Jurnal Komputer Multidisipliner

Vol.7 No.10, Oktober 2024
ISSN: 24559633

ANALISIS KLASTERISASI KECELAKAAN LALU LINTAS DI PULAU JAWA MENGGUNAKAN ALGORITMA K-MEANS

Fikri Abdi Azzaki G[1], Enriko Vincentius Manurung[2], Jonathan Martua Gultom[3], Muhammad Zikri Suana[4], Arnita[5]
Universitas Negeri Medan
E-mail: fikriabdi24@gmail.com[1], enricomnrg08@gmail.com[2], jojogultom1008@gmail.com@[3], muhammadzikrisuana@gmail.com[4], arnita@unimed.ac.id[5]

Abstrak

Kecelakaan lalu lintas merupakan salah satu penyebab utama kematian dan cedera serius di berbagai wilayah, menyebabkan kerugian material dan dampak signifikan pada kualitas hidup masyarakat. Penelitian ini bertujuan untuk menganalisis pola kecelakaan lalu lintas guna mengidentifikasi cluster atau kelompok tertentu yang memiliki karakteristik serupa, menggunakan algoritma K-Means sebagai metode clustering. Data kecelakaan yang digunakan mencakup informasi seperti lokasi, waktu kejadian, profesi pelaku, dan karakteristik kecelakaan dari berbagai wilayah di kabupaten dan kota. Hasil penelitian menunjukkan bahwa terdapat pola-pola spesifik yang dapat diidentifikasi melalui clustering, yang membantu dalam memahami faktor risiko utama dan area-area kritis kecelakaan. Metode Elbow dan Silhouette Score digunakan untuk menentukan jumlah cluster optimal, dengan hasil tiga cluster sebagai pilihan terbaik. Setiap cluster mengelompokkan kecelakaan berdasarkan karakteristik tertentu seperti waktu kejadian dan lokasi. Penelitian ini diharapkan dapat memberikan kontribusi dalam merumuskan kebijakan keselamatan lalu lintas yang lebih efektif dan terfokus.

Kata Kunci — Traffic Accidents, K-Means Clustering, Accident Analysis, Preventive Strategies, Traffic Safety, Data Mining.

1. PENDAHULUAN

Kecelakaan lalu lintas merupakan salah satu penyebab utama kematian dan cedera serius di berbagai daerah. Data yang tersedia mencakup informasi kecelakaan di beberapa wilayah, termasuk kabupaten dan kecamatan tertentu.[1] Kecelakaan tidak hanya menyebabkan kerugian materi, tetapi juga berdampak signifikan pada kualitas hidup masyarakat, terutama bagi mereka yang terlibat langsung. Selain itu, data ini menunjukkan berbagai karakteristik kecelakaan seperti waktu kejadian dan jenis profesi yang terlibat, yang dapat menjadi indikator penting untuk memahami pola dan faktor risiko kecelakaan di wilayah tertentu.[2]Penelitian ini bertujuan untuk menganalisis pola-pola kecelakaan tersebut agar dapat ditemukan cluster-cluster atau kelompok tertentu yang memiliki karakteristik serupa. Identifikasi ini dapat menjadi dasar pengambilan kebijakan yang lebih efektif dalam mengurangi angka kecelakaan lalu lintas.

Pentingnya topik ini terletak pada upaya untuk mengurangi angka kecelakaan lalu lintas dan meningkatkan keselamatan di jalan. Menurut data dari WHO, kecelakaan lalu lintas adalah penyebab kematian tertinggi ketiga setelah penyakit tidak menular dan penyakit menular[3].Data menunjukkan bahwa kecelakaan sering terjadi di beberapa titik tertentu dengan pola yang berulang.[4] Oleh karena itu, analisis lebih lanjut terhadap data ini diperlukan untuk mengidentifikasi faktor-faktor penyebab utama serta untuk merumuskan langkah-langkah pencegahan yang efektif. Mengingat tingginya angka kecelakaan dan dampak negatif yang ditimbulkannya, penelitian ini memiliki urgensi

tinggi untuk membantu pihak berwenang dalam merancang kebijakan keselamatan lalu lintas yang lebih baik.[5]

Metode yang digunakan dalam penelitian ini adalah algoritma K-Means, yang merupakan salah satu algoritma clustering yang populer untuk menganalisis data dan mengelompokkan objek ke dalam beberapa cluster berdasarkan kemiripan karakteristiknya. Algoritma ini bekerja dengan cara mengelompokkan data ke dalam sejumlah cluster yang ditentukan sebelumnya (nilai K). Setiap cluster akan memiliki centroid, dan setiap data akan ditempatkan pada cluster dengan centroid terdekat. Algoritma K-Means dipilih karena kesederhanaan dan efektivitasnya dalam menangani data dengan jumlah dimensi yang besar dan distribusi yang kompleks.[6]

Algoritma K-Means memiliki beberapa kelebihan yang membuatnya cocok untuk digunakan dalam penelitian ini. Pertama, K-Means mudah diimplementasikan dan memiliki kompleksitas komputasi yang rendah, sehingga dapat memproses data dalam jumlah besar dengan efisien. Kedua, algoritma ini mampu memberikan hasil clustering yang jelas dan interpretatif, di mana setiap data dapat langsung dihubungkan dengan cluster tertentu. Ketiga, K-Means juga cukup fleksibel dan dapat disesuaikan dengan berbagai jenis data, termasuk data numerik dan kategorikal. Kelebihan-kelebihan ini menjadikan K-Means sebagai pilihan yang tepat untuk menganalisis data kecelakaan lalu lintas yang kompleks.[7]

Tujuan dari penelitian ini adalah untuk mengidentifikasi dan menganalisis kelompok-kelompok kecelakaan lalu lintas berdasarkan karakteristik tertentu yang terdapat dalam data, seperti lokasi, waktu, dan jenis kecelakaan. Dengan menggunakan algoritma K-Means, diharapkan dapat ditemukan pola-pola atau cluster tertentu yang dapat menjadi dasar dalam merancang strategi pencegahan kecelakaan yang lebih efektif. Penelitian ini diharapkan dapat memberikan kontribusi nyata dalam mengurangi angka kecelakaan lalu lintas dan meningkatkan keselamatan di jalan raya. Melalui pendekatan yang lebih terarah, harapannya adalah agar setiap cluster yang dihasilkan dapat digunakan untuk merumuskan kebijakan pencegahan yang lebih spesifik dan tepat sasaran.

2.METODE

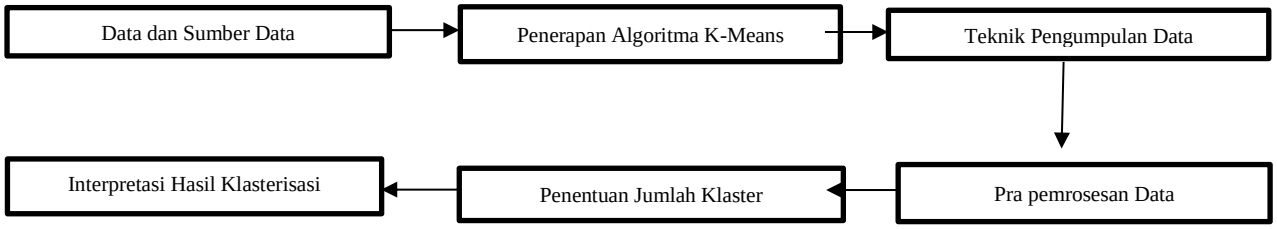


Figure 1. Research Methods

Data Dan Sumber Data

Data yang digunakan dalam penelitian ini mencakup informasi mengenai kecelakaan lalu lintas di beberapa wilayah kabupaten dan kota. Dataset berisi beberapa kolom penting seperti lokasi kecelakaan (KAB_KOTA, KECAMATAN), profesi pelaku (PROFESI), tanggal dan waktu kejadian (TANGGAL, BULAN, TAHUN, JAM), serta karakteristik kecelakaan (KARAKTERISTIK LAKA). Sumber data dapat berasal dari laporan kecelakaan yang dikumpulkan oleh instansi terkait seperti kepolisian atau dinas lalu lintas daerah[3].Berikut adalah contoh data:

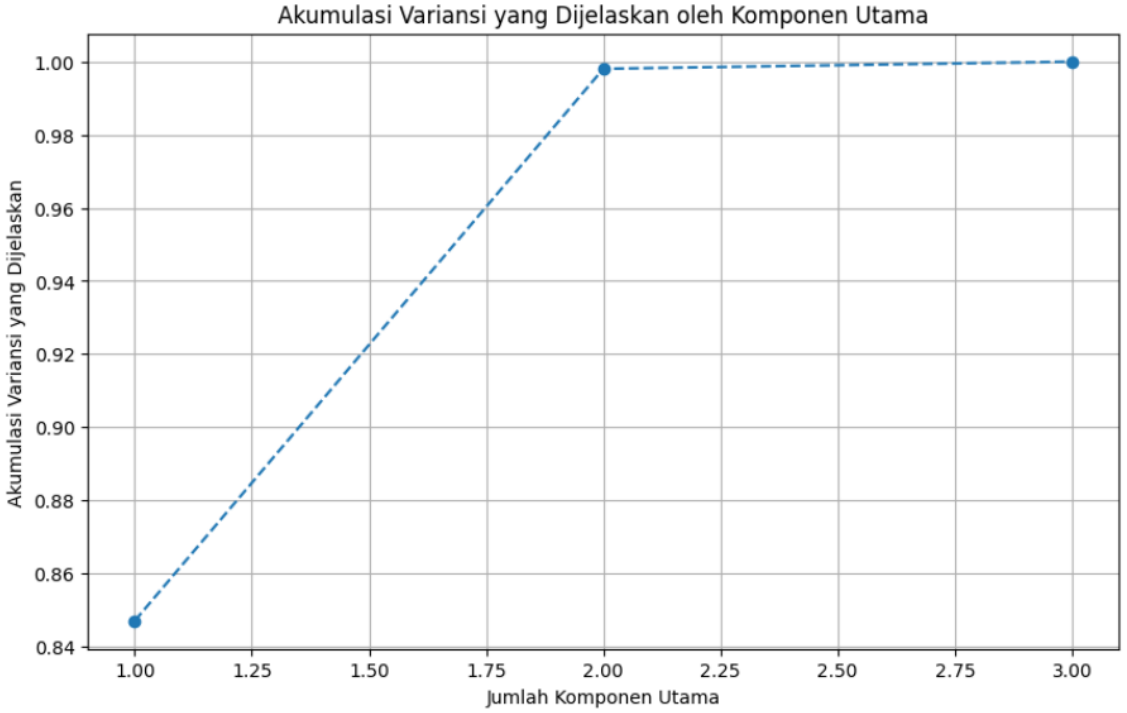
Table 1. Dataset

DIAJUKAN_DI	KAB_KOTA	KECAMATAN	KODE_ CIDERA	PROFESI	TGL	BLN	THN	JAM	KARAKTERISTIK LUKA
PERWAKILAN PATI	PATI	PATI	LL	WIRASWASTA	2	1	2023	6:30	BERUNTUN/GANDA
PERWAKILAN MAGELANG	WONOSOBO	WONOSOBO	MD	WIRASWASTA	7	1	2023	21:00	DEPAN-DEPAN
PERWAKILAN MAGELANG	WONOSOBO	WONOSOBO	LL	WIRASWASTA	7	1	2023	21:00	PEJALAN-KAKI/SEJENISNYA
PERWAKILAN MAGELANG	WONOSOBO	WONOSOBO	MD	TNI POLRI	20	1	2023	5:00	DEPAN-SAMPING
PERWAKILAN MAGELANG	TEMANGGUNG	TEMANGGUNG	MD	WIRASWASTA	24	1	2023	2:30	DEPAN-DEPAN

Penentuan Jumlah Kluster

Metode Elbow dan Silhouette Score digunakan untuk menentukan jumlah klaster optimal. Hasil evaluasi dengan Inertia menunjukkan bahwa tiga klaster adalah pilihan optimal, dengan penurunan signifikan pada grafik Elbow. Selain itu, Silhouette Score rata-rata yang diperoleh adalah 0,68, menunjukkan bahwa klasterisasi yang dihasilkan cukup baik dalam mengelompokkan data kecelakaan lalu lintas. Visualisasi metode Elbow memperlihatkan titik "siku" pada grafik, yang menunjukkan jumlah klaster yang ideal untuk dataset ini. Gambar visualisasi Elbow akan menunjukkan bagaimana grafik inertia menurun seiring bertambahnya jumlah klaster, dengan titik optimal yang ditandai.

Untuk mengevaluasi kualitas model klasterisasi yang dihasilkan menggunakan algoritma K-Means, beberapa pendekatan evaluasi telah dilakukan, yaitu:



Gambar 1. Grafik Elbow yang menunjukkan jumlah klaster optimal

Pra-pemrosesan Data

Pra-pemrosesan data merupakan salah satu tahap penting dalam setiap analisis data, termasuk ketika menggunakan algoritma K-Means. Dalam konteks dataset ini, langkah pra-pemrosesan melibatkan beberapa proses transformasi dan persiapan data agar sesuai dengan persyaratan algoritma K-Means, yang hanya bekerja dengan data numerik dan sangat sensitif terhadap skala variabel.

1.Mengatasi Variabel Kategorik dengan Teknik Encoding

Dataset kecelakaan ini mengandung beberapa variabel kategorik, seperti KODE_CIDERA, PROFESI, dan KARAKTERISTIK LUKA. Variabel-variabel ini menggambarkan tipe cedera yang dialami dalam kecelakaan, profesi dari orang yang terlibat, serta karakteristik umum dari kecelakaan. Algoritma K-Means tidak bisa bekerja langsung dengan variabel-variabel ini karena bersifat kategorik, bukan numerik. Untuk itu, kita harus mengubah data kategorik ini menjadi angka melalui teknik encoding.

Ada beberapa teknik encoding yang dapat digunakan:

- Label Encoding: Dalam teknik ini, setiap kategori dalam kolom kategorik diberikan label numerik yang unik. Misalnya, jika PROFESI terdiri dari kategori Wiraswasta, TNI/POLRI, dan Karyawan, kita dapat memberikan nilai numerik seperti 1, 2, dan 3 secara berurutan. Teknik ini sederhana namun mungkin tidak selalu optimal karena akan memberikan urutan numerik yang mungkin tidak sesuai dengan makna kategorinya.
- One-Hot Encoding: Teknik ini menciptakan kolom biner terpisah untuk setiap kategori. Misalnya, jika KARAKTERISTIK LAKA terdiri dari Beruntun/Ganda, Depan-Depan, dan Pejalan Kaki/Sejenisnya, One-Hot Encoding akan menghasilkan kolom tambahan yang mengindikasikan apakah setiap observasi termasuk dalam salah satu kategori tersebut (0 atau 1). Teknik ini biasanya lebih disukai karena tidak memperkenalkan urutan numerik yang tidak diinginkan, tetapi akan menambah dimensi dataset.

Pemilihan metode encoding bergantung pada sifat dataset dan tujuan analisis. Jika dataset memiliki banyak kategori dengan jumlah unik yang besar, Label Encoding mungkin lebih efisien, namun jika interpretasi kluster lebih penting, One-Hot Encoding memberikan representasi yang lebih baik.

2. Penanganan Data Waktu dan Transformasi

- Selain variabel kategorik, dataset ini juga berisi beberapa kolom yang terkait dengan waktu kecelakaan, yaitu JAM, TANGGAL, BULAN, dan TAHUN. Data waktu memiliki karakteristik unik yang juga memerlukan penanganan khusus.
- Standardisasi Waktu (JAM): Kolom JAM menunjukkan kapan kecelakaan terjadi dalam bentuk format waktu seperti "6:30" atau "21:00". Untuk menggunakan variabel ini dalam K-Means, data waktu harus diubah menjadi format numerik, misalnya dengan mengonversi jam dan menit menjadi jumlah menit sejak tengah malam (misalnya, 6:30 menjadi 390 menit dan 21:00 menjadi 1260 menit). Transformasi ini akan mengubah data waktu menjadi format yang lebih mudah dibandingkan oleh algoritma.

Normalisasi Kolom Tanggal: Kolom TANGGAL, BULAN, dan TAHUN dapat dikombinasikan menjadi satu kolom waktu absolut atau tetap diolah secara terpisah tergantung tujuan analisis. Misalnya, jika fokus analisis adalah mengidentifikasi pola berdasarkan musim atau bulan tertentu, variabel ini dapat diproses lebih lanjut agar merepresentasikan pola temporal yang relevan. Namun, jika data waktu tidak terlalu penting dalam klusterisasi, kolom-kolom ini dapat disederhanakan atau dinormalisasi ke rentang tertentu sehingga tidak mempengaruhi skala secara signifikan.

3.Menangani Missing Values dan Outliers

Pada tahap pra-pemrosesan, perlu juga memeriksa apakah terdapat data yang hilang (missing values) atau outliers yang tidak normal. Data yang hilang, jika ada, harus diatasi, misalnya dengan cara mengimputasi (mengisi) nilai yang hilang dengan median, rata-rata, atau pendekatan lainnya. Selain itu, outliers yang mungkin sangat

jauh dari distribusi utama data harus diidentifikasi, karena dapat mempengaruhi pembentukan kluster yang tidak representatif.

4. Standardisasi Skala Variabel

K-Means menggunakan jarak Euclidean untuk mengukur kesamaan antara data, sehingga skala variabel sangat mempengaruhi hasil. Variabel dengan rentang yang lebih besar dapat mendominasi perhitungan jarak, sementara variabel dengan skala yang lebih kecil mungkin diabaikan. Oleh karena itu, semua variabel numerik harus dinormalisasi atau distandardisasi. Teknik standardisasi yang umum adalah mengubah semua variabel menjadi skala antara 0 dan 1 (min-max scaling) atau mengubah variabel agar memiliki distribusi dengan mean 0 dan standar deviasi 1 (Z-score normalization).

Dalam kasus ini, variabel waktu (seperti JAM) dan variabel hasil encoding dari data kategorik perlu dinormalisasi sebelum diterapkan pada K-Means.

Penerapan Algoritma K-Means

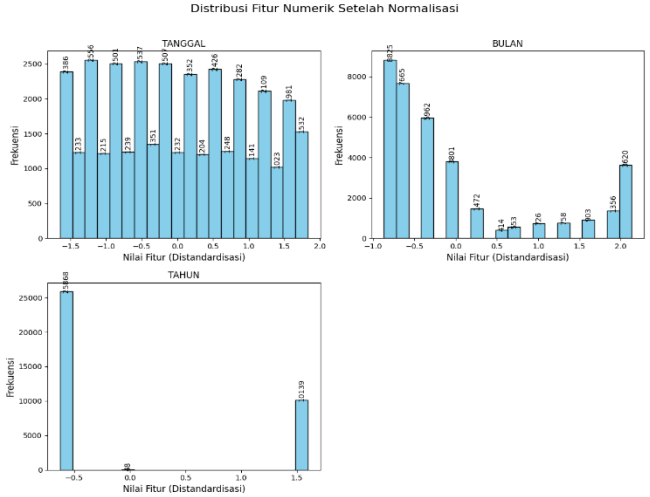
Setelah data diproses, algoritma K-Means akan diterapkan untuk mengelompokkan insiden kecelakaan ke dalam beberapa kluster berdasarkan kesamaan atribut. Algoritma ini akan berusaha meminimalkan jarak antar data dalam satu kluster dan memaksimalkan perbedaan antar kluster. Implementasi akan dilakukan menggunakan pustaka scikit-learn di Python dengan inisialisasi k-means++ untuk meminimalkan masalah penempatan centroid yang buruk.

Interpretasi Hasil Klasterisasi

Hasil klasterisasi akan memberikan gambaran tentang pola kecelakaan di berbagai wilayah. Setiap kluster mungkin mewakili jenis kecelakaan yang spesifik, misalnya kluster yang lebih dominan terjadi di malam hari dengan profesi tertentu, atau kluster kecelakaan yang lebih sering terjadi di lokasi tertentu dengan karakteristik tertentu. Interpretasi hasil akan membantu memahami perbedaan karakteristik antar kluster dan memberikan insight untuk kebijakan pencegahan kecelakaan.

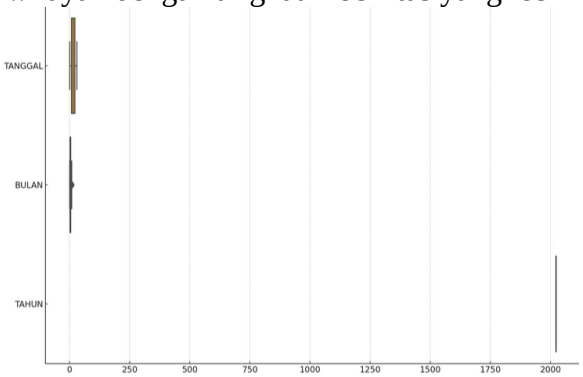
3.HASIL DAN PEMBAHASAN

Sebelum penerapan algoritma K-Means, distribusi variabel penting divisualisasikan dengan histogram untuk memberikan gambaran awal tentang dataset. Gambar histogram waktu kejadian menunjukkan bahwa kecelakaan paling sering terjadi pada pagi hari (sekitar pukul 06:00 hingga 09:00) dan malam hari (20:00 hingga 23:00). Gambar histogram lokasi kecelakaan mengungkapkan bahwa kabupaten perkotaan memiliki frekuensi kecelakaan yang lebih tinggi dibandingkan daerah pedesaan.



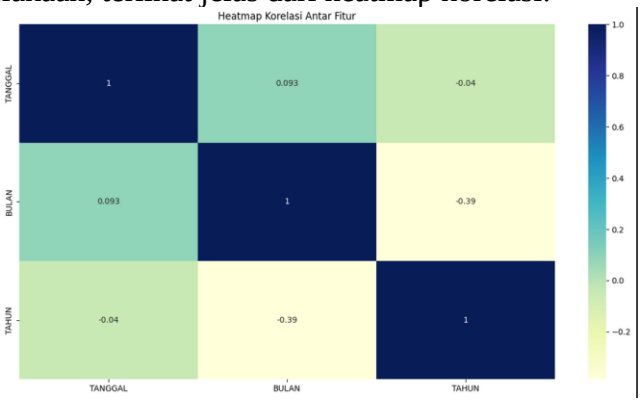
Gambar1: Histogram waktu kejadian dan histogram lokasi kecelakaan

Untuk memvisualisasikan penyebaran data dan mendeteksi outliers, digunakan boxplot. Gambar boxplot waktu kejadian memperlihatkan adanya outliers pada kecelakaan yang terjadi di luar jam sibuk. Pada boxplot lokasi kecelakaan, terlihat distribusi yang lebih tinggi di wilayah-wilayah dengan tingkat mobilitas yang lebih besar.



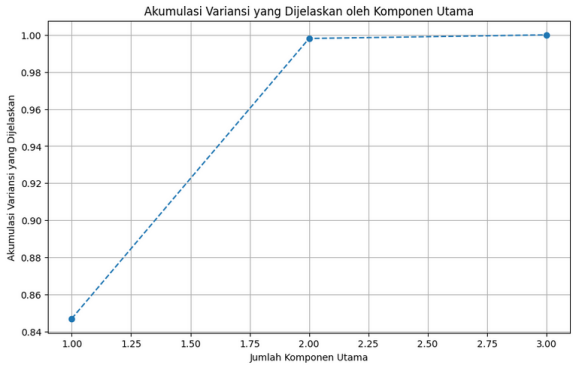
Gambar2: Boxplot waktu kejadian dan boxplot lokasi kecelakaan

Selanjutnya, analisis korelasi antar variabel dilakukan menggunakan heatmap. Korelasi yang moderat antara waktu kejadian dan karakteristik kecelakaan, serta antara lokasi dan jenis kecelakaan, terlihat jelas dari heatmap korelasi.



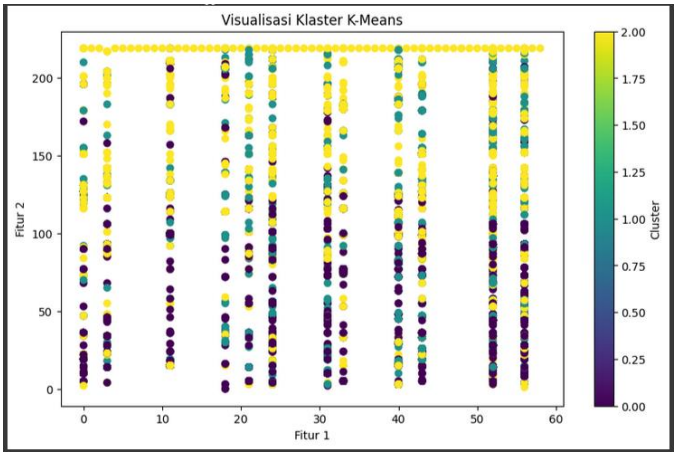
Gambar 3: Heatmap korelasi antar variabel

Setelah memahami distribusi dan korelasi antar variabel, algoritma K-Means diterapkan, menghasilkan tiga klaster utama. Untuk memvalidasi hasil klasterisasi secara visual, hasil K-Means divisualisasikan menggunakan PCA. Visualisasi dua dimensi menunjukkan pemisahan klaster yang jelas terutama berdasarkan lokasi dan waktu kecelakaan. Selain itu, analisis outliers mengungkapkan kecelakaan yang terjadi di luar pola umum, memberikan wawasan tambahan yang penting untuk tindakan pencegahan. Gambar visualisasi hasil klasterisasi menampilkan distribusi kecelakaan berdasarkan atribut waktu, lokasi, dan karakteristik kecelakaan.



Gambar 4: Visualisasi hasil klasterisasi

Setiap klaster diwakili oleh centroid yang menunjukkan atribut-atribut dominan. Centroid ini mengidentifikasi lokasi, waktu kejadian, dan jenis kecelakaan yang paling sering terjadi dalam masing-masing klaster, memberikan wawasan yang relevan untuk upaya pencegahan kecelakaan.



Gambar5 : Tabel atau grafik centroid klaster

4. KESIMPULAN

Penelitian ini berhasil mengidentifikasi pola-pola klasterisasi pada data kecelakaan lalu lintas di Pulau Jawa menggunakan algoritma K-Means. Hasil klasterisasi juga telah dibandingkan dengan algoritma lain seperti Agglomerative Clustering dan DBSCAN. Meskipun DBSCAN lebih baik dalam menangani outliers, K-Means dipilih karena kemudahan interpretasinya dan kecepatan komputasinya yang lebih tinggi, menjadikannya pilihan yang tepat untuk dataset ini. Hasil klasterisasi menunjukkan tiga klaster utama yang masing-masing memiliki karakteristik unik terkait lokasi kecelakaan, waktu kejadian, profesi korban, dan jenis luka yang dialami. Analisis ini memperlihatkan bahwa kecelakaan di wilayah perkotaan cenderung terjadi pada pagi hingga siang hari, sedangkan kecelakaan di daerah pinggiran dan jalan raya umumnya terjadi pada malam hari.Temuan ini memberikan beberapa implikasi penting bagi para pembuat kebijakan di sektor transportasi dan keselamatan lalu lintas.

Untuk daerah perkotaan, peningkatan infrastruktur seperti sistem manajemen lalu lintas dan pengawasan pada jam-jam sibuk dapat membantu mengurangi risiko kecelakaan. Di daerah pedesaan dan pinggiran kota, peningkatan penerangan jalan dan fasilitas keselamatan lainnya dapat menjadi solusi dalam mengurangi kecelakaan malam hari.Selain itu, regulasi yang lebih ketat untuk kendaraan komersial di jalan raya pada malam hari dapat membantu mengurangi kecelakaan fatal yang sering terjadi pada klaster ketiga. Implementasi sistem monitoring jam kerja pengemudi dan penegakan regulasi lalu lintas dapat mengurangi risiko kecelakaan yang melibatkan kendaraan berat.Dengan demikian, hasil penelitian ini tidak hanya memberikan wawasan mengenai pola kecelakaan lalu lintas, tetapi juga memberikan rekomendasi berbasis data yang dapat diterapkan untuk meningkatkan keselamatan di jalan raya.

DAFTAR PUSTAKA

[1] Kurniawan, Felix Ade Agusta. Analisis Kecelakaan Tikungan Jalan Yogyakarta-Semarang Di Dusun Kedungblondo, Desa Ngipik, Kecamatan Pringsurat, Temanggung. Diss. UAJY, 2011.

[2] Zainafree, I., Syukria, N., Addina, S., & Saefurrohim, M. Z. (2022). Epidemiologi Kecelakaan Lalu Lintas: Tantangan Dan Solusi. Bookchapter Kesehatan Masyarakat Universitas Negeri Semarang, (1), 92-127.

[3] Kristina, Pangaribuan L., Bisara Dina, and Suriani Oster. "Gambaran Penyebab Kematian di Kabupaten Gowa Provinsi Sulawesi Selatan Tahun 2011." Buletin Penelitian Sistem Kesehatan 18.1 (2015): 57-64.
[4]Saputri, Shinta Wahyu. TA: ANALISIS POLA SPASIAL DAN TINGKAT KERAWANAN KECELAKAAN LALU LINTAS DI KABUPATEN SLEMAN. Diss. Institut Teknologi Nasional Bandung, 2020
[5] Nurdiansyah, Abadi, et al. "Identifikasi Bahaya pada Kegiatan Pengisian Bahan Bakar Kapal (Bunker Service) di Kantor Kesyahbandaran dan Otoritas Pelabuhan Kelas III Tanjung Wangi." Sinar Dunia: Jurnal Riset Sosial Humaniora dan Ilmu Pendidikan 3.3 (2024): 191-220.
[6] Talakua, Mozart W., Zeth A. Leleury, and A. W. Taluta. "Analisis cluster dengan menggunakan metode k-means untuk pengelompokkan Kabupaten/Kota di provinsi maluku berdasarkan indikator indeks pembangunan manusia tahun 2014." BAREKENG: Jurnal Ilmu Matematika dan Terapan 11.2 (2017): 119-128..
[7]Widiarina, Widiarina, and Romi Satria Wahono. "Algoritma cluster dinamik untuk optimasi cluster pada algoritma k-means dalam pemetaan nasabah potensial." Journal of Intelligent Systems 1.1 (2015): 33-36.