K-Means Clustering for Road Traffic Accident Data in Samarinda City

Agnes Novalista Koban[1)], Eka Arriyanti[2)], dan **Pitrasacha Adytia[3)]**

[1,2,3]Teknik Informatika, STMIK Widya Cipta Dharma
1,2,3 Jl. M. Yamin No.25, Gn. Kelua, Samarinda, Kodepos 75123
E-mail: agnesnovaa20@gmail.com[1)], ekaashuri@gmail.com[2)], pitra@wicida.ac.id[3)]

ABSTRACT

Traffic accidents are one of the major issues affecting public safety, making it necessary to analyze the characteristics of accident causes as a basis for decisionmaking in prevention efforts. This study aims to implement the K-Means Clustering algorithm to classify road traffic accident data in Samarinda City based on environmental and human factors.
The research employed the Knowledge Discovery in Databases (KDD) methodology, which consists of data selection, data cleaning, data transformation, data mining, and evaluation. Categorical attributes were first converted into numerical values using the Attribute Value Weighting method and then summarized into two main variables: Environmental Factors (X), consisting of road geometry, road surface condition, road gradient, lighting condition, and weather, and Human Factors (Y), consisting of behavior, age, and gender. The clustering process was subsequently performed using the K-Means algorithm.
The results indicate that the optimal number of clusters is five, with each cluster exhibiting distinct characteristics. Cluster 1 is characterized by high environmental and high human factors. Cluster 2 is dominated by high environmental factors and moderate human factors. Cluster 3 is characterized by low environmental factors and high human factors. Cluster 4 exhibits moderate environmental factors and low human factors, while Cluster 5 is characterized by moderate environmental factors and high human factors. The clustering performance was evaluated using the Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI), yielding values of 0.5012, 0.6800, and 141.8924, respectively. These results indicate that the proposed model produced clusters with good quality. The findings of this study are expected to provide supporting information for the Samarinda Police Department (Polresta Samarinda) and other relevant agencies in determining priority locations for accident prevention and developing more effective road traffic accident prevention strategies.
***Keywords**: K-Means Clustering, Knowledge Discovery in Databases (KDD), traffic accidents, Elbow method, cluster evaluation.*

K-MEANS CLUSTERING UNTUK DATA KECELAKAAN LALU LINTAS JALAN RAYA DI KOTA SAMARINDA

ABSTRAK

Kecelakaan lalu lintas merupakan salah satu permasalahan yang berdampak terhadap keselamatan masyarakat sehingga diperlukan analisis untuk mengetahui karakteristik penyebab kecelakaan sebagai dasar pengambilan keputusan dalam upaya pencegahan. Penelitian ini bertujuan menerapkan algoritma K-Means Clustering untuk mengelompokkan data kecelakaan lalu lintas jalan raya di Kota Samarinda berdasarkan faktor lingkungan dan faktor manusia.
Metode penelitian menggunakan tahapan Knowledge Discovery in Database (KDD) yang meliputi *data selection*, *data cleaning*, *data transformation*, *data mining*, dan *evaluation*. Atribut kategorikal terlebih dahulu dikonversi menjadi data numerik menggunakan *Attribute Value Weighting*, kemudian diringkas menjadi dua variabel utama, yaitu Faktor Lingkungan (X) yang terdiri atas bentuk geometri jalan, kondisi permukaan jalan, kemiringan jalan, kondisi cahaya, dan cuaca, serta Faktor Manusia (Y) yang terdiri atas perilaku, usia, dan jenis kelamin. Selanjutnya dilakukan proses clustering menggunakan algoritma K-Means.
Hasil penelitian menunjukkan bahwa jumlah cluster optimal adalah lima cluster, dengan karakteristik yang berbeda pada setiap kelompok. Cluster 1 memiliki faktor lingkungan dan faktor manusia yang sama-sama tinggi, Cluster 2 didominasi oleh faktor lingkungan tinggi dan faktor manusia sedang, Cluster 3 menunjukkan faktor lingkungan rendah dan faktor manusia tinggi, Cluster 4 memiliki faktor lingkungan sedang dan faktor manusia rendah, sedangkan Cluster 5 menunjukkan faktor lingkungan sedang dan faktor manusia tinggi. Evaluasi hasil clustering menggunakan Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI) menghasilkan nilai masing-masing sebesar 0,5012, 0,6800, dan 141,8924, yang menunjukkan bahwa model mampu membentuk cluster dengan kualitas yang baik. Hasil penelitian ini diharapkan dapat menjadi informasi pendukung bagi Polresta Samarinda dan instansi terkait dalam menentukan prioritas penanganan lokasi rawan kecelakaan serta menyusun strategi pencegahan kecelakaan lalu lintas secara lebih efektif.

Kata Kunci: *K-Means Clustering*, *Knowledge Discovery in Database* (KDD), kecelakaan lalu lintas, metode elbow, evaluasi *cluster*.

1. PENDAHULUAN

Kota Samarinda sebagai ibu kota Provinsi Kalimantan Timur mengalami pertumbuhan penduduk dan mobilitas masyarakat yang semakin meningkat, yang turut mendorong bertambahnya jumlah kendaraan. Kondisi tersebut berpotensi meningkatkan risiko kecelakaan lalu lintas, terutama karena dipengaruhi oleh faktor seperti kepadatan kendaraan, kondisi jalan, infrastruktur, dan kondisi geografis.

Tingginya kecelakaan lalu lintas menyebabkan munculnya daerah rawan kecelakaan (black site) yang dapat mengurangi kenyamanan serta membahayakan keselamatan masyarakat. Oleh karena itu, diperlukan analisis terhadap data kecelakaan untuk mengetahui karakteristik dan pola penyebab kecelakaan serta mengidentifikasi ruas jalan yang memiliki tingkat kerawanan tertentu.

Penelitian ini menerapkan algoritma K-Means Clustering untuk mengelompokkan data kecelakaan lalu lintas berdasarkan faktor-faktor penyebab kecelakaan. K-Means dipilih karena sederhana, efisien, dan mudah diimplementasikan. Data yang digunakan berasal dari Polantas Polsek Samarinda Kota dan memiliki lebih dari 1.600 baris yang masih perlu diolah dan ditata. Selanjutnya, proses Knowledge Discovery in Database (KDD) digunakan untuk memperoleh pengetahuan atau *data insight* dari data kecelakaan tersebut sehingga dapat memberikan gambaran mengenai karakteristik dan kerentanan ruas jalan terhadap kecelakaan di Kota Samarinda.

2. RUANG LINGKUP

Dalam penelitian ini permasalahan mencakup:

1. Bagaimana proses analisis data untuk *K-Means Clustering* terhadap data kecelakaan lalu lintas jalan raya di kota Samarinda menggunakan pendekatan *Knowladge Discovery In Database* (KDD)
2. Variabel (X, Y) klasterisasi didasarkan pada analisis terhadap variabel-variabel dataset terkait penyebab kecelakaan di jalan raya, yaitu

Bentuk Geometri, Kondisi Permukaan Jalan, Kemiringan Jalan, Kondisi Cahaya, Cuaca, Usia, Perilaku, serta Jenis Kelamin Pengendara. Jumlah K dari *K-Means* ditentukan dengan *Elbow method.*

3. Dengan adanya hasil *clustering* data kecelakaan lalu lintas di kota Samarinda ini diharapkan dapat menjadi acuan kepada pemerintah kota Samarinda khususnya dinas pekerjaan umum (PU) dan dinas perhubungan untuk memperbaiki jalan rusak atau memperlebar jalan dan juga memasang lampu penerangan jalan umum (LPJU) serta menambah rambu lalu lintas agar meminimalisir terjadinya laka.
4. K-Means merupakan salah satu teknik dalam data mining yang digunakan untuk mengelompokkan objek berdasarkan kemiripan karakteristiknya (Silitonga , dkk., 2024). Sehingga objek dengan pola serupa berada dalam kelompok yang sama, sedangkan objek dengan ciri yang berbeda akan tergabung dalam kelompok lain (Izzuddin & Wijayanto, 2024).

Langkah-langkah dalam metode *K-Means* adalah sebagai berikut:

1. Tentukan nilai k sebagai jumlah klaster yang ingin dibentuk.
2. Bangkitkan k centroid (titik pusat *cluster*) awal.
3. Hitung jarak setiap data ke masing-masing centroid menggunakan rumus Manhattan Distance seperti pada persamaan berikut:

$$D(X,Y) = \sum |Xi - Yi|$$

Keterangan :
D(X,Y) = jarak objek antara Xi dan Yi
Xi = Koordinat dari objek Xi pada dimensi i
Yi = Koordinat dari objek Yi pada dimensi i

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroid*nya.
5. Tentukan centroid baru dengan cara menghitung nilai rata-rata dari data-data yang ada pada centroid yang sama dengan rumus seperti persamaan berikut. Ci=((∑x)/n)
6. Dimana n adalah banyaknya dokumen dalam *Cluster* i dan x adalah data yang akan dihitung.
7. Kembali ke langkah 3 jika posisi centroid barudengan centroid lama tidak sama.

5. *Clustering* adalah membagi data ke dalam grup-grup yang mempunyai objek dengan karakteristiknya sama. *Clustering* memegang peran penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan text mining, aplikasi basis data spesial, dananalisis web. (Relita, dkk., 2021: h. 45).Pada dasarnya metode pengelompokan ada 2 yakni *Hierarchical clustering* method dan Non *Hierarchical clustering* method. Metode hirarki digunakan jika jumlah kelompok tidak diketahui sebelumnya sedangkan non hirarki digunakan jika jumlah kelompok sudah diketahui dari sejumlah objek. Salah satu Algoritma yang termasuk dalam non hirarki adalah Algoritma *K-means*. (Relita, dkk., 2021: h. 46)

Berikut formula yang digunakan untuk menghitung jarak dengan euclidian distance :

$$(p, q) = (\sum_k^n \mu k|pk - qk|r)1/r$$

Keterangan:
N = Jumlah record data
K = Urutan field data
r = 2

Jarak adalah pendekatan yang umum dipakai untuk menentukan kesamaan atau ketidaksamaan dua vektor fitur yang dinyatakan dengan ranking. Apabila nilai ranking yang dihasilkan semakin kecil nilainya maka semakin dekat/tinggi kesamaan antara kedua vektor tersebut. Teknik pengukuran jarak dengan metode Euclidean menjadi salah satu metode yang paling umum

digunakan. Pengukuran jarak dengan metode euclidean dapat dituliskan dengan persamaan berikut :

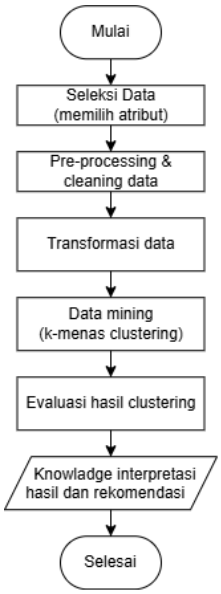
$$j(v_1, v_2) = \sum_{k=1}^{N} (v_1(k) - v_2(k))^2$$

dimana v_1 dan v_2 adalah dua vektor yang jaraknya akan dihitung dan N menyatakan panjang vektor.

3. BAHAN DAN METODE

3.1 Bahan yang dipakai pada penelitian ini yaitu data kecelakaan lalu lintas jalan raya di kota Samarinda pada tahun 2017-2022.

3.2 Metode yang dipakai pada penelitian ini yaitu knowladge discovery in database dengan alur sebagai berikut:



Gambar 1. Alur KDD

3.2.1 Data *Selection*

Pada proses ini dilakukan pemilihan himpunan data, menciptakan himpunan data target, atau memfokuskan pada subset variable (sampel data) dimana penemuan (discovery) akan dilakukan. Data yang dipilih terdiri dari 9 atribut yaitu nama jalan, bentuk geometri, kondisi permukaan jalan, kemiringan jalan, kondisi cahaya, cuaca, usia, perilaku serta jenis kelamin pengendara.

3.2.2 *Pre-Processing* dan *Cleaning* Data

Pre-Processing dan *Cleaning* Data dilakukan dengan membuang data yang tidak konsisten dan noise, duplikasi data, memperbaiki kesalahan data, dan bisa diperkaya dengan data eksternal yang relevan.

3.2.3 *Transformastion*

Proses ini mentransformasikan atau mengubah data dari bentuk kategorikal menjadi data numerik untuk melakukan proses mining.

3.2.4 *Data Mining*

Proses Data Mining yaitu proses penerapan algoritma k-means clustering ke dalam data untuk mencari pola atau informasi menarik dalam data terpilih sesuai dengan tujuan dari proses KDD secara keseluruhan. Pada tahap ini peneliti juga menggunakan metode elbow untuk mencari nilai k yang optimal.

3.2.5 *Interpretation* / Evaluasi

Proses untuk menerjemahkan pola-pola yang dihasilkan dari Data Mining. Mengevaluasi (menguji)

4. PEMBAHASAN

Pada penelitian ini data yang digunakan adalah data sekunder. Data sekunder sendiri merupakan sumber data yang didapatkan peneliti dengan media perantara atau tidak secara langsung. Data sekunder yang diperoleh dari Kepolisian Sektor Kota Samarinda yaitu Data kecelakaan pada tahun 2017 sampai tahun 2022. Data yang didapat sebanyak 1.640 data dan yang digunakan sebanyak 151 data. Kenapa data yang digunakan berjumlah sekian karena banyak data yang ada pada variabel tertentu bernilai null sehingga data tersebut tidak dapat digunakan untuk proses data mining. Berikut adalah potongan data awal, untuk bagian lainnya ada pada lampiran

No.Laka	Polres	Tanggal Kejadian	Tingkat Kecelakaan	Jumlah Meninggal Dunia
	KOTA SAMARINDA	2017-07-02 09:30:00	Ringan	0
	KOTA SAMARINDA	2017-10-11 07:30:00		0
	KOTA SAMARINDA	2017-02-22 17:50:00	Ringan	0
	KOTA SAMARINDA	2017-10-13 16:00:00	Ringan	0
	KOTA SAMARINDA	2017-10-23 14:30:00		0
	KOTA SAMARINDA	2017-01-10 20:00:00	Ringan	0
	KOTA SAMARINDA	2017-05-08 11:00:00		0
	KOTA SAMARINDA	2017-07-15 22:30:00	Ringan	0
	KOTA SAMARINDA	2017-05-21 15:00:00	Ringan	0
	KOTA SAMARINDA	2017-08-08 07:00:00		0
	KOTA SAMARINDA	2017-10-16 13:00:00	Ringan	0
	KOTA SAMARINDA	2017-10-24 16:00:00	Ringan	0
	KOTA SAMARINDA	2017-04-09 17:15:00		0
	KOTA SAMARINDA	2017-04-09 15:00:00		0
	KOTA SAMARINDA	2017-04-20 07:15:00	Ringan	0
	KOTA SAMARINDA	2017-10-16 16:00:00	Ringan	0
	KOTA SAMARINDA	2017-03-03 15:00:00		0
	KOTA SAMARINDA	2017-03-03 17:00:00		0

Gambar 2. Tampilan Data Awal

Data asli yang didapatkan penulis memiliki 31 atribut diantaranya yaitu No. Laka, Polres, Tanggal Kejadian, Tingkat Kecelakaan, Jumlah Meninggal Dunia, Jumlah Luka Berat, Jumlah Luka Ringan, Koordinat GPS – Lintang, Koordinat GPS – Bujur, Titik Acuan/Referensi, Jarak Ke Lokasi Kecelakaan, Arah Dari Titik Acuan Ke Lokasi Kecelakaan, Informasi Khusus, Tipe Kecelakaan, Kondisi Cahaya, Cuaca, Kecelakaan Menonjol, Nomor Jalan, Nama Jalan, Titik Jalan, Fungsi Jalan, Kelas Jalan, Tipe Jalan, Bentuk geometri, Kondisi Permukaan Jalan, Batas Kecepatan Di Lokasi, Kemiringan Jalan, Status Jalan, Perkiraan Nilai Rugi Material, Total Nilai Rugi Material Kendaraan, Laporan Export Keterangan Kerugian.

4.1 Seleksi Data (*Data Selection*)

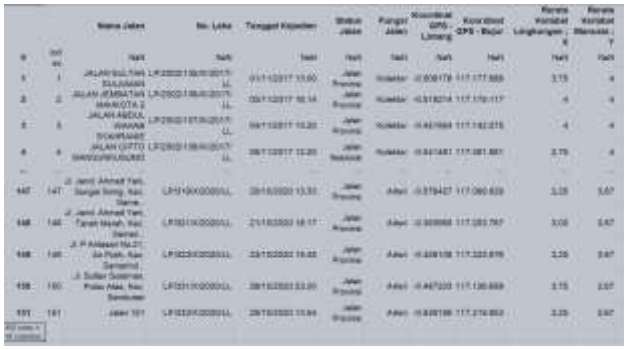
Pada tahap seleksi data, peneliti memilih data apa saja yang akan dipakai untuk dianalisis. Dari seluruh data yang didapatkan dengan total 31 atribut, hanya 9 atribut yang termasuk dalam variabel x (lingkungan) dan variabel y (manusia), kedua variabel tersebut yang penulis pakai untuk dianalisis. Berikut adalah tampilan seleksi data pada excel. Atribut-atribut tersebut antara lain, nama jalan, bentuk geometri, kondisi permukaan jalan, kemiringan jalan, kondisi cahaya, cuaca, usia, perilaku, serta jenis kelamin pengendara.

Tabel 1 Seleksi Data

Nama Jalan	Bentuk Geometri	Kondisi Permukaan Jalan	Kemiringan Jalan	Kondisi Cahaya	Cuaca	Usia	Perilaku	Jenis Kelamin
Jl. Sultan Sulaiman	Tikungan	Baik	Datar	Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
Jl. Jembatan Mahkota 2	Lurus	Baik	Datar	Terang/ jelas	Hujan/ gerimis	16-30	Tidak tertib	Laki - laki
Jl. Abdul Wahab Syahranie	Lurus	Baik	Datar	Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
Jl. Cipto Mangunkusumo	Tikungan			Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
Jl. Jendral A. Yani	Lurus	Baik	Datar	Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
Jl. Harun Nafsi	Lurus	Baik	Datar	Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
	T (Simpang tiga)	Berlubang		Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
Jl. Rapak Indah	Lurus	Baik	Datar	Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki
Jl. Wahid Hasyim 1	Lurus	Baik	Menanjak/menurun	Terang/ jelas	Cerah	16-30	Tidak tertib	Laki - laki

4.2 *Pre-Processing* dan *Cleaning Data*

Pre-processing dan *Cleaning* data adalah proses dimana peneliti melakukan pembersihan data agar tidak ada duplikasi data, lalu memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data, sehingga data tersebut dapat diolah dan dilakukan proses data *mining*. Pada proses ini penulis menggunakan excel untuk menghapus data-data yang tidak ada nilainya yang awalnya berjumlah 1.640 hanya tersisa 152 data. Kemudian data yang berjumalh 152 itu penulis *Cleaning* lagi menggunakan *python* untuk memastikan bahawa data yang bernilai null benar-benar sudah bersih. Data awal yang berjumlah 155 setelah melalui tahap *Cleaning* dengan python maka hanya tersisa 135 data yang siap dianalisis.



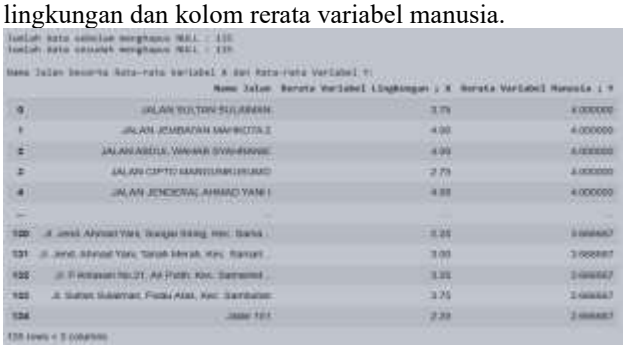
Gambar 3. Tampilan Data sebelum *cleaning*

Gambar 3 menampilkan data pada excel sebelum proses *Cleaning* menggunakan *python*. Data sebelum di*Cleaning* berjumlah 152 data, lalu proses *Cleaning* menghapus data yang bernilai null .

```
jumlah_sebelum = len(df)

df = df.dropna().reset_index(drop=True)

jumlah_sesudah = len(df)

print("Jumlah data sebelum menghapus NULL :", jumlah_sebelum)
print("Jumlah data sesudah menghapus NULL :", jumlah_sesudah)

print("\nNama Jalan beserta Rata-rata Variabel X dan Rata-rata Variabel Y:")
display(df[['Nama Jalan', 'Rerata Variabel Lingkungan ; X', 'Rerata Variabel Manusia ; Y']])
```

Gambar 4 Kode Program untuk Proses *Cleaning*

Gambar 4 menampilkan kode program untuk menghapus null menggunakan perintah df.dropna. Kode ini untuk perintah menampilkan ata setelah cleaning dan hanya menampilkan kolom nama jalan, kolom rerata variabel lingkungan dan kolom rerata variabel manusia.



Gambar 5 Tampilan Data setelah proses *Cleaning*

Gambar 4.4 menampilkan output dari proses *Cleaning* dengan kode df.dropna(). Data sebelum *cleaning* berjumlah 152 data, lalu proses *cleaning* menghapus data yang bernilai null atau data kosong yang tidak bisa diolah sehingga data yang tersisa hanya berjumlah 135 data. Pada tahap ini penulis hanya memilih data yang dipakai untuk ditampilkan. Kenapa hanya ada 3 kolom, karena dari 8 atribut yang dipilih pada tahap seleksi data 7 atribut lainnya sudah dihitung reratanya dan sudah tergabung kedalam kolom rerata variabel lingkungan (x) dan rerata variabel manusia (y).

4.3 Transformasi Data (*Transformastion*)

Pada tahap transformasi data, penulis melakukan perubahan data yang semula
berbentuk kategori atau teks (huruf) menjadi data numerik (angka) melalui proses pembobotan (*encoding*). Tujuannya adalah agar data dapat diproses oleh algoritma *k-means clustering*, karena algoritma tersebut hanya dapat melakukan perhitungan terhadap data yang bersifat numerik. Setiap kategori diberikan nilai atau bobot tertentu sesuai dengan klasifikasi yang telah ditentukan, sehingga seluruh atribut memiliki format yang seragam dan siap untuk dianalisis. Pembobotan dilakukan manual pada *microsoft excel* dengan menggunakan teknik *Attribute Value Weighting* yaitu memberi nilai bobot berdasarkan kepentingan kategori tersebut, semakin besar pengaruh kategori tersebut maka semakin tinggi nilai bobotnya. Hasil dari proses transformasi ini adalah dataset dalam bentuk numerik yang dapat digunakan pada tahap normalisasi dan proses *clustering*. Berikut adalah tampilan data pada excel setelah dilkukan tahap encoding.

Tabel 2 Hasil Generate Atribut baru dengan pembobotan

Bentuk Geometri		Kondisi Permukaan Jalan		Kemiringan Jalan		Kondisi Cahaya		Cuaca		Usia		Perilaku		Jenis Kelamin	
Tikungan	5	Baik	5	Datar	2	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Lurus	6	Baik	5	Datar	2	Terang/ Jelas	3	Hujan/ gerimis	3	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Lurus	6	Baik	5	Datar	2	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Tikungan	5		2		1	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Lurus	6	Baik	5	Datar	2	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Lurus	6	Baik	5	Datar	2	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
T (Simpang tiga)	4	Berlubang	3		1	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Lurus	6	Baik	5	Datar	2	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2
Lurus	6	Baik	5	Menanjak/ menurun	0	Terang/ Jelas	3	Cerah	4	16 - 30th	6	Tidak Tertib	4	Laki -laki	2

Tabel 2 menampilkan data setelah proses encoding atau pembobotan. Pembobotan dilakukan berdasarkan tingkat mendominasi kategori, semakin tinggi jumlah kategori dalam suatu atribut maka nilai pembobotan atribut tersebut semakin besar. Data yang akan di-inisialisasi merupakan data dari atribut bentuk geometri, kondisi permukaan jalan, kemiringan jalan, kondisi cahaya, cuaca, usia, perilaku, dan jenis kelamin seperti tabel berikut ini

Tabel 3 Inisialisasi Keterangan Bentuk Geometri

Keterangan Bentuk Geometri	Inisialisasi (Pembobotan)
Lurus	6
Tikungan	5
T (Simpang Tiga)	4
X atau + (Simpang Empat)	3
Y (Simpang Tiga)	2
Z (Tikungan Tajam)	1
Jembatan	1
Terowongan	1
Tidak Diketahui	0

Proses pembobotan pada tabel 3 dilakukan dengan menyesuaikan tingkat pengaruh kategori-kategori yang ada pada kejadian kecelakaan lalu lintas. Pada atribut bentuk geometri itu sendiri kategori jalan lurus menjadi kategori yang paling berpengaruh terjadinya kecelakaan, maka dari itu bobot pada kategori lurus lebih tinggi. Mengapa jalan lurus menjadi kategori paling tinggi untuk kejadian laka, karena berdasaran hasil wawancara para pengendara sering melampaui batas kecepatan pada jalan lurus. Dari tabel 4.2 juga menunjukan bahwa ada 3 kategori dengan bobot yang sama, hal itu terjadi karena dari ketiga kategori tersebut mempunyai jumlah kejadian yang sama terhadap pengaruh kecelakaan.

Tabel 4 Inisialisasi Kondisi Permukaan Jalan

Keterangan Permukaan Jalan	Inisialisasi (Pembobotan)
Baik/mulus/rata	5
Keriting	4
Berlubang	3
Berdebu	2
Tidak Diketahui	2
Basah	1

Proses pembobotan pada tabel 4 dilakukan dengan menyesuaikan tingkat pengaruh kategori-kategori yang mendominasi pada kejadian kecelakaan. Pada atribut kondisi permukaan jalan, kategori jalan baik/mulus/rata menjadi kategori yang paling berpengaruh terjadinya kecelakaan, maka dari itu bobot pada kategori lurus lebih tinggi. Dan dari tabel 4 juga menunjukan bahwa ada 2 kategori dengan bobot yang sama, hal itu terjadi karena dari kedua kategori tersebut mempunyai angka kejadian yang sama terhadap pengaruh kecelakaan setelah dilakukan proses *cleaning.*

Tabel 5 Inisialisasi Kemiringan Jalan

Keterangan Kemiringan Jalan	Inisialisasi (Pembobotan)
Datar	2
Menanjak/Menurun	1
Tidak Diketahui	0

Proses pembobotan pada tabel 5 langkahnya sama seperti proses pada tabel 3 dan tabel 4. Pada atribut kemiringan jalan, kategori jalan datar menjadi kategori yang paling berpengaruh terjadinya kecelakaan. Maka dari itu bobot pada kategori jalan datar lebih tinggi.

Tabel 6 Inisialisasi Kondisi Cahaya

Keterangan Kondisi Cahaya	Inisialisasi (Pembobotan)
Terang/jelas	4
Redup/Samar	3
Gelap	2
Tidak diketahui	1

Pada tabel 6 langkah pembobotan sama seperti tabel-tabel sebelumnya yaitu menentukan nilai bobot dengan melihat kategori mana yang paling berpengaruh untuk mendapatkan nilai bobot tertinggi. Dilihat dari atribut kondisi cahaya, kondisi cahaya terang menjadi ategori mendominasi dalam terjadinya laka, oleh karena itu bobot pada kategori cahaya terang/jelas menjadi bobot dengan nilai tertinggi. Alasannya karena pada saat cahaya terang para pengendara lebih sering berkendara dengan batas kecepatan yang tinggi.

Tabel 7 Inisialisasi Kondisi Cuaca

Keterangan Kondisi Cuaca	Inisialisasi (Pembobotan)
Cerah	3
Hujan/Gerimis	2
Berawan/Mendung	1
Hujan Disertai Angin Kencang	1
Tidak Diketahui	1

Pada tabel 7 langkah pembobotan sama seperti tabel sebelumnya yaitu menentukan nilai bobot dengan melihat kategori mana yang paling berpengaruh untuk mendapatkan nilai bobot yang lebih tinggi. Dilihat dari

atribut kondisi cuaca, cuaca terang menjadi kategori yang mendominasi dalam terjadinya kecelakaan. Tiga kategori lain yang memiliki nilai bobot sama pada tabel 4.6 karena angka pengaruhnya terhadap kecelakaan sama jumlahnya. Mengapa ada kategori tidak diketahui, karena pada data yang diterima peneliti ada data dengan kategori tersebut.

Tabel 8 Inisialisasi Perilaku

Keterangan Perilaku	Inisialisasi (Pembobotan)
Tidak Tertib	4
Lengah	3
Batas Kecepatan	2
Mabuk	1
Kantuk	1

Tabel 8 menunjukan angka bobot pada kategori tidak tertib menjadi kategori dengan bobot paling tinggi dari kategori lain karena pada atribut perilaku kategori tidak tertib menjadi faktor paling mendominasi saat terjadinya kecelakaan

Tabel 9 Inisialisasi Usia

Keterangan Usia	Inisialisasi (Pembobotan)
16-30 tahun	5
31-40 tahun	4
10-15 tahun	3
41-50 tahun	2
51 keatas	1

Tabel 9 menunjukan angka bobot pada kategori usia 16-30 tahun paling tinggi dari kategori lain. Mengapa demikian, karena berdasarkan data yang penulis dapatkan, pelaku kecelakaan atau usia pengendara penyebab kecelakaan didominasi oleh usia remaja sampai usia dewasa (16-30 tahun).

Tabel 10 Inialisasi Jenis Kelamin

Keterangan Jenis Kelamin	Inisialisasi (Pembobotan)
Laki-laki	2
Perempuan	1
Tidak Diketahui	0

Tabel 10 menunjukan kategori laki-laki menjadi kategori jenis kelamin pengendara yang paling mendominasi. Hal ini mempengaruhi nilai pembobotan. Mengapa ada kategori tidak diketahui, karena pada data yang diterima peneliti ada data dengan kategori tersebut.

Selanjutnya data yang sudah di inisialisasi dihitung reratanya. Dalam konteks penelitian kali ini dibuatlah menggunakan 2 variabel yang dibandingkan. Diantaranya ada kondisi lingkungan yang nantinya akan disebut dengan nilai X dan kondisi pengemudi yang nantinya disebut dengan nilai Y. Kondisi lingkungan dapat didefinisikan dengan beberapa atribut yang berkaitan dengan hal ini diantaranya:

1. Kondisi lingkungan, kondisi lingkungan itu terdiri dari dua faktor yang pertama kondisi lingkungan itu sendiri, dan selanjutnya faktor cuaca. Untuk faktor lingkungan terdapat beberap atribut diantaranya atribut Bentuk Geometri, Kondisi Permukaan Jalan dan Kemiringan Jalan. Kemudian untuk faktor cuaca terdiri dari Kondisi cahaya dan Cuaca itu sendiri.
2. Pengemudi, faktor pengemudi terdiri dari tiga atribut yaitu atribut perilaku, usia dan jenis kelamin.

Untuk menggambungkan atribut-atribut tadi maka diperlukan nilai rerata untuk mencari nilai X dan Y dengan hasil seperti dibawah ini

Tabel 11 Mencari Nilai X

Bentuk Geometri	Kondisi Permukaan Jalan	Kemiringan Jalan	Kondisi Cahaya	Cuaca	X
5	5	2	3	4	**3,75**
6	5	2	3	3	**3,67**
6	5	2	3	4	**3,92**
5	2	1	3	4	**3,08**
6	5	2	3	4	**3,92**
…	…	…	…	…	**…**
0	4	2	3	4	**2,75**
5	4	0	3	4	**3,25**
6	4	2	1	4	**3,25**
6	4	2	3	4	**3,75**
0	4	2	3	4	**2,75**

untuk mendapatkan nilai x maka yang dilakukan penelliti adalah menghitung nilai rata-ratanya menggunakan perintah average pada excel dan bisa juga menggunakan kode phyton pada google colab. Yang dihitung pertama adalah atribut kondisi lingkungan yaitu bobot bentuk geometri, bobot kondisi permukaan jalan dan bobot kemiringan jalan dan hasilnya adalah 3,3 .Selanjutnya peneliti menghitung rerata nilai pada kondisi cuaca yang meliputi bobot kondisi cahaya dan bobot cuaca menghasilkan nilai 3,5. Langkah terakhir, kedua hasil dari perhitungan tersebut dihitung lagi rata-ratanya dan hasilnya seperti yang tertera pada kolom X. Contoh rata-rata dari nilai 3,3 dan 3,5 adalah 3,4, maka nilai rata rata untuk variabel X adalah 3,4.

Tabel 12 Mencari Nilai Y

Usia	Perilaku	Jenis Kelamin	Variabel Y (Manusia)
6	4	2	**3,50**
6	4	2	**3,50**
6	4	2	**3,50**
6	4	2	**3,50**
6	4	2	**3,50**
6	4	2	**3,50**
….	….	….	**….**
5	4	2	**3,25**
5	4	2	**3,25**
5	4	2	**3,25**
2	4	2	**2,50**
2	4	2	**2,50**

Selanjutnya untuk mencari nilai y, peneliti melakukan perhitungan terhadap bobot perilaku, bobot usia dan bobot jenis kelamin. contoh bobot perilaku 4, bobot usia 6 bobot jenis kelamin 3, maka nilai rata-rata variabel manusia adalah untuk menghitung nilai y, sama seperti cara sebelumnya, tambahkan kolom bobot usia, kolom bobot perilaku dan juga kolom bobot jenis kelamin lalu direrata menggunakan perintah average maka hasilnya akan sama seperti di kolom Y.

4.4 *Data Mining*

Proses data mining adalah sebagai berikut:

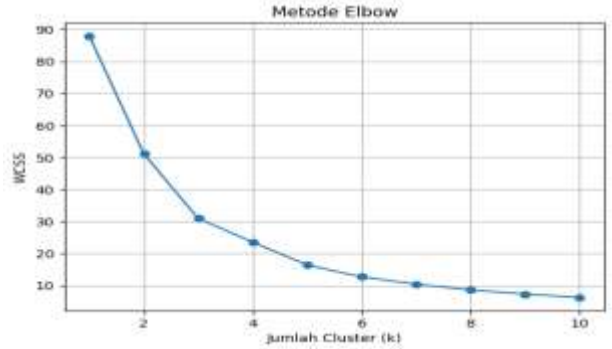
Pertama penulis mencari nilai k yang optimal dengan *Elbow method* atau menggunakan nilai *Sum of Squared Errors* (SSE) atau sering juga disebut *Within Cluster Sum of Squares* (WCSS) dengan menentukan k = 1, 2, 3, 4,

5,…10 lalu dihitung menggunakan nilai WCSS seperti pada kode program berikut:

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

X = df[["Rerata Variabel Lingkungan ; X", "Rerata Variabel Manusia ; Y"]].values

#Cari jumlah cluster optimal (Metode Elbow)
wcss = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, init='k-means++', random_state=42, n_init=10)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 11), wcss, marker='o')
plt.title("Metode Elbow")
plt.xlabel("Jumlah Cluster (k)")
plt.ylabel("WCSS")
plt.grid(True)
plt.show()
```

Gambar 6. Kode Program untuk menentukan nilai K

Setelah dihitung menggunakan nilai SSE pada tahap sebelumnya maka penulis mendapkan nilai optimal k dengan k=4 seperti pada gamabr dibawah. Kenapa nilanya 4 karena pada grafik elbow, titik siku (elbow *point*) terlihat pada K = 4, yaitu saat penurunan nilai *Sum of Squared Errors* (SSE) mulai melambat. Hal ini menunjukkan bahwa empat *Cluster* sudah mampu mengelompokkan data dengan baik.



Gambar 7. Grafik *Elbow method*

Pada grafik, sumbu **X** menunjukkan **jumlah *Cluster* (K)** yang diuji, yaitu dari **K = 1 hingga K = 10**, sedangkan sumbu **Y** menunjukkan nilai **Within *Cluster* Sum of Squares (WCSS)**. Nilai WCSS merupakan jumlah kuadrat jarak setiap data terhadap centroid *Cluster* masing-masing. Semakin kecil nilai WCSS, semakin dekat data terhadap centroid sehingga variasi di dalam *Cluster* semakin kecil.

Berdasarkan grafik metode *Elbow* terlihat bahwa nilai WCSS mengalami penurunan**yang sangat tajam dari K = 1 hingga K = 3**. Hal ini menunjukkan bahwa penambahan jumlah *Cluster* pada tahap awal memberikan peningkatan kualitas pengelompokan yang signifikan karena data menjadi lebih homogen di dalam setiap *cluster*.

Setelah **K = 3**, penurunan nilai WCSS mulai melambat. Pada rentang **K = 4 hingga K = 10**, kurva cenderung lebih landai. Kondisi ini menunjukkan bahwa penambahan jumlah *Cluster* setelah titik tersebut hanya memberikan penurunan WCSS yang relatif kecil sehingga peningkatan kualitas *clustering* tidak lagi signifikan. Dengan demikian, titik siku (elbow) pada grafik berada di sekitar k=3 hingga k=5. Oleh karena itu, jumlah *Cluster* pada rentang tersebut menjadi kandidat terbaik untuk dilakukan evaluasi lebih lanjut menggunakan metrik validasi lain.

Pada penelitian ini, evaluasi lanjutan dilakukan menggunakan **Silhouette Score** terhadap nilai **K = 3, K = 4, dan K = 5**, dengan kode program sebagai berikut:

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import pandas as pd

hasil = []

for k in [3, 4, 5]:
    # Membentuk model K-Means
    kmeans = KMeans(
        n_clusters=k,
        random_state=42,
        n_init=10
    )

    # Melakukan clustering
    labels = kmeans.fit_predict(X)

    # Menghitung Silhouette Score
    score = silhouette_score(X, labels)

    hasil.append([k, score])

# Membuat tabel hasil
hasil_df = pd.DataFrame(
    hasil,
    columns=["Jumlah Cluster (K)", "Silhouette Score"]
)

# Menampilkan hasil
print(hasil_df)
```

Gambar 8. Kode Program Untuk Evaluasi Nilai K

Titik siku (*elbow*) berada pada rentan k=3 sampai k=5, maka dari itu perlu melakukan evaluasi nilai k untuk mendapatkan nilai k yang optimal menggunakan silhouette score. Berikut adalah output atau hasil dari evaluasi untuk nilai k

```
   Jumlah Cluster (K)  Silhouette Score
0                   3          0.457845
1                   4          0.468391
2                   5          0.501195

==================================
HASIL EVALUASI SILHOUETTE SCORE
==================================
K Terbaik           : 5
Silhouette Score    : 0.5012
```

Gambar 9 Hasil Evaluasi Nilai K

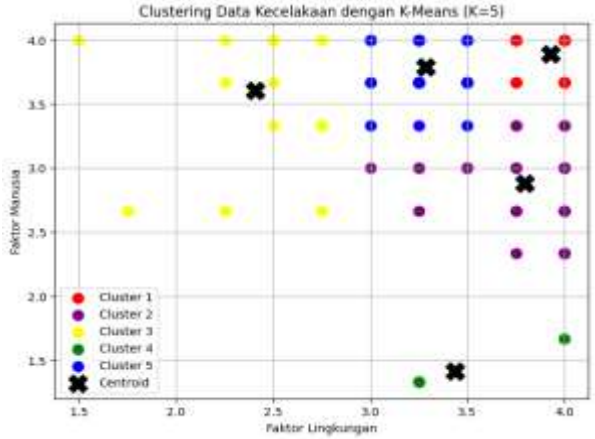
Berdasarkan grafik Metode *Elbow* pada gambar 4.9, titik siku berada pada rentang **K = 3 hingga K = 5**, yang menandakan bahwa penambahan jumlah *Cluster* setelah rentang tersebut tidak lagi memberikan penurunan WCSS yang signifikan. Setelah dilakukan validasi menggunakan **Silhouette Score**, diperoleh bahwa **K = 5** menghasilkan nilai tertinggi sebesar **0,501195**, sehingga dipilih sebagai jumlah *Cluster* yang paling optimal. Dengan demikian, penentuan jumlah *Cluster* pada penelitian ini tidak hanya didasarkan pada Metode *Elbow*, tetapi juga diperkuat

melalui evaluasi *Silhouette Score* sehingga keputusan pemilihan **K = 5** memiliki dasar yang lebih kuat.

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# K-Means
kmeans = KMeans(
    n_clusters=5,
    init='k-means++',
    random_state=42,
    n_init=10
)

y_kmeans = kmeans.fit_predict(X)

# Warna tiap cluster
colors = ['red', 'purple', 'yellow', 'green', 'blue']

plt.figure(figsize=(8,6))

for i in range(5):
    plt.scatter(
        X[y_kmeans == i, 0],      # Faktor Lingkungan
        X[y_kmeans == i, 1],      # Faktor Manusia
        s=80,
        c=colors[i],
        label=f'Cluster {i+1}'
    )
```

Gambar 10 Kode Program untuk Visualisasi Hasil *Cluster*

Gambar 10 menampilkan kode program yang digunakan untuk visualisasi hasil *Cluster* dengan nilai k=5. *Cluster* 1 berwarna merah, *Cluster* 2 berwarna ungu, *Cluster* 3 berwarna kuning, *Cluster* 4 berwarna hijau dan *Cluster* 5 berwarna biru. Maka berikut adalah output dari kode program diatas:



Gambar 11 Visualisasi Hasil *Cluster*

Gambar 11 menunjukkan bahwa data kecelakaan berhasil dikelompokkan menjadi **lima *Cluster*** dengan karakteristik yang berbeda berdasarkan kombinasi faktor lingkungan dan faktor manusia. Terlihat bahwa sebagian besar data berada pada area dengan nilai faktor lingkungan dan faktor manusia yang relatif tinggi, sedangkan terdapat satu *Cluster* kecil (*Cluster* 4) yang memiliki karakteristik berbeda dengan jumlah anggota yang lebih sedikit.

Hasil visualisasi ini juga konsisten dengan hasil evaluasi menggunakan **Metode** *Elbow* dan **Silhouette Score**, di mana **K = 5** dipilih sebagai jumlah *Cluster* yang paling optimal. Dengan demikian, pembagian menjadi lima *Cluster* mampu memberikan pemisahan data yang cukup baik sehingga memudahkan identifikasi karakteristik masing-masing kelompok kecelakaan dan dapat digunakan sebagai dasar dalam menentukan prioritas penanganan maupun penyusunan strategi pencegahan kecelakaan lalu lintas di Kota Samarinda.

Cluster **1 (Merah)** berada pada bagian kanan atas grafik dengan nilai faktor lingkungan sekitar **3,8–4,0** dan faktor manusia sekitar **3,7–4,0**. *Cluster* ini menunjukkan kelompok kecelakaan yang memiliki nilai tinggi pada kedua variabel. Artinya, kejadian kecelakaan dalam *Cluster* ini dipengaruhi oleh faktor lingkungan dan faktor manusia yang sama-sama tinggi. *Cluster* **2 (Ungu)** berada pada nilai faktor lingkungan yang tinggi (sekitar **3,0–4,0**) namun faktor manusia berada pada tingkat sedang (sekitar **2,3–3,3**). Hal ini menunjukkan bahwa kelompok ini memiliki kondisi lingkungan yang relatif serupa dengan *Cluster* 1, tetapi tingkat pengaruh faktor manusia tidak setinggi *Cluster* 1. *Cluster* **3 (Kuning)** berada pada sisi kiri grafik dengan nilai faktor lingkungan lebih rendah (sekitar **1,5–2,8**) namun faktor manusia relatif tinggi (sekitar **2,7–4,0**). Hal ini menunjukkan bahwa pada kelompok ini, kecelakaan lebih didominasi oleh faktor manusia dibandingkan faktor lingkungan. *Cluster* **4 (Hijau)** hanya terdiri dari sedikit data yang terletak pada bagian bawah grafik. *Cluster* ini memiliki nilai faktor manusia paling rendah (sekitar **1,3–1,7**) dengan faktor lingkungan berada pada kisaran **3,2–4,0**. Jumlah anggotanya yang sedikit menunjukkan bahwa *Cluster* ini merupakan kelompok khusus (*minor cluster*) dengan karakteristik yang berbeda dari kelompok lainnya.

***Cluster* 5 (Biru)** berada pada bagian kanan atas hingga tengah dengan nilai faktor lingkungan tinggi (sekitar **3,0–3,5**) dan faktor manusia juga tinggi (sekitar **3,3–4,0**). *Cluster* ini memiliki karakteristik yang mirip dengan *Cluster* 1, namun centroid-nya menunjukkan bahwa rata-rata kedua variabel sedikit lebih rendah sehingga terbentuk sebagai kelompok yang berbeda. Berikutnya penulis akan menampilkan jumlah data yang terbagi dalam setiap *Cluster* dengan kode program:

```
# ============================================================
# Menampilkan jumlah dan persentase anggota setiap cluster
# ============================================================

print("Jumlah anggota setiap cluster:\n")

jumlah_cluster = df_cluster['Cluster'].value_counts().sort_index()

total_data = len(df_cluster)

for cluster, jumlah in jumlah_cluster.items():
    persentase = (jumlah / total_data) * 100
    print(f"Cluster {cluster} : {jumlah} data ({persentase:.2f}%)")
```

Gambar 12 Kode Untuk Menampilkan Jumlah Data Tiap *Cluster*

```
Jumlah anggota setiap cluster:

Cluster 1 : 42 data (31.11%)
Cluster 2 : 33 data (24.44%)
Cluster 3 : 16 data (11.85%)
Cluster 4 : 4 data (2.96%)
Cluster 5 : 40 data (29.63%)
```

Gambar 13 Output Jumlah Data

Dari gambar 13 dapat dilihat jumlah data dari masing-masing *cluster*. *Cluster* 1 memiliki jumlah data sebanyak 42 data dengan total 31.11% dari seluruh data. *Cluster* 2 memiliki jumlah data sebanyak 33 data dengan total 24.44%, *Cluster* 3 memiliki 16 data dengan total 11.85%, *Cluster* 4 memiliki 4 data dengan total 2.96%, sedangkan *Cluster* 5 memiliki 40 data dengan total 29.63%. Berikut penulis akan membuat kode program untuk menampilkan rata-rata serta karakteristik setiap *cluster*.

```
df_cluster = df.copy()
df_cluster['Cluster'] = y_kmeans

summary = df_cluster.groupby("Cluster")[
    ["Rerata Variabel Lingkungan ; X", "Rerata Variabel Manusia ; Y"]
].mean().round(2)

# Rename columns to match the later usage in the loop
summary = summary.rename(columns={
    "Rerata Variabel Lingkungan ; X": "Faktor Lingkungan",
    "Rerata Variabel Manusia ; Y": "Faktor Manusia"
})

for cluster, row in summary.iterrows():

    print("="*60)
    print(f"Cluster {cluster}")

    print(f"Rata-rata Faktor Lingkungan : {row['Faktor Lingkungan']:.2f}")
    print(f"Rata-rata Faktor Manusia    : {row['Faktor Manusia']:.2f}")

    if row["Faktor Lingkungan"] >= 3.5:
        lingkungan = "tinggi"
    elif row["Faktor Lingkungan"] >= 2.5:
        lingkungan = "sedang"
    else:
        lingkungan = "rendah"

    if row["Faktor Manusia"] >= 3.5:
        manusia = "tinggi"
```

Gambar 14 kode untuk menampilkan rata-rata serta karakteristik *cluster*

Kode pada gambar 14 adalah perintah untuk menampilkan nilai rata-rata setiap *cluster*, mengubah nama "rerata variabel" menjadi "faktor"dan menampilkan karakteristik setiap *cluster*. Kategori untuk karakteristik setiap *cluster* yaitu jika rata-rata faktor lingkungan dan manusia nilainya lebih besar atau sama dengan 3.5 maka faktor tersebut berada pada kategori tinggi. Jika nilai rata-rata faktor lingkungan dan faktor manusia kurang dari atau sama dengan 2.5 maka berada pada kategori sedang. Jika nilai rata-rata untuk faktor lingkungan dan faktor manusia kurang dari 2.5 maka kategori untuk nilai rata-rata tersebut adalah rendah.

```
Cluster 1
Rata-rata Faktor Lingkungan : 3.93
Rata-rata Faktor Manusia    : 3.89
Karakteristik :
- Faktor Lingkungan tinggi
- Faktor Manusia tinggi
============================================================
Cluster 2
Rata-rata Faktor Lingkungan : 3.80
Rata-rata Faktor Manusia    : 2.88
Karakteristik :
- Faktor Lingkungan tinggi
- Faktor Manusia sedang
============================================================
Cluster 3
Rata-rata Faktor Lingkungan : 2.41
Rata-rata Faktor Manusia    : 3.60
Karakteristik :
- Faktor Lingkungan rendah
- Faktor Manusia tinggi
============================================================
Cluster 4
Rata-rata Faktor Lingkungan : 3.44
Rata-rata Faktor Manusia    : 1.42
Karakteristik :
- Faktor Lingkungan sedang
- Faktor Manusia rendah
============================================================
Cluster 5
Rata-rata Faktor Lingkungan : 3.29
Rata-rata Faktor Manusia    : 3.79
Karakteristik :
- Faktor Lingkungan sedang
- Faktor Manusia tinggi
```

Gambar 15 Nilai rata-rata dan karakteristik setiap *cluster*

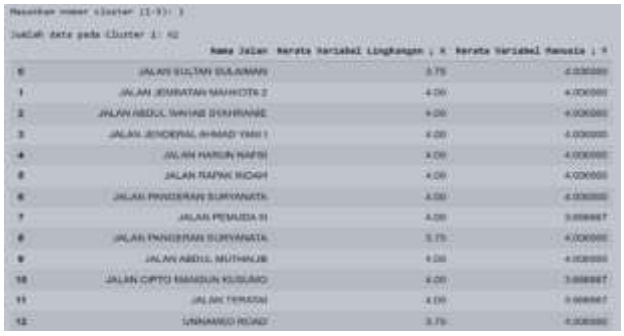
Dari gambar diatas dapat dilihat bahwa *cluster* 1 dengan nilai rata-rata pada faktor lingkungan tinggi dan faktor manusia tinggi. *Cluster* 2 dengan nilai rata-rata pada faktor lingkungan tinggi dan faktor manusia sedang. *Cluster* 3 dengan nilai rata-rata pada faktor lingkungan rendah dan faktor manusia tinggi. *Cluster* 4 dengan nilai rata-rata pada faktor lingkungan sedang dan faktor manusia rendah. *Cluster* 5 dengan nilai rata-rata pada faktor lingkungan sedang dan faktor manusia tinggi.

Berikut adalah tampilkan kode untuk melihat keseluruhan data yang ada pada *cluster* dengan memilih ingin melihat *cluster* apa.

```
nomor_cluster = int(input("Masukkan nomor cluster (1-5): "))

# Memilih data sesuai cluster
data_cluster = df_cluster[df_cluster['Cluster'] == nomor_cluster]

print(f"\nJumlah data pada Cluster {nomor_cluster}: {len(data_cluster)}")

# Menampilkan hanya kolom yang dibutuhkan
display(
    data_cluster[
        ['Nama Jalan',
         'Rerata Variabel Lingkungan ; X',
         'Rerata Variabel Manusia ; Y']
    ].reset_index(drop=True)
)
```

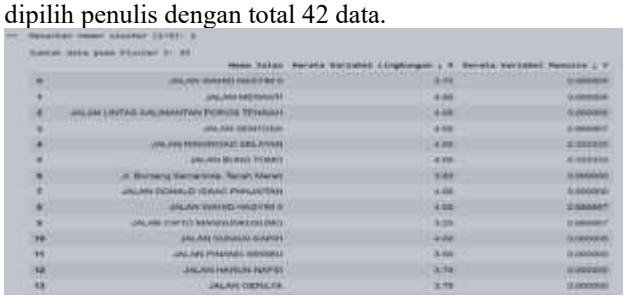
Gambar 16 Kode Untuk Memilih *Cluster* secara Interaktif

Kode program pada gambar 16 untuk menampilkan seluruh data yang ada pada *cluster* 1, *cluster* 2, *cluster* 3, *cluster* 4 dan *cluster* 5



Gambar 17 Data Dari *Cluster* 1

Gambar 17 merupakan output dari kode program pada gambar 16. Gambar diatas menampilkan *cluster* 1 yang dipilih penulis dengan total 42 data.



Gambar 18 Menampilkan Data Dari *Cluster* 2

Gambar 18 merupakan output dari kode program pada gambar 16. Gambar diatas menampilkan *cluster* 2 yang dipilih penulis dengan total 33 data.

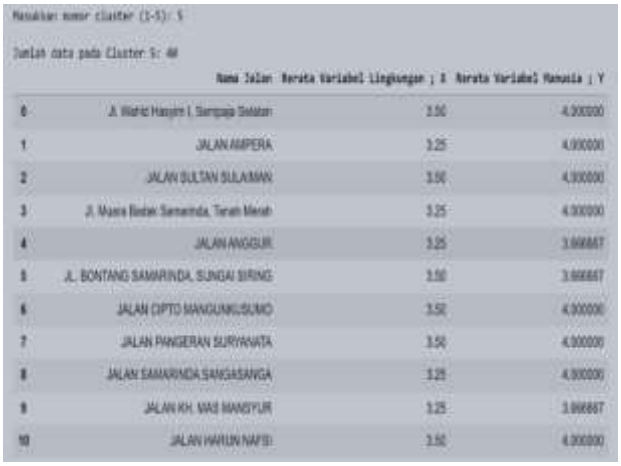
Gambar 19 Menampilkan Data Dari *Cluster* 3

Gambar 19 merupakan output dari kode program pada gambar 16. Gambar diatas menampilkan *cluster* 3 yang dipilih penulis dengan total 16 data.

	Nama Jalan	Rerata Variabel Lingkungan ; X	Rerata Variabel Manusia ; Y
0	Unnamed Road, Sungai Siring, Samarinda Utara	4.00	1.666667
1	Jl. Teuku Umar No.38, Karang Anyar, Kec. Sungai...	3.25	1.333333
2	Jl. Poros Kebun Agung No.38, Lempake, Kec. Sam...	3.25	1.333333
3	Rw., Rw. Makmur, Kec. Palaran	3.25	1.333333

Gambar 20 Menampilkan Data Dari *Cluster* 4

Gambar 20 merupakan output dari kode program pada gambar 16. Gambar diatas menampilkan *cluster* 4 yang dipilih penulis dengan total 4 data.



	Nama Jalan	Rerata Variabel Lingkungan ; X	Rerata Variabel Manusia ; Y
0	Jl. Wahid Hasyim I, Sempaja Selatan	3.50	4.000000
1	JALAN AMPERA	3.25	4.000000
2	JALAN SULTAN SULAIMAN	3.50	4.000000
3	Jl. Muara Badak Samarinda, Teraih Merah	3.25	4.000000
4	JALAN ANGGUR	3.25	3.666667
5	JL. BONTANG SAMARINDA, SUNGAI SIRING	3.50	3.666667
6	JALAN CIPTO MANGUNKUSUMO	3.50	4.000000
7	JALAN PANGERAN SURYANATA	3.50	4.000000
8	JALAN SAMARINDA SANGASANGA	3.25	4.000000
9	JALAN KH. WAS MANSYUR	3.25	3.666667
10	JALAN HARUN NAFSI	3.50	4.000000

Gambar 21 Menampilkan Data Dari *Cluster* 5

Gambar 21 merupakan output dari kode program pada gambar 16. Gambar diatas menampilkan *cluster* 5 yang dipilih penulis dengan total 40 data.

Dari 5 *Cluster* diatas dapat disimpulkan bahwa *Cluster* dengan jumlah data tertinggi ada pada *Cluster* 1. *Cluster* 1 merupakan merupakan kelompok dengan tingkat nilai rata-rata faktor lingkungan dan faktor manusia yang tinggi. Nilai rerata faktor lingkungan berada pada kisaran **3,75–4,00**, sedangkan faktor manusia berada pada kisaran **3,67–4,00**. Hal ini menunjukkan bahwa kejadian kecelakaan pada kelompok ini dipengaruhi secara bersamaan oleh kondisi lingkungan jalan serta faktor manusia. Dengan demikian, *Cluster* ini dapat dikategorikan sebagai kelompok dengan tingkat risiko kecelakaan yang tinggi sehingga memerlukan prioritas penanganan melalui peningkatan keselamatan infrastruktur jalan dan pengendalian perilaku pengguna jalan. Berikut akan dibuatkan kode untuk menampilkan nama jalan dengan tingkat kecelakaan terbanyak berdasarkan hasil penelitian ini:

```
import pandas as pd
import re

# Salin dataframe agar data asli tidak berubah
df_cluster = df.copy()

# Fungsi normalisasi nama jalan
def normalisasi_jalan(nama):
    nama = str(nama).upper().strip()

    # Hilangkan tanda baca
    nama = re.sub(r'[.,/-]', ' ', nama)

    # Hilangkan spasi berlebih
    nama = re.sub(r'\s+', ' ', nama)

    # Samakan variasi penulisan
    pengganti = {
        "A WAHAB SYAHRANIE": "ABDUL WAHAB SYAHRANIE",
        "A. WAHAB SYAHRANIE": "ABDUL WAHAB SYAHRANIE",
        "JALAN A WAHAB SYAHRANIE": "JALAN ABDUL WAHAB SYAHRANIE",
        "JALAN A. WAHAB SYAHRANIE": "JALAN ABDUL WAHAB SYAHRANIE",

        "CIPTO MANGUNKUSUMO": "CIPTO MANGUN KUSUMO",
        "JALAN CIPTO MANGUNKUSUMO": "JALAN CIPTO MANGUN KUSUMO",

        "APT PRANOTO": "APT. PRANOTO",
        "APT. PRANOTO": "APT. PRANOTO",

        "PM NOOR": "P M NOOR",
        "PM. NOOR": "P M NOOR"
    }

    for lama, baru in pengganti.items():
        nama = nama.replace(lama, baru)

    return nama
```

Gambar 22 Kode Untuk Menampilkan Nama Jalan

Gambar 22 merupakan kode untuk menampilkan nama jalan yang paling sering terjadi kecelakaan. Karena ada beberapa nama jalan dari data asli yang penulisannya berbeda walaupun merupakan nama jalan yang sama maka peneli membuatkan kode untuk samakan variasi penulisan seperti kode diatas. Berikut adalah 10 nama jalan dengan 2 jaan teratas sebagai jalan dengan jumlah laka terbanyak sesuai data yang penulis dapatkan.

Nama Jalan Normal	Jumlah Kecelakaan
JALAN ABDUL WAHAB SYAHRANIE	5
JALAN HARUN NAFSI	5
JALAN JAKARTA	2
JALAN CIPTO MANGUN KUSUMO	2
JALAN PATTIMURA	2
JALAN PANGERAN SURYANATA	2
JALAN APT. PRANOTO	1
JALAN AMPERA	1
JALAN GAJAH MADA	1
JALAN DONALD ISAAC PANJAITAN	1

Gambar 23 Menampilkan Nama Jalan

Gambar 23 merupakan hasil dari output gambar 4.22. Gambar ini menampilkan 10 data teratas nama jalan dengan 2 nama jalan yang memiliki jumlah kecelakaan tertinggi yaitu **"jalan Abdul Wahab Syahranie**" dan **"jalan Harun Nafsi".** Jalan Abdul Wahab Syahrani adalah salah satu jalan utama dan pusat aktivitas penting di Kota Samarinda, Kalimantan Timur, yang melintasi kawasan strategis seperti Samarinda Ulu dan Samarinda Utara. Jalan ini dikenal sebagai jalur penghubung padat yang dikelilingi oleh area perumahan, pusat bisnis, serta fasilitas publik. Sedangkan jalan Harun Nafsi adalah jalan yang terletak di Kelurahan Rapak Dalam, Kecamatan Loa Janan Ilir, Kota Samarinda, Kalimantan Timur, di wilayah Samarinda Seberang.

4.5 Interpretation / Evaluasi

Untuk mengevaluasi hasil *clustering* maka penulis menggunakan tiga metrik evaluasi internal, yaitu *Silhouette Score* untuk mengukur kualitas *clustering* berdasarkan tingkat kekompakan (cohesion) dan keterpisahan (separation) antar*cluster* Nilainya berkisar dari -1 hingga 1, semakin dekat dengan 1 maka kualitas *clustering* semakin baik. *Davies-Bouldin Index* (DBI), mengukur rata-rata kemiripan antar *cluster*. Nilai yang lebih kecil menunjukkan *cluster* yang lebih kompak dan terpisah dengan baik. Calinski-Harabasz Index (CHI) Mengukur rasio antara variasi antar*cluster* (*between-cluster dispersion*) dan variasi di dalam *cluster* (*within-cluster dispersion*). Nilainya selalu positif. Semakin besar nilai CHI, maka semakin baik kualitas *clustering*, karena menunjukkan *cluster* yang kompak dan memiliki pemisahan yang jelas. Berikut akan penulis tampilkan kode untuk evaluasi hasil *clustering*;

```
from sklearn.metrics import (
    silhouette_score,
    silhouette_samples,
    davies_bouldin_score,
    calinski_harabasz_score
)

print("="*60)
print("EVALUASI HASIL CLUSTERING")
print("="*60)


silhouette_avg = silhouette_score(X, y_kmeans)

print(f"\nSilhouette Score : {silhouette_avg:.4f}")

if silhouette_avg >= 0.70:
    print("Interpretasi : Struktur cluster sangat baik.")
elif silhouette_avg >= 0.50:
    print("Interpretasi : Struktur cluster baik.")
elif silhouette_avg >= 0.25:
    print("Interpretasi : Struktur cluster cukup baik.")
else:
    print("Interpretasi : Struktur cluster kurang baik.")


dbi = davies_bouldin_score(X, y_kmeans)

print(f"\nDavies-Bouldin Index : {dbi:.4f}")

if dbi < 1:
    print("Interpretasi : Cluster sangat baik.")
elif dbi < 2:
    print("Interpretasi : Cluster cukup baik.")
else:
    print("Interpretasi : Cluster kurang baik.")


chi = calinski_harabasz_score(X, y_kmeans)
```

Gambar 24 Kode Untuk Mengukur Kualitas Hasil *Clustering*

Gambar 24 merupakan kode program yang dipakai untuk mengukur kualitas hasil *clustering* dengan menggunakan Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI).

```
============================================================
EVALUASI HASIL CLUSTERING
============================================================

Silhouette Score : 0.5012
Interpretasi : Struktur cluster baik.

Davies-Bouldin Index : 0.6800
Interpretasi : Cluster sangat baik.

Calinski-Harabasz Index : 141.8924

Semakin besar nilai Calinski-Harabasz Index
menunjukkan kualitas cluster semakin baik.
```

Gambar 25 Output Evaluasi Hasil *Clustering*

Hasil evaluasi menunjukkan bahwa algoritma *K-Means* dengan jumlah *Cluster* sebanyak lima menghasilkan kualitas pengelompokan yang baik. Hal ini ditunjukkan oleh nilai **Silhouette Score sebesar 0,5012**, yang mengindikasikan bahwa struktur *Cluster* telah terbentuk dengan baik. Selain itu, nilai **Davies-Bouldin Index sebesar 0,6800** menunjukkan bahwa setiap *Cluster* memiliki tingkat kekompakan yang baik serta pemisahan yang cukup jelas antar*cluster*. Sementara itu, **Calinski-Harabasz Index sebesar 141,8924** menunjukkan bahwa variasi antar*cluster* lebih besar dibandingkan variasi di dalam *cluster*. Secara keseluruhan, ketiga metrik evaluasi tersebut menunjukkan bahwa hasil *clustering* yang diperoleh layak digunakan untuk menganalisis karakteristik kecelakaan lalu lintas di Kota Samarinda.

5. KESIMPULAN

Kesimpulan pada penelitian ini yaitu:

1. Penelitian ini berhasil menerapkan tahapan Knowledge Discovery in Database (KDD) yang meliputi *data selection*, *data cleaning*, *data transformation*, *data mining*, dan *evaluation* dalam

proses pengelompokan data kecelakaan lalu lintas. Setelah dilakukan proses pembersihan data, diperoleh data yang layak digunakan untuk proses *clustering*.

2. Metode *Attribute Value Weighting* berhasil mengubah atribut kategorikal menjadi data numerik sehingga dapat diproses menggunakan algoritma K-Means. Selanjutnya atribut tersebut diringkas menjadi dua variabel utama, yaitu Faktor Lingkungan (X) yang terdiri atas bentuk geometri jalan, kondisi permukaan jalan, kemiringan jalan, kondisi cahaya, dan cuaca, serta Faktor Manusia (Y) yang terdiri atas perilaku, usia, dan jenis kelamin. Kedua variabel tersebut digunakan sebagai dasar dalam proses pengelompokan data.
3. Proses *clustering* menggunakan algoritma *K-Means* berhasil mengelompokkan data kecelakaan lalu lintas menjadi lima cluster yang memiliki karakteristik berbeda, yaitu: Cluster 1 memiliki rata-rata Faktor Lingkungan sebesar 3,93 dan Faktor Manusia sebesar 3,89, sehingga termasuk kategori faktor lingkungan tinggi dan faktor manusia tinggi. Cluster ini menunjukkan bahwa kecelakaan dipengaruhi secara bersamaan oleh kondisi lingkungan jalan dan faktor manusia, sehingga memerlukan penanganan yang menyasar kedua aspek tersebut. Cluster 2 memiliki rata-rata Faktor Lingkungan sebesar 3,80 dan Faktor Manusia sebesar 2,88, sehingga termasuk kategori faktor lingkungan tinggi dan faktor manusia sedang. Hal ini menunjukkan bahwa kondisi lingkungan jalan menjadi faktor yang lebih dominan dibandingkan faktor manusia. Cluster 3 memiliki rata-rata Faktor Lingkungan sebesar 2,41 dan Faktor Manusia sebesar 3,60, sehingga termasuk kategori faktor lingkungan rendah dan faktor manusia tinggi. Cluster ini mengindikasikan bahwa kecelakaan lebih banyak dipengaruhi oleh perilaku dan karakteristik pengguna jalan dibandingkan kondisi lingkungan. Cluster 4 memiliki rata-rata Faktor Lingkungan sebesar 3,44 dan Faktor Manusia sebesar 1,42, sehingga termasuk kategori faktor lingkungan sedang dan faktor manusia rendah. Karakteristik ini menunjukkan bahwa faktor lingkungan memiliki pengaruh yang lebih besar dibandingkan faktor manusia terhadap kejadian kecelakaan. Cluster 5 memiliki rata-rata Faktor Lingkungan sebesar 3,29 dan Faktor Manusia sebesar 3,79, sehingga termasuk kategori faktor lingkungan sedang dan faktor manusia tinggi. Hal ini menunjukkan bahwa faktor manusia masih menjadi penyebab yang dominan meskipun kondisi lingkungan juga memberikan kontribusi terhadap terjadinya kecelakaan.
4. Hasil evaluasi menggunakan Silhouette Score, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI) menunjukkan bahwa model *K-Means* menghasilkan kualitas *clustering* yang baik. Nilai Silhouette Score sebesar 0,5012 menunjukkan bahwa struktur cluster telah terbentuk dengan baik, Davies-Bouldin Index sebesar 0,6800 menunjukkan cluster yang kompak dan memiliki pemisahan yang cukup jelas, sedangkan Calinski-Harabasz Index sebesar 141,8924 menunjukkan bahwa variasi antarcluster lebih besar dibandingkan variasi di dalam cluster. Dengan demikian, hasil *clustering* layak digunakan untuk menganalisis karakteristik kecelakaan lalu lintas di Kota Samarinda.
5. Secara keseluruhan, hasil penelitian menunjukkan bahwa algoritma *K-Means Clustering* mampu mengelompokkan data kecelakaan lalu lintas berdasarkan karakteristik faktor lingkungan dan faktor manusia. Informasi yang dihasilkan dari masing-masing cluster dapat dimanfaatkan sebagai dasar bagi Polresta Samarinda, Dinas Perhubungan, dan instansi terkait dalam menentukan prioritas penanganan, peningkatan keselamatan jalan, serta penyusunan strategi pencegahan kecelakaan yang lebih tepat sasaran sesuai karakteristik setiap kelompok data.

6. SARAN

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat diberikan untuk pengembangan penelitian selanjutnya, yaitu sebagai berikut:

1. Bagi Polresta Samarinda dan instansi terkait, hasil pengelompokan data kecelakaan menggunakan algoritma *K-Means* dapat dimanfaatkan sebagai bahan pendukung dalam menentukan prioritas penanganan pada ruas jalan yang memiliki tingkat risiko kecelakaan lebih tinggi. Informasi hasil *clustering* diharapkan dapat menjadi dasar dalam penyusunan strategi pencegahan kecelakaan, seperti meningkatkan pengawasan dengan melakukan patroli serta pelaksanaan program edukasi keselamatan berlalu lintas kepada masyarakat.
2. Bagi pemerintah daerah, hasil penelitian ini dapat dijadikan salah satu referensi dalam perencanaan pembangunan dan pemeliharaan infrastruktur jalan. Ruas jalan yang berada pada *Cluster* dengan tingkat risiko tinggi dapat diprioritaskan untuk dilakukan evaluasi terhadap kondisi geometrik jalan, permukaan jalan, penerangan jalan, maupun fasilitas keselamatan lainnya guna mengurangi potensi terjadinya kecelakaan.
3. Bagi peneliti selanjutnya, disarankan untuk menggunakan jumlah data yang lebih besar dan lebih mutakhir sehingga hasil *clustering* dapat menggambarkan kondisi kecelakaan lalu lintas secara lebih representatif. Selain itu, penelitian dapat menambahkan variabel lain, seperti volume lalu lintas, kecepatan kendaraan, tingkat kepadatan lalu lintas, kondisi kendaraan, maupun karakteristik pengemudi agar hasil analisis menjadi lebih komprehensif.

4. Hasil penelitian ini masih berupa pengelompokan berdasarkan karakteristik data kecelakaan. Oleh karena itu, penelitian selanjutnya dapat mengembangkan sistem berbasis Sistem Informasi Geografis (SIG) atau dashboard interaktif yang menampilkan persebaran lokasi kecelakaan berdasarkan hasil *clustering* sehingga informasi yang dihasilkan lebih mudah dipahami dan dimanfaatkan oleh pihak terkait dalam proses pengambilan keputusan.

5. REFERENSI

Agneresa, L. H., Hilabi, S. S., Hananto, A., & Tukino. (2022). *Strategi promosi penerapan data mining mahasiswa baru dengan metode K-Means clustering*. *Jurnal Manajemen dan Sistem Informasi*, *2*(2), 25–34.

Arriyanti, E., Arfyanti, I., & Adytia, P. (2019, September). Spatial coordinatetrial: Converting non-spatial data dimension for DBSCAN. In *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (pp. 223-228). IEEE.

1.

Baldah, A., Duarisah, A. V., & Maulana, R. A. (2023). *Clustering* daerah rawan bencana alam di Indonesia berdasarkan provinsi dengan metode K-means. *Jurnal Ilmiah Informatika Global*, *14*(2), 31–36.

Fatimah, K. N., Pratiwi, H., & Arriyanti, E. (2026). *Penerapan algoritma K-Means clustering dalam pengelompokan hasil belajar siswa di SMP Budi Luhur Samarinda*. *Jurnal Ilmiah STMIK Widya Cipta Dharma*, 1–10.

Hartono, B., & Lusiana, V. (2026). *Analisis metode Elbow SSE, Silhouette Score, dan Jaccard Stability dalam pemilihan jumlah klaster data yang optimal*. *TIN: Terapan Informatika Nusantara*, *6*(8), 1521–1532.

Hendrastuty, N. (2024). *Penerapan data mining menggunakan algoritma K-Means clustering dalam evaluasi hasil pembelajaran siswa*. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, *3*(1), 46–56.

Janner Lubis, D., & Tamam, M. B. (2022). *Penerapan K-Means untuk pengelompokan beasiswa santri di pondok pesantren Miftahul Huda Bogor*. *Jurnal Ilmiah Teknologi Informasi & Sains (TEKNOIS)*, *12*, 7–20.

Kurniawan, T., & Jajuli, M. (2022). *Clustering* data kecelakaan lalu lintas di Kecamatan Cileungsi menggunakan metode *K-means*. *Generation Journal*, *6*(1), 1–12.

Maori, N. A., & Evanita. (2023). *Metode Elbow dalam optimasi jumlah Cluster pada K-Means clustering*. *Jurnal SIMETRIS*, *14*(2), 277–287.

Maulana, M. A., & Subekti, A. T. (2024). *Analisis faktor penyebab kecelakaan lalu lintas pada kurir ekspedisi X di Kecamatan Bojong Kabupaten Tegal. BOHSEJ Bhamada (Bhamada Occupational Health Safety Environment Journal)*, *2*(1), 7–13.

Rahma, H. E., & Latifah, A. J. (2024). Analisis pola kecelakaan lalu lintas menggunakan metode clustering: Studi kasus Polresta Samarinda. *Jurnal Publikasi Teknik Informatika (JUPTI), 3*(1), 13–20. https://doi.org/10.55606/jupti.v3i1.2494

Rohaniyah, S., & Purnamasari, A. I. (2023). *Clustering* data pencari kerja menurut tingkat pendidikan menggunakan algoritma K-means. *JVP (Jurnal Minfo Polgan)*, *12*(1), 1–14.

Siregar, H. A., Azlan, & Gaol, N. Y. L. (2023). *Penerapan data mining pada penjualan rumah makan kasih ibu menggunakan metode K-Means clustering*. *Jurnal Sistem Informasi TGD*, *2*(5), 750–760.

Wijayanti, S., & Kesumawati, A. (2025). *Perbandingan analisis K-Means dan Hierarchical clustering dalam mengelompokkan kecamatan di Kabupaten Grobogan berdasarkan jumlah titik kejadian bencana alam Emerging Statistics and Data Science Journal*, *3*(2), 597–617.

Yasin, M., Garancang, S., & Hamzah, A. A. (2024). *Metode dan instrumen pengumpulan data (kualitatif dan kuantitatif). Banjarese: Journal of International Multidisciplinary Research*, *2*(3), 161–173.