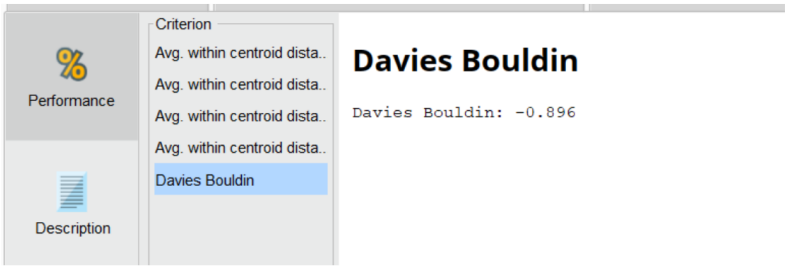

Figure 15. Results of the Davies Bouldin Index Evaluation

The lower the value produced (closer to zero) in the evaluation calculation using the Davies Bouldin Index, the number of clusters formed is considered optimal. Based on the Davies Bouldin Index evaluation results generated by the RapidMiner program, which is 0.896, this indicates that the grouping is in a fairly good category.

3.5. Visualization

Results of the information generated from this clustering process is about the characteristics of accidents in each district regarding the distribution of accident frequency over the past three years In figure 16, there is a bar chart created from the grouping results that have been conducted.

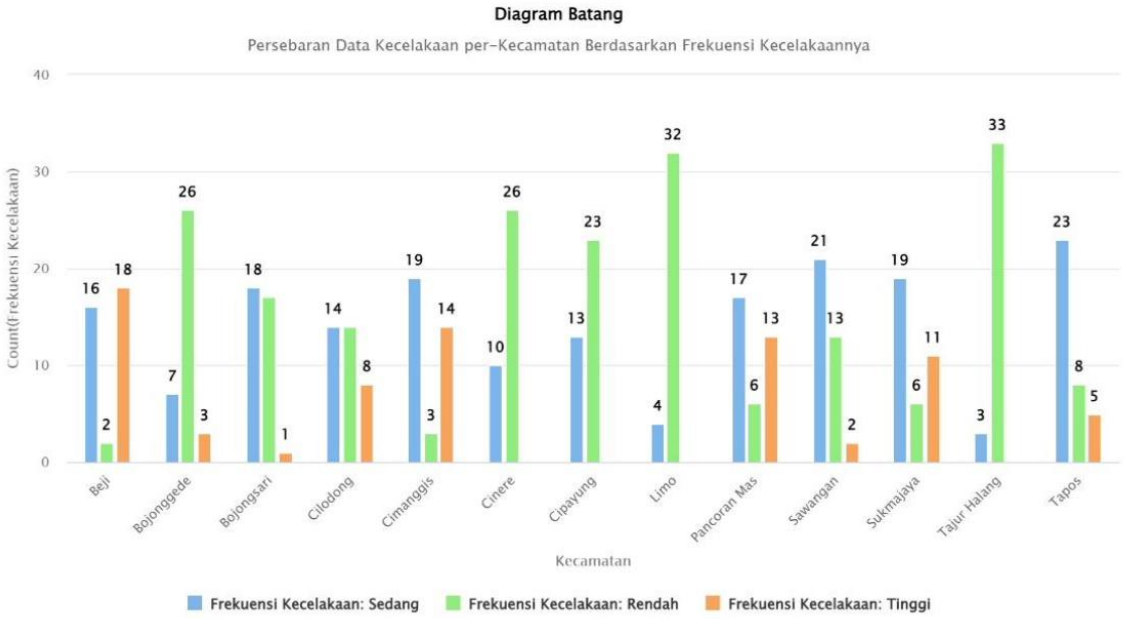

Figure 16 Bar Chart of Accident Data Distribution

Bar Chart Description
Blue = Medium Accident Frequency.
Green = Low Accident Frequency.
Orange = High Accident Frequency.

In the displayed bar chart, information regarding the distribution of accident frequency in each district can be seen. This information is used by the author to help conclude the level of accidents in each district based on the distribution of accident frequency. However, the information presented in the bar chart does not yet provide a detailed depiction of the distribution of accident frequency in each district. Therefore, the author created a pie chart with the aim of illustrating in more detail the distribution of accident frequency in each of those districts.

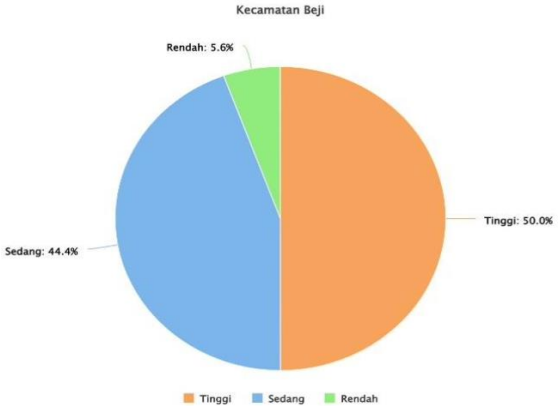

Figure 17. Frequency of accident in Beji

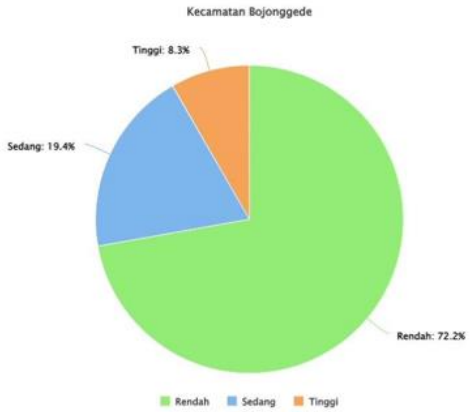

Figure 18. Frequency of accident in Bojonggede

The research applies the K-Means Clustering algorithm to identify traffic accident-prone areas in Depok City over the past three years (2020-2022). The results show that Beji, Cimanggis, Pancoran Mas, and Sukmajaya districts have high accident rates, while Bojongsari, Cilodong, Sawangan, and Tapos have medium rates. Bojonggede, Cinere, Cipayung, Limo, and Tajur Halang districts fall into the low accident rate category. The frequency distribution of accidents in each district was visualized using demographic maps created with QGIS software, highlighting the varying levels of accident risk across the city.

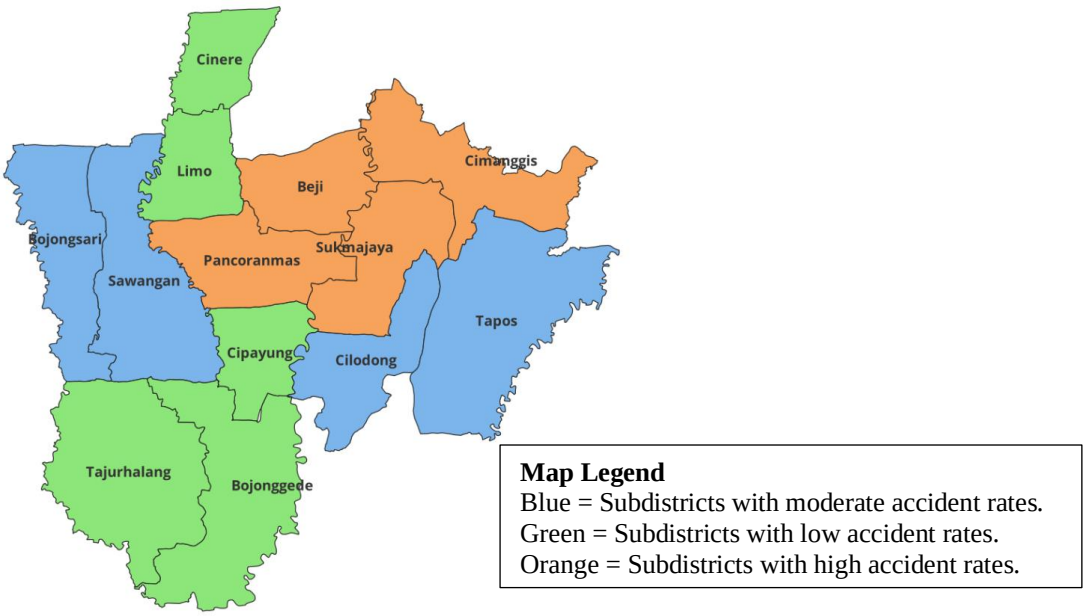

Figure 19. Demographic Map of Accident Rates in the City of Depok

Based on the results of the demographic map visualization shown in figure 19, the author concludes that out of the thirteen districts in Depok City, there are four districts identified as areas prone to traffic accidents. The four sub-districts are Beji Sub-district, Cimanggis Sub-district, Pancoran Mas Sub-district, and Sukmajaya Sub-district. This is based on the information that these four districts have had a high rate of accidents over the past three years.

4. CONCLUSION

Based on the research conducted, it can be concluded that the clustering method using the K-Means algorithm is effectively applicable in supporting decision-making processes regarding traffic accident-prone areas in Depok City over the past three years (2020 to 2022). The clustering results provide valuable information about the distribution of accident frequencies in each district. This information has been instrumental in identifying the level of accident risk in each district, based on the frequency distribution of accidents. The study concludes that out of the thirteen districts in Depok City, four districts—Beji, Cimanggis, Pancoran Mas, and Sukmajaya—have been identified as high-risk areas for traffic accidents. This conclusion is drawn from the observation that these four districts have exhibited consistently high accident rates over the past three years.

The findings align with the expectations outlined in the introduction, demonstrating that the application of K-Means clustering can indeed yield insights that are critical for improving road safety and guiding interventions in accident-prone areas. The results of this study suggest a promising direction for future research and practical applications. The methodology can be further developed by incorporating additional variables or employing more advanced clustering techniques to enhance the precision of the risk assessment. Furthermore, these findings can be leveraged by local authorities to implement targeted traffic safety measures, ultimately reducing the incidence of traffic accidents in Depok City.

REFERENCES

[1] F. Fatmah, V. P. Dewi, and Y. Priotomo, "Developing age-friendly city readiness: A case study from Depok city, Indonesia," *SAGE Open Med.*, vol. 7, p. 2050312119852510, 2019.

[2] M. Hafizha, "Preferensi Masyarakat Dalam Memilih Apartemen Dengan Menggunakan Metode Fuzzy Analytical Hierarchy Process (F-Ahp) Di Kota Depok Population Preferences In Choosing Apartment Model Using Fuzzy Analytical Hierarchy Process (F-Ahp) Method In Depok City," 2020.

[3] E. Rustiadi, A. E. Pravitasari, Y. Setiawan, S. P. Mulya, D. O. Pribadi, and N. Tsutsumida, "Impact of continuous Jakarta megacity urban expansion on the formation of the Jakarta-Bandung conurbation over the rice farm region," *Cities*, vol. 111, p. 103000, 2021.

[4] E. Macioszek and A. Granà, "The analysis of the factors influencing the severity of bicyclist injury in bicyclist-vehicle crashes," *Sustainability*, vol. 14, no. 1, p. 215, 2021.

[5] O. Tengilimoglu, O. Carsten, and Z. Wadud, "Implications of automated vehicles for physical road environment: A comprehensive review," *Transp. Res. part E Logist. Transp. Rev.*, vol. 169, p. 102989, 2023.

[6] M. R. F. Amrozi and R. P. Isheka, "Optimizing the functional performance of road network using vulnerability assessment to cope with unforeseen road incidents," in *Journal of the civil engineering forum*, 2022, vol. 8, no. 1, pp. 67–80.

[7] M. Isradi, "The impact of Covid-19 on areas prone to traffic accidents in Depok City: Margonda Raya Road case study," in *Journal of World Conference (JWC)*, 2021, vol. 3, no. 2, pp. 241–251.

[8] A. Supriadi and T. Oswari, "Analysis of Geographical Information System (GIS) design aplication in the Fire Department of Depok City," *Tech. Soc. Sci. J.*, vol. 8, p. 1, 2020.

[9] M. Thibenda, D. M. P. Wedagama, and D. Dissanayake, "Drivers' attitudes to road safety in the South East Asian cities of Jakarta and Hanoi: Socio-economic and demographic characterisation by Multiple Correspondence Analysis," *Saf. Sci.*, vol. 155, p. 105869, 2022.

[10] T. Tjahjono, B. Swantika, A. Kusuma, R. Purnomo, and G. H. Tambun, "Determinant contributing variables to severity levels of pedestrian crossed the road crashes in three cities in Indonesia," *Traffic Inj. Prev.*, vol. 22, no. 4, pp. 318–323, 2021.

[11] W. H. Organization, *Pedestrian safety: a road safety manual for decision-makers and practitioners*. World Health Organization, 2023.

[12] A. Z. Siregar, M. Awaluddin, and Y. Wahyuddin, "Identification of Traffic Accidents Vulnerability Level Using Kernel Density And K-Medoids Methods (Case Study: Depok and Kalasan Districts, Sleman Regency)," *J. Ilm. Geomatika*, vol. 3, no. 1, pp. 23–35, 2023.

[13] A. Sulhi, "Data mining technology used in an Internet of Things-based decision support system for information processing intelligent manufacturing," *Int. J. Informatics Inf. Syst.*, vol. 4, no. 3, pp. 168–

179, 2021.
[14] G. Zhang, Y. Li, and X. Deng, "K-means clustering-based electrical equipment identification for smart building application," *Information*, vol. 11, no. 1, p. 27, 2020.
[15] B. Lund and J. Ma, "A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering," *Perform. Meas. Metrics*, vol. 22, no. 3, pp. 161–173, 2021.
[16] C. Wu, F. Zhou, J. Ren, X. Li, Y. Jiang, and S. Ma, "A selective review of multi-level omics data integration using variable selection," *High-throughput*, vol. 8, no. 1, p. 4, 2019.
[17] N. Naheed, M. Shaheen, S. A. Khan, M. Alawairdhi, and M. A. Khan, "Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review," *Comput. Model. Eng. Sci.*, vol. 125, no. 1, pp. 314–344, 2020.
[18] A. Zimmermann, "Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 2, p. e1330, 2020.
[19] R. Rahim, J. T. Santoso, S. Jumini, G. Bhawika, D. Susilo, and D. Wibowo, "Unsupervised data mining technique for clustering library in Indonesia," *Libr. Philos. Pract.*, vol. 4866, 2021.
[20] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering evaluation by davies-bouldin index (dbi) in cereal data using k-means," in *2020 Fourth international conference on computing methodologies and communication (ICCMC)*, 2020, pp. 306–310.
[21] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, pp. 178–199, 2023.
[22] L. Zappia and A. Oshlack, "Clustering trees: a visualization for evaluating clusterings at multiple resolutions," *Gigascience*, vol. 7, no. 7, p. giy083, 2018.
[23] L. Zhao *et al.*, "K-means cluster analysis of characteristic patterns of allergen in different ages: Real life study," *Clin. Transl. Allergy*, vol. 13, no. 7, p. e12281, 2023.
[24] M. Rezaei, I. Cribben, and M. Samorani, "A clustering-based feature selection method for automatically generated relational attributes," *Ann. Oper. Res.*, vol. 303, no. 1, pp. 233–263, 2021.
[25] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020.
[26] A. Maghawry, R. Hodhod, Y. Omar, and M. Kholief, "An Approach to Optimize Multi-objective Problems Using Hybrid Genetic Algorithms Supported by Initial Centroid Selection Optimization Enhanced K-Means Based Selection Operator," in *Artificial Intelligence in Intelligent Systems: Proceedings of 10th Computer Science On-line Conference 2021, Vol. 2*, 2021, pp. 64–87.