**JURNAL TEKNOLOGI DAN OPEN SOURCE**
Vol. 5, No. 1, June 2022, pp. 81~88
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 36/E/KPT/2019
DOI: 10.36378/jtos. v5i1.2427

# Implementation of K-Means Clustering Algorithm for Grouping Traffic Violation Levels in Siak

**Bias Arbi Fauzan[1], M Jamaris[2], Junadhi[3], Hadi Asnal[4]**
[1,2,3,4]Teknik Informatika, STMIK Amik Riau, Pekanbaru, Riau, Indonesia

**Article Info**

***Article history:***

Received 05 15, 2022
Revised 06 20, 2022
Accepted 06 24,2022

***Keywords:***

K-Means Clustering
Violation
Grouping
Traffic

**ABSTRACT**

Traffic offences often occur in different regions, ranging from mild to moderate to severe. The categories of offences include not carrying a Driver's Licence, stnk (Vehicle Number Certificate) or stck (Vehicle Trial Certificate) is invalid, not wearing a seat belt, not turning on headlights during the day and under certain conditions, disobeying traffic signs, disobeying traffic signals. Moderate offences include not having a Driver's Licence, not concentrating while driving and breaking the door of the drawbar. Serious violations include deviating from other vehicles on the road, damaging and interfering with road functions, not insuring one's own responsibility and not insuring staff and passengers. In this study, the K-Means algorithm was used with the aim of obtaining information on data groups of traffic violations based on the time of the incident so that the cause of the traffic violations that occurred in Siak City is known. Based on the validation with Davies Bouldin Index metric, 4 clusters were identified which can group the data well. The PerformanceVector results from the assessment of the clusters resulted in 4 clusters with a value of 0.134. Cluster 1 with the most data violations amounting to 74 violations occurred at night, Cluster 2 with the most violations amounting to 16 violations occurred during the day, Cluster 3 with the most violations amounting to 6 violations occurred in the afternoon and Cluster 4 with the most violations amounting to 113 violations occurred in the morning.

***Corresponding Author:***

Bias Arbi Fauzan
Teknik Informatika
STMIK AMIK Riau
Penaknbaru, Riau
Email: biasarbi2022@gmail.com

## 1. Introduction

The number of vehicles is increasing rapidly, according to data from the Central Statistics Agency (BPS), which shows that the stock of all vehicles in Indonesia will exceed 133 million units in 2019. The number of vehicles, broken down by type, includes 15,592,419 passenger cars, 231,569 buses, 5,021,888 freight cars and 112,771,136 motorbikes. The cumulative increase in traffic, both two-wheeled and four-wheeled, leads to a traffic situation that is increasingly congested and uncontrollable, and indirectly increases the risk of growing traffic problems, as the increase in traffic is not proportional to the number of road widenings. From the perspective of social psychology, the problem makes drivers look for shortcuts or the fastest roads, and it certainly leads to traffic violations when no one is watching.

Traffic violations often occur in different regions, ranging from mild to moderate to severe. The categories of offences include not carrying a Driver's Licence, stnk (Vehicle Number Certificate) or stck (Vehicle Trial Certificate) is invalid, not wearing a seat belt, not turning on headlights during the day and under certain conditions, disobeying traffic signs, disobeying traffic signals. Moderate offences include not having a Driver's Licence, not concentrating while driving and breaking the door of the drawbar. Serious offences include deviating from other vehicles on the road, damaging and interfering with road functions, failing to insure oneself and failing to insure staff and passengers.

Traffic violations are cultivated in the community, including in the Siak Regency area. Every day, the number of road users who do not obey traffic rules can increase the number of road accidents and traffic violations in the Siak Regency region, so people do not understand the order of the road.

The application of K-Means clustering for accident data analysis has been done by Iswari (2015), Fajar (2015) and Rahmat et al (2017). Iswari (2015) uses the K-Means algorithm for mapping accident-prone areas in Sleman, Yogyakarta Special Region. Rahmat et al. (2017) used K-Means clustering to analyse the frequency of accident rates at each location with the potential for accident occurrence in Kendari city. Iswari and Rahmat classified the road accident data into clusters based on accident prone areas. Fajar (2015) also used K-Means in his study to classify the accident data in Semarang into several clusters of accident rate categories based on the age of the accident victims.

Based on the existing problems, the clustering method is used to group the traffic violation data using the K-Means algorithm, where the expected result is information about the traffic violation data cluster based on the time of the incident, so that the cause of the violation that occurred in Siak regency is known.

## 2. Reseach Methodology

The researchers used a clustering method using the K-Means algorithm to analyse data on traffic violations in the Siak Regency area. So that the research can be conducted in a more targeted manner. Here is the methodology of the research conducted by the researcher:

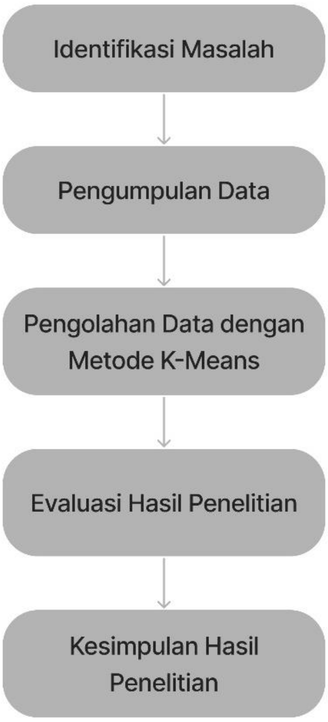
Identifikasi Masalah
Pengumpulan Data
Pengolahan Data dengan Metode K-Means
Evaluasi Hasil Penelitian
Kesimpulan Hasil Penelitian

Figure 1. Steps of Research

The steps carried out in the stages of the study can be described as follows:

1. Problem Identification

   The identification of problems in this study is to analyze the incidence of traffic customers in the Siak Regency area using the K-Means method.
2. Data Collection

   The data collected was sourced from the Siak District Attorney's Office in 2019.

3. Data Processing
   Data processing in this study used a clustering method with a K-means algorithm on traffic violation data in the Siak Regency area. So that the results of this study can be useful for evaluation and in order to minimize the incidence of traffic violations.
4. Evaluation of Research Results
   At this step, the author tested the results of the study using RapidMiner software in connecting the database to be tested. So that the information generated by the data mining process can be displayed in a form that is easy to understand for research.
5. Conclusion of the Research Results
   At this step, the completed research will be given conclusions from the existing problems so that they can be used by the authorities.

The method used in this study was clustering with the K-Means algorithm. Clustering is one of the analytical techniques in data mining that groups data based on similar features. By the similarity of the features.

Meanwhile, K-Means is one of the algorithms or methods of non-hierarchical cluster analysis that attempts to divide existing objects into one or more clusters or groups of objects based on their characteristics and determine their closest points. Figure 2 shows a K-Means flowchart.

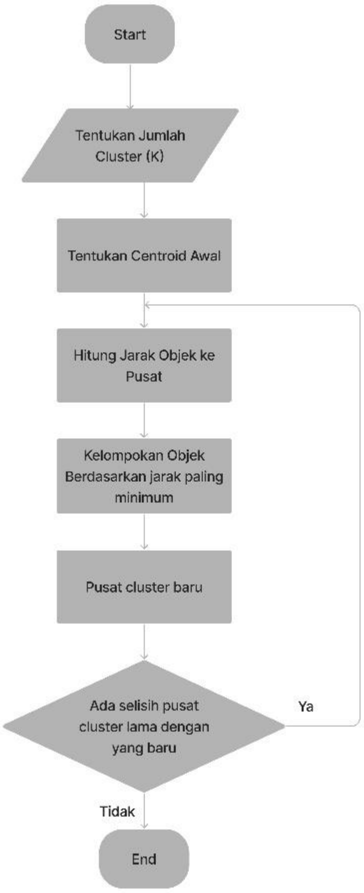
Start
Tentukan Jumlah Cluster (K)
Tentukan Centroid Awal
Hitung Jarak Objek ke Pusat
Kelompokan Objek Berdasarkan jarak paling minimum
Pusat cluster baru
Ada selisih pusat cluster lama dengan yang baru
Ya
Tidak
End

Figure 2. Flowchart K-Means Clustering

Here are the steps of the K-means algorithm process:

1. randomly determine the cluster as the centre (centroid).
2. using the Euclidean distance formula, calculates the least distance of the data to the central point. The Euclidean formula is as follows:

$$d\,(i\,,j) = \sqrt{\textstyle\sum_{j}^{m} (Cij - Ckj)^2} \qquad (1)$$

where :
$xi$ = criteria data
$\mu j$ = Centroid at the Cluster to – j

3. Group the data in clusters with the smallest distances.

$$\text{Min} \textstyle\sum_{k=1}^{k} dik = \sqrt{\textstyle\sum_{j}^{m} (Cij - Ckj)^2} \qquad (2)$$

4. Update the value of the center point by performing the cluster average.

$$C_{kj} = \frac{\sum_{i=1}^{p} Xij}{p} \qquad (3)$$

5. Repeat steps 1, 2, 3 until all members of the cluster are no longer changed.
6. If the fifth step is already fulfilled, which is used as a parameter to determine the accuracy of the data, the last value of the cluster centre.

## 3. Result and Discussion

### 3.1 Result

This step ensures that the selected data on victims of violations are suitable for processing. The original data obtained by the author amounted to 445 traffic violation data with a total of 9 attributes.

Table 1. Accident Data Attributes

| No | Attribut | Information |
|---|---|---|
| 1 | Location | Location of the violation incurred |
| 2 | Time | Time of the violation incurred |
| 3 | Age | Age of violator |
| 4 | Gender | Gender of the violator |
| 5 | Victim's Condition | Violator's Condition |
| 6 | Profession | The profession of an accident violation |
| 7 | Type of Vehicle | Circumstances involved in the violation |
| 8 | Way/Mode | Violator’s way when the violation incured |
| 9 | Material Losses | Material losses from violations incurred |

1. Data Cleanning

Data cleaning or data cleansing is the process of selecting data attributes to be used in research. Mode and material losses are then eliminated for data attributes that are not used.

2. Data Transformation

Data of a nominal nature such as crime location, time, age, sex, condition of the victim, occupation and vehicle involved must first be initialised in the form of numbers or numerical values. This initialisation can be done by sorting the numbers according to their frequency.

Table 2. Transformed Datasets

| No | Location | Age | Gender | Condition | Profession | Types of Vehicle |
|---|---|---|---|---|---|---|
| 1 | 5 | 2 | 2 | 1 | 2 | 5 |
| 2 | 5 | 2 | 2 | 1 | 1 | 5 |
| 3 | 2 | 3 | 1 | 1 | 2 | 9 |
| 4 | 2 | 3 | 2 | 2 | 3 | 9 |
| 5 | 2 | 3 | 3 | 1 | 1 | 2 |
| 6 | 1 | 4 | 1 | 2 | 1 | 4 |
| 7 | 1 | 4 | 3 | 1 | 1 | 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | 1 | 4 | 1 | 1 | 1 | 1 |
| 9 | 3 | 4 | 2 | 1 | 1 | 1 |
| 10 | 9 | 3 | 2 | 1 | 2 | 1 |
| ... | ... | ... | ... | ... | .... | ... |
| 208 | 5 | 3 | 2 | 1 | 3 | 1 |
| 209 | 2 | 4 | 3 | 1 | 1 | 6 |

The first step of the k-means algorithm is to determine the number of clusters. In this study, there are 4 clusters according to the formation of time groups for traffic violations, both morning, afternoon, evening and night. The initial cluster was randomly determined with the attributes of offence location, time, age, gender, victim's circumstances, occupation and type of vehicle involved.

Below is a dataset of traffic violations that has gone through the pre-processing steps, including the data cleaning step, the nominal type data initialisation step and the final step of transforming all data that has gone through the initialisation step.

Table 3. Preprocessing Result Data

| No | Location | Age | Gender | Condition | Profession | Types of Vehicle |
|---|---|---|---|---|---|---|
| 1 | 5 | 2 | 2 | 1 | 2 | 5 |
| 2 | 5 | 2 | 2 | 1 | 1 | 5 |
| 3 | 2 | 3 | 1 | 1 | 2 | 9 |
| 4 | 2 | 3 | 2 | 2 | 3 | 9 |
| 5 | 2 | 3 | 3 | 1 | 1 | 2 |
| 6 | 1 | 4 | 1 | 2 | 1 | 4 |
| 7 | 1 | 4 | 3 | 1 | 1 | 2 |
| 8 | 1 | 4 | 1 | 1 | 1 | 1 |
| 9 | 3 | 4 | 2 | 1 | 1 | 1 |
| 10 | 9 | 3 | 2 | 1 | 2 | 1 |
| ... | ... | ... | ... | ... | .... | ... |
| 200 | 11 | 3 | 3 | 1 | 3 | 2 |
| 201 | 8 | 4 | 1 | 2 | 1 | 5 |
| 202 | 6 | 2 | 1 | 1 | 2 | 1 |
| 203 | 7 | 3 | 3 | 1 | 2 | 2 |
| 204 | 11 | 4 | 2 | 1 | 2 | 1 |
| 205 | 4 | 4 | 3 | 1 | 2 | 2 |
| 206 | 5 | 1 | 3 | 1 | 1 | 2 |
| 207 | 2 | 2 | 1 | 1 | 2 | 22 |
| 208 | 5 | 3 | 2 | 1 | 3 | 1 |
| 209 | 2 | 4 | 3 | 1 | 1 | 6 |

The next step is to determine the initial centre of the cluster (centroid), which is chosen at random. In this study it was chosen from the 38th, 9th, 4th and 57th dates.

Table 4. Initial Centroids

| Centroid | No | Location | Age | Gender | Condition | Profession | Types of Vehicle |
|---|---|---|---|---|---|---|---|
| Cluster 0 | 38 | 8 | 3 | 1 | 2 | 1 | 4 |
| Cluster 1 | 9 | 11 | 3 | 2 | 2 | 2 | 10 |
| Cluster 2 | 4 | 1 | 1 | 2 | 1 | 1 | 26 |
| Cluster 3 | 57 | 1 | 4 | 2 | 1 | 1 | 1 |

After determining the initial centroid, the next step is to calculate the distance of each data to the nearest centroid to determine the cluster that the data follows using the Euclidean distance formula. Here you can see the complete calculation of iteration 1.

Table 5. Calculation Results Using Euclidean's formula in the 1st iteration

| **Data to-i** | **Centroid Distance** | | | | **Nearest** | **Followed Cluster** |
|---|---|---|---|---|---|---|
| | **C0** | **C1** | **C2** | **C3** | | |
| 1 | **38,7** | 7,94 | 21,42 | 6,16 | 3,87 | C0 |
| 2 | **3,74** | **8,25** | **21,49** | **6,08** | **3,74** | **C0** |
| 3 | **8,00** | **9,17** | **17,20** | **8,31** | **8,00** | **C0** |
| 4 | 8,12 | 9,17 | 17,32 | 8,43 | 8,12 | C0 |
| 5 | 6,78 | 12,17 | 24,12 | 2,24 | 2,24 | C3 |
| 6 | 7,14 | 11,79 | 22,25 | 3,44 | 3,46 | C3 |
| 7 | 7,75 | 12,96 | 24,21 | 1,73 | 1,73 | C3 |
| 8 | 7,81 | 13,60 | 25,20 | 1,41 | 1,41 | C3 |
| 9 | 6,08 | 12,21 | 25,28 | 2,00 | 2,00 | C3 |
| 10 | 3,74 | 9,27 | 26,34 | 8,19 | 3,74 | C1 |
| ... | ... | ... | ... | ... | ... | ... |
| 200 | 4,69 | 8,25 | 26,19 | 10,34 | 4,69 | C0 |
| 201 | 1,41 | 6,16 | 22,41 | 8,19 | 1,41 | C0 |
| 202 | 4,00 | 10,49 | 25,57 | 5,57 | 4,00 | C0 |
| 203 | 3,32 | 9,11 | 24,88 | 6,32 | 3,32 | C0 |
| 204 | 4,80 | 9,11 | 27,11 | 10,10 | 4,80 | C0 |
| 205 | 5,20 | 10,82 | 24,43 | 3,46 | 3,46 | C3 |
| 206 | 4,80 | 10,34 | 24,35 | 5,29 | 4,80 | C0 |
| 207 | 19,08 | 15,10 | 4,47 | 21,19 | 4,47 | C2 |
| 208 | 5,00 | 10,91 | 25,48 | 4,69 | 4,69 | C3 |
| 209 | 6,86 | 10,05 | 20,27 | 5,29 | 5,29 | C3 |

In addition, the data group corresponds to the nearest cluster. From this data, a new centroid is determined based on the average results of each cluster.

### 3.2 Discussion

Based on the clustering process with the k-means algorithm using the Rapidminer application, the following information is obtained:

Table 4. Initial Centroid

| **Variable** | **No** | **Location** | **C0** | **C1** | **C2** | **C3** |
|---|---|---|---|---|---|---|
| Time | 1 | Night | 30 | 2 | 2 | 22 |
| | 2 | Evening | 17 | 3 | 2 | 23 |
| | 3 | Afternoon | 16 | 9 | 1 | 33 |
| | 4 | Morning | 11 | 2 | 1 | 35 |
| Location | 1 | Jl. Raya Rajapolah | - | 2 | 3 | 32 |
| | 2 | Jl. AH Nasution | - | 5 | 1 | 19 |
| | 3 | Jl. Raya Jamanis | - | - | - | 2 3 |
| | 4 | Jl. Raya Ciawi | - | 1 | 1 | 1 9 |
| | 5 | Jl. Raya Kadipaten | - | 1 | - | 20 |
| | 6 | Jl. Ir. H. Juanda | 1 6 | 1 | - | - |
| | 7 | Jl. Syekh Abdul Muhyi | 1 6 | - | - | - |
| | 8 | Jl. Raya Manonjaya | 1 1 | 1 | 1 | - |
| | 9 | Jl. Letjen Ibrahim Adjie | 1 0 | 2 | - | - |
| | 1 0 | Jl. Perintis | 1 1 | 1 | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Kemerdekaan | | | | |
| | 11 | Jl. Raya Cisayong | 10 | 2 | - | - |
| Age | 1 | 27 – 38 | 24 | 5 | 3 | 39 |
| | 2 | 15 – 26 | 28 | 8 | 3 | 44 |
| | 3 | 39 – 50 | 21 | - | - | 29 |
| | 4 | 51 – 62 | 1 | 1 | - | 1 |
| | 5 | 87 – 92 | - | 2 | - | - |
| Gender | 1 | Male | 58 | 11 | 6 | 79 |
| | 2 | Female | 16 | 5 | - | 34 |
| Condition | 1 | Minor Injuries | 55 | 8 | 5 | 78 |
| | 2 | Die | 13 | 5 | 1 | 28 |
| | 3 | Major Injuries | 6 | 3 | - | 7 |
| Profession | 1 | Enterpreneur | 22 | - | - | 27 |
| | 2 | Student of school | 9 | - | - | 21 |
| | 3 | Worker | 12 | - | - | 16 |
| | 4 | Housewife | 9 | - | - | 16 |
| | 5 | Pegawai Swasta | 8 | - | - | 24 |
| | 6 | Student of college | 11 | - | - | 4 |
| | 7 | Jobless | 1 | - | - | 3 |
| | 8 | Seller | 2 | - | - | 2 |
| | 9 | Civil servant | - | 4 | - | - |
| | 10 | Teacher | - | 4 | - | - |
| | 11 | Pensioner | - | 2 | - | - |
| | 12 | Farmer | - | 1 | - | - |
| | 13 | Nurse | - | 1 | - | - |
| | 14 | Rider | - | 1 | - | - |
| | 15 | Driver | - | 2 | - | - |
| | 17 | Employee DINAS PU | - | 1 | - | - |
| | 18 | Employee of BUMN | - | - | 1 | - |
| | 19 | Head of Village | - | - | 1 | - |
| | 22 | Mechanics | - | - | 1 | - |
| Types of Vehicle | 1 | R2 | 50 | 11 | 5 | 74 |
| | 2 | R4 | 23 | 5 | 1 | 38 |
| | 3 | R6 | 1 | - | - | 1 |

From the above description, it can be concluded that in cluster 0, violations often occur at night, one of the causes is insufficient light intensity or illumination. In cluster 1, many violations occur during the day because it is school time. In cluster 2, many violations occur in the afternoon, which is due to the fact that the afternoon is the time when people return from work. In Cluster 3, many violations also occur in the morning because morning activities begin for students, employees, and others.

## 4. Conclusion

Based on the research results on injury data grouping using the k-means algorithm, the following conclusions were drawn: Based on the validation of the Davies-Bouldin Index (DBI), with a total sample of 209 data formed into 4 clusters with an accuracy value of 1.201; From the grouping of traffic violation data based on the time of occurrence of violations, which is divided into 4 clusters, it produces information on cluster 0 with the number of data breaches as many as 15 violations that occur at night, cluster 1 with the number of violations as many as 26 violations that occur during the day, cluster 2 with the number of violations as many as 24 violations that occur in the afternoon, and cluster 3 with the number of violations as many as 34 violations that occur during the day morning; Based on the information about the time of violation, there can be a knowledge for the Laka Lantas unit of siak regency police by connecting with the place where the traffic violation occurred to carry out appropriate treatment to reduce the number of traffic violations in Siak.

## References

[1] Aprianti, W., & Permadi, J. (2018). K-Means Clustering Untuk Data Kecelakaan Lalu Lintas K-Means Clustering for Highway Traffic Accident Data in Pelaihari Sub District, 5(5), 613–620. https://doi.org/10.25126/jtiik2018551113

[2] Ayu, D., Wati, M., Puspitasari, D., & Purwaningsih, E. (2019). Metode Clustering Pada Model Algoritma K-Means Untuk Pemilihan Alat Kontrasepsi, 3(2), 129–138.

[3] Azzirrahman, M., Normelani, E., & Arisanty, D. (2015). Faktor Penyebab Terjadinya Kecelakaan Lalu Lintas pada Daerah Rawan Kecelakaan di Kecamatan Banjarmasin Tengah Kota Banjarmasin. Jurnal Pendidikan Geografi, 2(3), 20 37.

[4] Eko, W. A. (2016). Implementasi data mining dalam pengelompokan data peserta didik di sekolah untuk memprediksi calon penerima beasiswa dengan menggunakan algoritma k- means (studi kasus sman 16 bekasi), 21(3).

[5] Endra, F. (2017). Pengantar Metodologi Penelitian (Statiska Praktis). zifatama jawara.

[6] Iswari, L., & Ayu, E. G. (2015). Pemanfaatan Algoritma K-Means Untuk Pemetaan Hasil Klasterisasi Data Kecelakaan Lalu Lintas. Pemanfaatan Algoritma K-Means Untuk Pemetaan Hasil Klasterisasi Data Kecelakaan Lalu Lintas, 21(1), 1–13. https://doi.org/10.20885/teknoin.vol21.iss1.art7

[7] Purwaningsih, E. (2019). Laporan Akhir Penelitian PDY: Analisis Kecelakaan Berlalu Lintas Di Kota Jakarta Dengan Menggunakan Metode KMeans. Jakarta.

[8] Rahmat C.T. I, B., Gafar, A. A., Fajriani, N., Ramdani, U., Uyun, F. R., P, Y. P., & Ransi, N. (2017).IMPLEMETASI K-MEANS CLUSTERING PADA RAPIDMINER UNTUK ANALISIS DAERAH RAWAN KECELAKAAN. In Prosiding Seminar Nasional Riset Kuantitatif Terapan 2017 (pp.58–62). Kendari: Lembaga Pengembangan Sistem Informasi.

[9] Saragih, P. G. G. (2013). ANALISA KECELAKAAN LALU LINTAS DI KOTA PEMATANG SIANTAR. JURNAL TEKNIK SIPIL USU, 2(3). Retrieved from https://jurnal.usu.ac.id/index.php/jts/article /view/5676

[10] Saragih, R., & Sitompul, J. N. (2019). Perbandingan Data Mining Mengidentifikasi Pola Keterkaitan Variabel Kecelakaan Lalu Lintas Di Polresta Kota Medan. Journal Information System Development (ISD), 4(1), 39–45. Retrieved from https://ejournal.medan.uph.edu/index.php/i sd/article/view/281

[11] Vulandari, R. T. (2017). Data Mining Teori dan Aplikasi Rapidminer. Yogyakarta: Gaava Media.

[12] Kesuma, D. P. (2018, Desember 20). Polisi Akui Angka Kecelakaan Lalu Lintas Sepanjang 2018 Cenderung Naik. Diambil kembali dari Tribunnews.com: https://www.tribunnews.com

[13] Lestian , A. C., & Ahmad, Z. F. (2017). Implementasi Algoritma K-Means Pada Data pelanggaran Lalu Lintas Di Pengadilan Negri Purwodadi. Jurnal UDINUS.

[14] Mulyani, E. D., Agustin, S., & Surgawi, N. (2018). Implementasi Algoritma KMeans dan Fp-Growth Untuk Rekomendasi Bimbingan Belajar Berdasarkan Segmentasi Akademik Siswa. Informatic Technique) Journal, 160-173.

[15] Ramadhani , N., Rahman, A. F., & Riskiyati, D. (2017). Analisis Cluster Data Register Perkara lalu Lintas Menggunakan Algoritma K-Means. SESINDO 9.