Clustering Data Kecelakaan Lalu Lintas melalui Algoritma *K-Means* dengan Seleksi Fitur *Chi-Square*

Adellia Putri Margaretha[1], Nurissaidah Ulinnuha[2], Putroue Keumala Intan[3]
[1]Jurusan Matematika, Fakultas Sains dan Teknologi, UIN Sunan Ampel, Surabaya
[2]Jurusan Matematika, Fakultas Sains dan Teknologi, UIN Sunan Ampel, Surabaya
[3]Jurusan Matematika, Fakultas Sains dan Teknologi, UIN Sunan Ampel, Surabaya
*Email: 09010221001@student.uinsa.ac.id

Abstract. *Traffic accidents are a significant problem in Indonesia, with fatalities and huge economic losses. This study aims to apply the K-means algorithm to cluster traffic accident data using feature selection. The traffic accident data used was obtained from an accident insurance company in Sidoarjo and processed to generate relevant features. The feature selection process is carried out to determine the features that have importance and the most relevant information in the clustering process. The feature selection method used in this research is Chi-Square feature selection, which aims to select features that have a significant relationship with the target accident variable. The results show that the data is divided into 2 clusters with feature selection and without feature selection, namely areas with high and low accident rates. The value of the cluster silhouette coefficient before feature selection is 0.57. While after applying feature selection with Chi-Square, a better result is obtained, which is 0.72. This research shows that applying the feature selection method can improve the performance of clustering traffic accident data with the K-means algorithm.*

Keywords: *Chi-Square, Clustering, Feature Selection, K-Means, Traffic accidents*

Abstrak. *Kecelakaan lalu lintas merupakan permasalahan signifikan di Indonesia, dengan dampak fatal dan kerugian ekonomi yang besar. Penelitian ini bertujuan untuk menerapkan algoritma K-means untuk mengelompokkan data kecelakaan lalu lintas dengan menggunakan seleksi fitur. Data kecelakaan lalu lintas yang digunakan diperoleh dari sebuah perusahaan asuransi kecelakaan di Sidoarjo dan diproses untuk menghasilkan fitur-fitur yang relevan. Proses seleksi fitur dilakukan untuk menentukan fitur-fitur yang memiliki kepentingan dan informasi yang paling relevan dalam proses pengelompokkan. Metode seleksi fitur yang digunakan dalam penelitian ini adalah seleksi fitur Chi-Square, yang bertujuan untuk memilih fitur-fitur yang memiliki hubungan signifikan dengan variabel target kecelakaan. Hasil penelitian menunjukkan bahwa data terbagi menjadi 2 cluster dengan seleksi fitur maupun tanpa seleksi fitur, yaitu wilayah dengan tingkat kecelakaan tinggi dan rendah. Nilai koefisien silhouette cluster sebelum dilakukan seleksi fitur adalah sebesar 0,57. Sedangkan setelah diterapkan seleksi fitur dengan Chi-Square, diperoleh hasil yang lebih baik yaitu sebesar 0,72. Penelitian ini menunjukkan bahwa dengan menerapkan metode seleksi fitur dapat meningkatkan performa pengelompokkan data kecelakaan lalu lintas dengan algoritma K-means.*

Kata Kunci: *Chi-Square, Clustering, K-Means, Kecelakaan lalu lintas, Seleksi Fitur*

1. Pendahuluan

1.1. Latar Belakang

Seiring dengan pesatnya pertumbuhan populasi Indonesia, jumlah kendaraan di setiap daerah mengalami lonjakan yang signifikan. Fenomena ini berimbas pada padatnya jalanan dan ramainya lalu lintas, menciptakan situasi yang penuh sesak dan berisiko tinggi. Tak heran, statistik kecelakaan di Indonesia pun menunjukkan tren peningkatan yang memprihatinkan (Gurning et al., 2024). Kecelakaan

lalu lintas merujuk pada insiden yang terjadi dalam arus lalu lintas karena kesalahan yang terjadi pada komponen-komponen sistem lalu lintas, seperti pengemudi, kendaraan, infrastruktur jalan, dan lingkungan sekitarnya. Definisi kecelakaan lalu lintas berkaitan dengan situasi di mana kondisinya tidak memenuhi standar keselamatan atau mencapai tingkat keparahan yang diharapkan (Siregar et al., 2022).

Kecelakaan lalu lintas dapat mengakibatkan kerusakan kendaraan seperti benturan, rusakan, atau hancuran kendaraan. Kerugian yang diakibatkan kecelakaan lalu lintas antara lain adalah kerusakan kendaraan, biaya perbaikan kendaraan, biaya pengobatan, kerugian waktu, dan kerugian harga barang (J.W.Kaawoan, 2023). Dampak ini semakin parah jika melibatkan korban jiwa atau cedera serius. Biaya perbaikan kendaraan yang diakibatkan kecelakaan lalu lintas dapat menjadi kerugian besar, terutama jika kendaraan yang rusak adalah kendaraan yang baru atau masih berwujud pinjaman. Biaya pengobatan juga dapat menjadi kerugian besar, terutama jika kecelakaan lalu lintas memakan korban. Kerugian waktu yang diakibatkan kecelakaan lalu lintas dapat menjadi kerugian besar, terutama jika kecelakaan lalu lintas memakan korban waktu kerja atau kegiatan lainnya. Kerugian harga barang yang diakibatkan kecelakaan lalu lintas dapat menjadi kerugian besar, terutama jika kecelakaan lalu lintas menyebabkan hilangnya barang atau peralatan yang diperlukan (Siregar et al., 2022).

Melihat kerugian-kerugian yang timbul dari kecelakaan lalu lintas, penelitian ini difokuskan pada wilayah Sidoarjo sebagai area studi. Sidoarjo merupakan salah satu wilayah di Jawa Timur dengan tingkat kecelakaan lalu lintas yang tinggi. Pada tahun 2023, tercatat 2.102 kecelakaan lalu lintas di Sidoarjo, dengan 287 korban meninggal dunia (Yusuf Haafidh Nur Siddiq, 2024). Angka ini menunjukkan bahwa kecelakaan lalu lintas masih menjadi permasalahan serius di Sidoarjo. Kecelakaan lalu lintas tidak hanya mengakibatkan kerugian jiwa dan cedera pada korban, tetapi juga menimbulkan dampak ekonomi dan sosial yang besar bagi masyarakat dan pemerintah setempat. Kerugian ini mencakup biaya perawatan medis, kerusakan kendaraan dan infrastruktur, hingga penurunan produktivitas kerja (sitasi). Oleh karena itu, upaya untuk memahami pola-pola kecelakaan secara lebih mendalam menjadi hal yang krusial dalam merancang strategi pencegahan yang efektif. Dalam konteks ini, penerapan algoritma *K-Means Clustering* untuk menganalisis dan mengelompokkan data kecelakaan lalu lintas menjadi sangat penting. Dengan mengelompokkan data berdasarkan karakteristik tertentu seperti lokasi kejadian, waktu, jenis kendaraan, dan tingkat keparahan *K-Means* dapat membantu mengidentifikasi pola dan wilayah rawan kecelakaan secara lebih sistematis. Hasil klaster ini tidak hanya memberikan wawasan yang berguna bagi pengambil kebijakan, tetapi juga mendukung alokasi sumber daya yang lebih tepat sasaran dalam upaya peningkatan keselamatan di jalan raya.

K-Means adalah metode *clustering* partisi yang memisahkan data ke dalam k wilayah terpisah. Metode ini menggunakan algoritma iteratif untuk mencari *centroid cluster* yang terbaik dan mengelompokkan data ke *cluster* yang sesuai (Luh et al., 2022). Proses *clustering* bergantung pada jumlah data yang akan dikelompokkan, jumlah *cluster* yang diinginkan, dan jumlah fitur yang digunakan. Jumlah data yang besar dapat membuat proses *clustering* lebih lama dan membutuhkan lebih banyak komputasi. Jumlah *cluster* yang diinginkan juga dapat mempengaruhi panjang proses *clustering*, karena jumlah *cluster* yang lebih banyak dapat membutuhkan lebih banyak iterasi untuk mencari *centroid cluster* yang terbaik (Amalia & Arianto, 2024). Kelebihan dari metode K-Means antara lain sederhana dan mudah diterapkan, kecepatan dan efisiensi dalam menemukan *cluster*, adaptabilitas yang baik terhadap berbagai jenis data, serta fleksibilitas dalam penentuan jumlah *cluster* (Hendrawan et al., 2023). Metode *K-Means Clustering* digunakan untuk mengelompokkan data kecelakaan lalu lintas ke dalam beberapa *cluster* berdasarkan karakteristiknya. Data kecelakaan lalu lintas dipisahkan menjadi beberapa set data, seperti dataset 1 dan dataset 2. Setelah proses pengelompokkan, dilakukan pengujian menggunakan koefisien *silhouette* untuk menentukan *cluster* dengan kualitas terbaik (Vernanda et al., 2021). Metode seleksi fitur diterapkan sebelum proses *clustering* dilakukan pada penelitian ini. Penerapan metode seleksi fitur ini bertujuan untuk membandingkan performa yang diperoleh sebelum data diproses dengan seleksi fitur dan setelah data diproses tanpa seleksi fitur.

Urgensi penelitian ini semakin menguat karena masih terbatasnya studi yang secara eksplisit menggabungkan metode seleksi fitur dengan klasterisasi pada kasus kecelakaan lalu lintas secara spasial, khususnya di Sidoarjo. Penelitian terdahulu lebih banyak memfokuskan penggunaan K-Means untuk bidang-bidang lain. Penelitian sebelumnya yang menggunakan metode *K-Means* untuk menganalisis faktor penyebab *stunting* pada balita, menunjukkan bahwa metode *silhouette* efektif dalam

membantu proses *clustering* dengan algoritma K-Means. Penelitian tersebut menghasilkan 3 *cluster* optimum dengan nilai *silhouette* sebesar 0,37 (Amalia & Arianto, 2024). Penelitian lainnya menggunakan K-Means untuk membentuk *cluster* data siswa berdasarkan nilai sikap, disiplin, dan akademik. Pengujian menggunakan *elbow method* dan *silhouette coefficient* menghasilkan 3 *cluster* optimum dengan nilai *silhouette* sebesar 0,489 (Yudhistira & Andika, 2023). Berdasarkan penelitian terdahulu, K-Means termasuk metode clustering yang populer dan mendapatkan evaluasi hasil cluster yang cukup baik. Rahmansyah membandingkan metode *K-Means* dengan *Fuzzy C-Means* pada pengelompokan 63 Puskesmas di Surabaya berdasarkan data gizi 148,720 balita. Hasilnya menunjukkan bahwa 4.2% balita mengalami kekurangan gizi, 0.1% sangat kekurangan gizi, dan 4.5% mengalami *stunting*. Algoritma *K-Means* dengan normalisasi menghasilkan nilai *silhouette coefficient* terbaik sebesar 0.518, sedangkan *Fuzzy C-Means* dengan normalisasi mencapai 0.497. Hal ini menunjukkan *K-Means* lebih efektif dalam melakukan *clustering* (Rahmansyah et al., 2023).

Penelitian lain menggunakan metode *Chi Square* untuk mereduksi atribut yang kurang relevan dalam dataset, yang kemudian diklasifikasikan menggunakan *decision tre*e C4.5, pada data *South Germany Credit* yang berisi 1000 data dan 20 atribut. Hasil pengujian menunjukkan bahwa penerapan *Chi Square* meningkatkan akurasi *decision tree* C4.5 dengan rata-rata peningkatan sebesar 2.5% (Kandida Br et al., 2023). Penelitian selanjutnya membandingkan metode Chi-Square dengan beberapa metode optimasi lainnya, seperti Correlation Based Feature, Information Gain, dan ANOVA, pada data performa akademik mahasiswa. Hasil penelitian ini mengungkapkan bahwa Chi-Square memiliki nilai tertinggi dalam meningkatkan akurasi, dengan peningkatan sebesar 2,45 (Priantama et al., 2022).

Seleksi fitur dapat membantu mengurangi fitur-fitur yang tidak relevan dan mempermudah proses analisis data (Septianingrum et al., 2021). Ini dapat membantu untuk menghapus faktor-faktor atau variabel yang terlalu banyak dan kurang relevan terhadap data penelitian. Salah satu metode penyeleksian fitur yang dapat diterapkan adalah *Chi-Square*. Penerapan algoritma K-means untuk melakukan clustering data kecelakaan lalu lintas, disertai dengan seleksi fitur, menjadikan penelitian ini memiliki urgensi yang tinggi. Seleksi fitur bertujuan untuk mengidentifikasi fitur-fitur yang memiliki hubungan signifikan dengan variabel target dan mempertahankan fitur-fitur tersebut, sementara fitur-fitur yang kurang relevan atau terlalu banyak noise dapat dihilangkan (Iswanto et al., 2021). Penelitian ini bertujuan untuk klasterisasi data kecelakaan lalu lintas di Sidoarjo menggunakan *K-Means* yang dikombinasikan dengan seleksi fitur berbasis *Chi-Square*, serta membandingkannya dengan model tanpa seleksi fitur. Penelitian ini tidak hanya menghasilkan klaster wilayah rawan kecelakaan yang lebih representatif, tetapi juga memberikan landasan bagi perumusan kebijakan intervensi lalu lintas yang lebih fokus dan efisien. Dengan identifikasi pola kecelakaan secara lebih akurat, strategi pencegahan seperti pengaturan lalu lintas, edukasi keselamatan, dan alokasi sumber daya dapat dirancang secara lebih efektif untuk menekan angka kecelakaan di Sidoarjo.

2. Tinjauan Pustaka

2.1. Kecelakaan Lalu Lintas

Kecelakaan lalu lintas adalah insiden yang tidak diharapkan yang terjadi di jalan raya dan melibatkan kendaraan bermotor dan sering menyebabkan kerugian materi atau nyawa. Kecelakaan dapat dipengaruhi oleh berbagai faktor, yaitu faktor manusia seperti ketidakmampuan pengemudi dan pelanggaran peraturan lalu lintas, faktor kendaraan seperti kondisi teknis yang buruk dan kurangnya fitur keselamatan, faktor lingkungan seperti kondisi jalan yang rusak dan cuaca buruk, faktor sosial seperti budaya penggunaan kendaraan yang tidak aman dan kurangnya kesadaran akan keselamatan, serta faktor teknis seperti sistem penindakan pelanggaran yang tidak efektif dan infrastruktur jalan yang kurang memadai (Sidiq et al., 2023).

2.2. Normalisasi Data

Normalisasi data adalah proses transformasi nilai-nilai fitur ke dalam skala yang seragam, biasanya dalam rentang tertentu (0 hingga 1). Tujuannya adalah untuk menghindari dominasi fitur tertentu karena perbedaan skala, serta meningkatkan kinerja dan konvergensi algoritma pembelajaran mesin, seperti K-Means clustering. Normalisasi dihitung menggunakan persamaan (1) sebagai berikut:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Keterangan:
X' = Nilai ternormalisasi
X = Nilai asli data
X_{min}, X_{max}= Nilai minimum dan maximum data

2.3. Seleksi Fitur

Seleksi fitur merupakan langkah untuk memilih sejumlah fitur dari sekumpulan fitur awal atau fitur lengkap sehingga fitur yang dipilih memiliki dampak yang berarti terhadap ketepatan *clustering*. Dalam proses seleksi fitur, tujuan utamanya adalah untuk mengurangi jumlah fitur dan menghilangkan gangguan yang dianggap tidak relevan atau berlebihan, dengan harapan meningkatkan akurasi. Pengurangan dimensi pada fitur-fitur yang digunakan dalam analisis data, yang merupakan salah satu strategi dalam data mining, dapat dianggap sebagai konsep dari seleksi fitur (Syarif, 2023). Cara menerapkan seleksi fitur pada analisis data yaitu dengan menyiapkan data, melakukan *pre-processing* data, kemudian memilih metode seleksi fitur yang tepat.

2.4. *Chi*-Square

Metode seleksi fitur *Chi-Square* adalah teknik statistik yang digunakan dalam *machine learning* untuk menentukan relevansi fitur dengan menguji hubungan antara fitur-fitur dan variabel target, dengan melibatkan perhitungan statistik *Chi-Square* digunakan untuk mengevaluasi perbedaan antara frekuensi yang diamati dan yang diharapkan dari setiap kategori. Uji *Chi-Square* membantu dalam memilih fitur-fitur yang kuat terkait dengan variabel target, Semakin tinggi nilai chi-square pada suatu fitur maka semakin kuat juga hubungan antar fitur dengan variabel target tersebut (Ernayanti et al., 2023). Penggunaan Chi-Square sebelum proses klasterisasi dapat membantu menyaring fitur-fitur yang memiliki informasi kuat dan menghilangkan fitur yang mengandung noise atau redundan. Hal ini akan meningkatkan kualitas hasil klaster yang dihasilkan oleh K-Means, karena fitur-fitur yang digunakan telah melalui proses seleksi yang mempertimbangkan kekuatan hubungannya terhadap pola dalam data. Kaidah pengambilan keputusan untuk Chi-Square adalah tolak H_0 jika $X^2 \geq X^2_{a,(n-1)(m-1)}$ pada tingkat kepercayaan 95% dengan nilai alpha 0,05. Perhitungan untuk mencari nilai Chi-Square adalah seperti persamaan (1) dan (2) (Iswanto et al., 2021):

$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

$$E_{ij} = \frac{b_i K_j}{N} \quad (3)$$

Keterangan:
x^2 = Nilai *chi-square*
O_{ij} = Nilai observasi pada baris ke-i dan kolom ke-j
E_{ij}= Nilai expected pada baris ke-i dan kolom ke-j
b_i = Hasil penjumlahan baris ke-1
K_j = hasil penjumlahan kolom ke-j
N = Total observasi

2.5. Clustering

Clustering adalah proses mengelompokkan data atau objek yang terkait satu sama lain menjadi grup-grup yang disebut *cluster*. *Cluster* adalah grup data yang memiliki sama-sama kemiripannya. Klasterisasi dapat dilakukan menggunakan berbagai metode, di antaranya adalah metode *K-Means*.

Metode *K-Means* merupakan jenis *clustering* partisi yang membagi data ke dalam k wilayah terpisah. Metode ini menggunakan algoritma iteratif untuk mencari *centroid cluster* yang terbaik dan mengelompokkan data ke kluster yang sesuai (Syarif, 2023).

2.6. Algoritma K-Means

Pada analisis data, peneliti menerapkan teknik *clustering* menggunakan algoritma *K-Means* untuk melakukan analisis pada insiden kecelakaan lalu lintas di daerah Sidoarjo. Algoritma ini bertujuan untuk mencari *centroid-centroid* terbaik yang mengelompokkan kejadian kecelakaan lalu lintas ke dalam kelompok-kelompok yang sesuai. Algoritma ini berjalan melalui beberapa langkah, yaitu (Abdullah et al., 2022):

1. Menentukan jumlah kluster (k) yang diinginkan. Pilih k titik awal secara acak sebagai *centroid* untuk setiap *cluster*. *Centroid* awal ini dapat dipilih dari dataset atau secara acak dari ruang fitur.
2. Menggunakan rumus jarak Euclidean Distance untuk menghitung jarak pada setiap data. Perhitungan untuk mencari nilai jarak Euclidean Distance adalah seperti pada persamaan (3):

$$d(xi, \mu j) = \sqrt{\sum_{i=1}^{n} (xi - \mu j)^2} \quad (4)$$

   Keterangan:
   xi = Data kriteria
   μj = *Centroid* pada *cluster* ke-js
3. Mengelompokkan data berdasarkan jarak terdekat, lalu memperbarui *centroid* baru (rata-rata dari tiap data yang ada).
4. Langkah perhitungan pada langkah kedua dan ketiga dilakukan perulangan sampai tidak ada perubahan pada anggota *cluster*.

2.7. Koefisien Silhouette

Koefisien *silhouette* merupakan indikator evaluasi yang sering digunakan dalam analisis *cluster* untuk menentukan jumlah *cluster* optimal dalam data. Koefisien *silhouette* mendekati 1 menandakan bahwa data dalam *cluster* tersebut tergolong dalam *cluster* yang sesuai. Nilai negatif pada koefisien *silhouette* menandakan bahwa data ditempatkan secara tidak sesuai dalam klaster yang dimaksud. Kriteria subjektif untuk mengukur pengelompokan berdasarkan koefisien *silhouette* sesuai dengan Kauffman dan Roesseeuw dapat dilihat pada Tabel 1 (Izzaty et al., 2020):

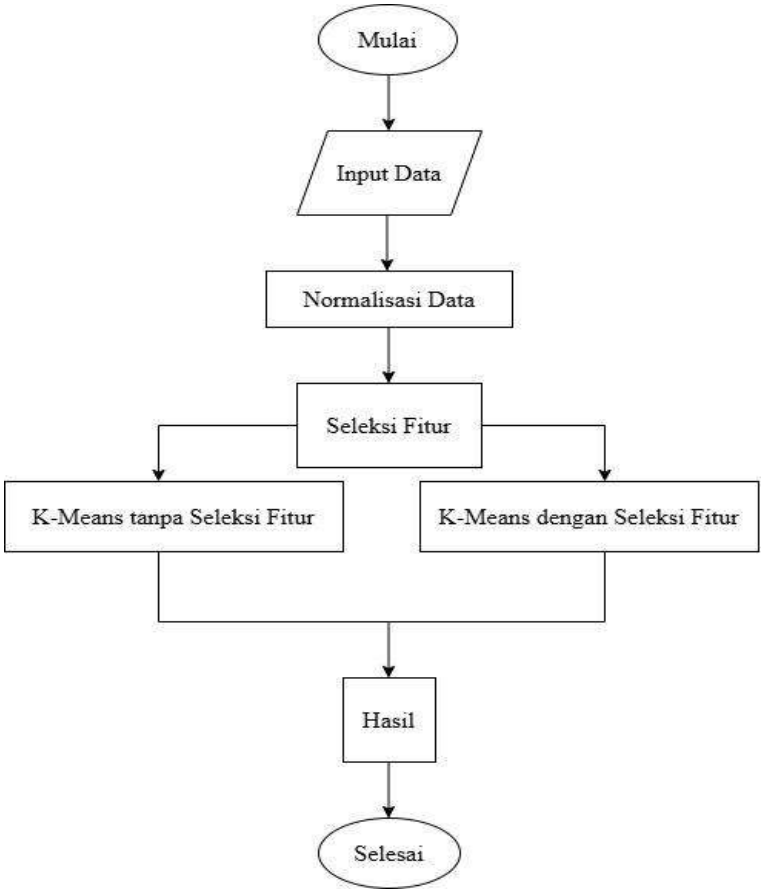
Tabel 1. Kriteria Pengukuran *Silhouette* Koefisien (Izzaty et al., 2020).

Nilai	Kriteria
0,71-1,00	Struktur Kuat
0,51-0,70	Struktur Baik
0,26-0,50	Struktur Lemah
≤ 0,25	Struktur Buruk

3. Metode Penelitian

Analisis cluster terkait kecelakaan lalu lintas ini masuk ke dalam kategori penelitian kuantitatif karena memanfaatkan perhitungan matematis dan menggunakan data berupa angka dalam prosesnya. Metode yang diterapkan dalam studi ini mencakup Chi-Square dan K-Means Clustering. Data yang dianalisis dalam cluster ini didapat dari sumber sekunder pada tahun 2024 melalui PT. Jasa Raharja (Persero) Kantor Cabang Kabupaten Sidoarjo, yang merupakan kantor pelayanan asuransi korban laka

lantas di wilayah tersebut. Penelitian ini mencakup dua belas variabel independen yang berkaitan dengan karakteristik kecelakaan lalu lintas. Pada Tabel 2 menyajikan variabel-variabel yang digunakan dalam penelitian, yang berasal dari data kecelakaan lalu lintas di wilayah Sidoarjo pada tahun 2023. Variabel-variabel ini mencakup informasi mengenai kondisi korban, waktu kejadian, jenis kendaraan, serta faktor-faktor lain yang dianggap relevan dalam membentuk pola kecelakaan (Titus & Jajuli, 2022).



Gambar 1. Alur penelitian

Tabel 2. Variabel Data Kecelakaan Tahun 2023

Variabel	Keterangan	Satuan
Pria	Korban kecelakaan berjenis kelamin pria	Jiwa
Wanita	Korban kecelakaan berjenis kelamin wanita	Jiwa
Remaja	Korban kecelakaan berusia remaja	Jiwa
Dewasa	Korban kecelakaan berusia dewasa	Jiwa
Waktu Terang	Banyaknya kecelakaan saat waktu terang	Kejadian
Waktu Gelap	Banyaknya kecelakaan saat waktu gelap	Kejadian
Panjang Jalan	Panjang jalan tiap kecamatan di Sidoarjo	Kilometer
Truk	Banyaknya kendaraan truk yang terlibat kecelakaan	Unit
Motor	Banyaknya kendaraan motor yang terlibat kecelakaan	Unit
Mobil	Banyaknya kendaraan mobil yang terlibat kecelakaan	Unit
Hari Kerja	Banyaknya kecelakaan yang terjadi saat hari kerja	Kejadian
Hari Libur	Banyaknya kecelakaan yang terjadi saat hari libur	Kejadian

Gambar 1 menunjukkan alur penelitian yang digunakan dalam studi ini. Penelitian dimulai dengan tahap input data kecelakaan lalu lintas dari 18 kecamatan di Sidoarjo, dilanjutkan dengan proses normalisasi data untuk memastikan keseragaman skala antar fitur yang dihitung menggunakan persamaan (1). Setelah itu, dilakukan seleksi fitur menggunakan metode Chi-Square untuk menentukan

variabel-variabel yang paling berpengaruh yang dihitung menggunakan persamaan (2) dan (3). Proses clustering dilakukan dalam dua skenario, yaitu dengan dan tanpa seleksi fitur, menggunakan algoritma K-Means yang dihitung menggunakan persamaan (4). Hasil dari kedua skenario tersebut kemudian dianalisis dan dibandingkan untuk memperoleh hasil clustering yang paling optimal.

4. Hasil dan Pembahasan

Pada Tabel 3 menunjukkan hasil normalisasi data karakteristik kecelakaan lalu lintas di 18 kecamatan di Sidoarjo. Nilai-nilai pada kedua tabel telah dinormalisasi ke dalam rentang 0 hingga 1 untuk memastikan bahwa seluruh fitur berada dalam skala yang sama dan tidak mendominasi proses analisis.

Tabel 3. Data Ternormalisasi

Pria	Wanita	Remaja	Dewasa	Waktu Terang	Waktu Gelap	Truk	Motor	Mobil	Hari Kerja	Hari Libur	Panjang Jalan
0.837	0.780	0.744	0.773	0.820	0.814	1.000	0.817	0.359	0.844	0.707	0.431
0.234	0.242	0.291	0.182	0.275	0.257	0.111	0.284	0.000	0.263	0.275	0.160
…	…	…	…	…	…	…	…	…	…	…	…
0.391	0.429	0.419	0.354	0.382	0.496	0.222	0.419	0.308	0.517	0.080	0.855

Keseluruhan data yang digunakan untuk *clustering* terdiri dari 18 kecamatan di wilayah Sidoarjo dan 12 variabel yang merupakan data kecelakaan lalu lintas. Sebelum dilakukan seleksi fitur, terdapat 12 variabel awal. Proses seleksi fitur dilakukan untuk memilih fitur terbaik dari variabel-variabel tersebut. Hasil uji coba menunjukkan pemilihan fitur dengan jumlah 3, 4, 5, dan 6 fitur. Dari hasil tersebut, nilai *Chi-Square* hitung ditampilkan pada Tabel 3. Sedangkan nilai untuk *Chi-Square* tabel pada taraf kepercayaan 95% (0,05) dengan nilai derajat kebebasan sebesar 187 menunjukkan angka 219,9, yang berarti bahwa tolak H_0. Dengan kata lain, variabel-variabel pada Tabel 3 berpengaruh signifikan.

Adapun fitur-fitur yang terpilih berdasarkan hasil seleksi *Chi-Square* yaitu jenis kelamin korban (Pria), usia (Remaja dan Dewasa), jenis kendaraan (Motor), waktu kejadian (Terang), dan hari kejadian (Hari Kerja) memiliki makna dalam menggambarkan karakteristik kecelakaan lalu lintas di Sidoarjo. Korban pria cenderung lebih dominan karena lebih aktif berkendara. Usia remaja dan dewasa merupakan kelompok paling aktif dalam mobilitas harian, baik untuk sekolah maupun bekerja, sehingga lebih rentan terhadap kecelakaan. Sepeda motor sebagai moda transportasi utama juga memiliki risiko tinggi karena minimnya perlindungan fisik. Kejadian pada waktu terang menunjukkan tingginya aktivitas lalu lintas di siang hari, sementara hari kerja mencerminkan intensitas mobilitas masyarakat yang lebih tinggi dibandingkan hari libur.

Tabel 4. Hasil Uji Coba Pemilihan Fitur

Jumlah Fitur Terpilih	Fitur yang Dipilih	Nilai *Chi-Square* Hitung
3	Pria, Motor, Hari Kerja	479,3; 597,7; 539,0
4	Pria, Motor, Waktu Terang, Hari Kerja	479,3; 597,7; 470,1; 539,0
5	Pria, Dewasa, Motor, Waktu Terang, Hari Kerja	479,3; 424,0; 597,7; 470,1; 539,0
6	Pria, Remaja, Dewasa, Motor, Waktu Terang, Hari Kerja	479,3; 322,9; 424,0; 597,7; 470,1; 539,0

Pada uji coba yang dilakukan, perbandingan antara K-Means tanpa seleksi fitur dan K-Means dengan seleksi fitur menunjukkan perbedaan performa yang cukup signifikan, khususnya dalam hal kualitas klaster yang terbentuk. *K-Means* tanpa seleksi fitur menggunakan keseluruhan 12 variabel awal secara langsung dalam proses klasterisasi. Hal ini berisiko menyebabkan adanya fitur yang kurang relevan ikut mempengaruhi pembentukan pusat klaster (centroid), sehingga hasil klaster menjadi kurang representatif. Hal ini tercermin dari rendahnya nilai *silhouette coefficient* yang dihasilkan, terutama ketika jumlah klaster lebih dari dua.

Sebaliknya, *K-Means* dengan seleksi fitur terlebih dahulu menggunakan metode *Chi-Square* untuk menyaring fitur-fitur yang paling relevan dan berpengaruh terhadap pola kecelakaan lalu lintas. Pemilihan fitur yang terbukti signifikan (seperti Pria, Motor, dan Hari Kerja untuk 3 fitur terbaik) menghasilkan klaster yang lebih padat dan terpisah dengan jelas, sebagaimana ditunjukkan oleh peningkatan nilai *silhouette coefficient* dari 0,57 menjadi 0,72 pada jumlah klaster 2. Tabel 4 menunjukkan nilai *silhouette* untuk *K-Means* tanpa seleksi fitur dan *K-Means* dengan seleksi fitur.

Tabel 5. Nilai Koefisien *Silhouette*

Cluster	*K-Means* tanpa *Chi-Square*	*K-Means* dengan *Chi-Square*			
		3 Fitur	4 Fitur	5 Fitur	6 Fitur
2	0,57	0,72	0,71	0,66	0,67
3	0,32	0,60	0,57	0,52	0,54
4	0,30	0,53	0,58	0,54	0,55

Selanjutnya, dilakukan pemetaan wilayah Sidoarjo dengan mewarnai kecamatan yang memiliki tingkat kecelakaan tinggi dengan warna merah dan tingkat kecelakaan rendah dengan warna hijau, yang ditunjukkan oleh Gambar 2. Berdasarkan Gambar 2, wilayah dengan tingkat kecelakaan tinggi ditunjukkan oleh *cluster* 0 yang terdiri dari 4 kecamatan. Sedangkan wilayah Sidoarjo yang termasuk ke dalam tingkat kecelakaan rendah ditunjukkan oleh *cluster* 1 yang terdiri dari 14 kecamatan. Persentase masing-masing *cluster* terbagi menjadi 22.22% untuk wilayah dengan tingkat kecelakaan tinggi, yaitu Kecamatan Balongbendo, Krian, Taman, dan Sidoarjo. Hal ini terlihat pada peta karena wilayah-wilayah tersebut berbatasan langsung dengan Kabupaten Gresik dan Kota Surabaya, serta berada di pusat Kota Sidoarjo. Kepadatan lalu lintas yang tinggi di wilayah tersebut meningkatkan peluang terjadinya kecelakaan, terutama karena jaringan jalan yang kompleks dan banyak dilalui oleh kendaraan dari berbagai kota. Sedangkan 77.78% untuk beberapa kecamatan lainnya di wilayah Sidoarjo seperti Kecamatan Buduran, Candi, Tulangan, Wonoayu, dan sebagainya dengan tingkat kecelakaan rendah.

Gambar 2. Peta Sidoarjo

5. Penutup

5.1. Kesimpulan

Proses seleksi fitur menghasilkan 3 variabel yang paling berpengaruh terhadap kecelakaan yaitu jenis kelamin (Pria), jenis kendaraan (Motor), serta Hari Kerja. Proses *clustering* menghasilkan dua *cluster* dengan tingkat kecelakaan tinggi dan rendah. Setelah diterapkan metode seleksi fitur, nilai validasi *clustering* meningkat dari 0,57 menjadi 0,72. *Cluster* dengan tingkat kecelakaan tinggi mencakup kecamatan Balongbendo, Taman, Sidoarjo, dan Krian. Sedangkan untuk wilayah dengan kecelakaan rendah terdiri dari kecamatan Waru, Tulangan, Tarik, dan seterusnya. Dengan demikian dapat disimpulkan bahwa metode seleksi fitur *Chi-Square* dapat diterapkan untuk meningkatkan nilai *silhouette* dari proses *clustering* menggunakan metode *K-Means*.

Referensi

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2022). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. *Quality and Quantity*, *56*(3), 1283–1291. https://doi.org/10.1007/S11135-021-01176-W/FIGURES/3

Amalia, M. F., & Arianto, D. B. (2024). Implementasi Algoritma K-Means Clustering Dalam Klasterisasi Kabupaten/Kota Provinsi Jawa Barat Berdasarkan Faktor Pemicu Stunting Pada Balita. *Jurnal Sistem Informasi Dan Sistem Komputer*, *9*(1), 36–46. https://doi.org/10.51717/SIMKOM.V9I1.356

Ernayanti, T., Rusgiyono, A., Rachman Hakim, A., Statistika, D., Sains dan Matematika, F., & Diponegoro, U. (2023). Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Pelanggan Tokopedia. *Jurnal Gaussian*, *11*(4), 562–571. https://doi.org/10.14710/J.GAUSS.11.4.562-571

Gurning, U. R., Octavia, S. F., Andriyani, D. R., Nurainun, N., & Permana, I. (2024). Prediksi Risiko Stunting pada Keluarga Menggunakan Naïve Bayes Classifier dan Chi-Square. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(1), 172–180. https://doi.org/10.57152/MALCOM.V4I1.1074

Hendrawan, S., Dewi, F. K. S., & Pranowo. (2023). Clustering Evaluasi Dosen Universitas Atma Jaya Yogyakarta Menggunakan Metode K-Means. *Jurnal Informatika Atma Jogja*, *4*(1), 1–8. https://doi.org/10.24002/JIAJ.V4I1.7436

Iswanto, A. P., Imron, N. A., & Priyanto, S. (2021). Analysis of Understanding and Violation of the Early Warning System (EWS) on Accident Rates at Level Crosses with the Chi-Square Method. *Jurnal Perkeretaapian Indonesia (Indonesian Railway Journal)*, *5*(1), 10–17. https://doi.org/10.37367/JPI.V5I1.133

Izzaty, U., Hg, I. R., & Devianto, D. (2020). Pengklasteran Kabupaten/Kota Di Provinsi Sumatera Barat Berdasarkan Indikator Kesejahteraan Masyarakat Dengan Validitas Koefisien Silhouette. *Jurnal Matematika UNAND*, *9*(2), 192–198. https://doi.org/10.25077/JMU.9.2.192-198.2020

J.W.Kaawoan, Y. (2023). Ganti Kerugian Oleh Pihak Yang Bertanggung Jawab Atas Terjadinya Kecelakaan Lalu Lintas. *Lex Privatum*, *11*(3). https://ejournal.unsrat.ac.id/v3/index.php/lexprivatum/article/view/47209

Kandida Br, A., #1, G., Silvi, M., #2, L., Muisa, E., & #3, Z. (2023). Reduksi Atribut Menggunakan Chi Square untuk Optimasi Kinerja Metode Decision Tree C4.5. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *9*(1), 44–49. https://doi.org/10.26418/JP.V9I1.56542

Luh, N., Dewi, P. P., Nyoman Purnama, I., & Utami, N. W. (2022). Penerapan Data Mining Untuk Clustering Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: STMIK Primakara). *Jurnal Ilmiah Teknologi Informasi Asia*, *16*(2), 105–112. https://doi.org/10.32815/JITIKA.V16I2.761

Priantama, Y., Azhima, T., & Siswa, Y. (2022). Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi Random Forest Classifier Dalam Prediksi Performa Akademik Mahasiswa. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(2), 251–260. https://doi.org/10.26798/JIKO.V6I2.651

Rahmansyah, A. K., Thufeil, A., Aziz, S., Novianto, N., & Rolliawati, D. (2023). Perbandingan Algoritma K-Means dan Fuzzy C-Means untuk Clustering Puskesmas Berdasarkan Gizi Balita di Surabaya. *Jurnal PROCESSOR*, *18*(1). https://doi.org/10.33998/PROCESSOR.2023.18.1.696

Septianingrum, F., Susilo, A., & Irawan, Y. (2021). *Metode Seleksi Fitur Untuk Klasifikasi Sentimen Menggunakan Algoritma Naive Bayes: Sebuah Literature Review*. https://doi.org/10.30865/mib.v5i3.2983

Sidiq, M., Kurniawan, D., Raharjo, S., & Nurharyanto, E. (2023). Penyelesaian Kecelakaan Lalu Lintas Yang Mengakibatkan Korban Meninggal Dunia Dengan Pendekatan Restorative Justice. *Kajian Hasil Penelitian Hukum*, *7*(1), 110–124. https://doi.org/10.37159/JMIH.V7I1.3031

Siregar, R. F., Paisah, N., & Patriotika, F. (2022). Analisis Kecelakaan Lalu Lintas (Black Site) Pada Ruas Jalan H.T. Rizal Nurdin Kota Padangsidimpuan. *STATIKA*, *5*(1), 14–30. https://jurnal.ugn.ac.id/index.php/statika/article/view/907

Syarif, Z. N. (2023). Penerapan Information Gain Dan Algoritma K-Means Untuk Klasterisasi Kedisiplinan Pegawai Menggunakan Rapidminer. *TeknoIS : Jurnal Ilmiah Teknologi Informasi Dan Sains*, *13*(1), 1–

12. https://doi.org/10.36350/JBS.V13I1.165

Titus, T. K., & Jajuli, M. (2022). Clustering Data Kecelakaan Lalu Lintas di Kecamatan Cileungsi Menggunakan Metode K-Means. *Generation Journal*, *6*(1), 1–12. https://doi.org/10.29407/GJ.V6I1.16103

Vernanda, A. A., Faisol, A., & Vendyansyah, N. (2021). Penerapan Metode K-Means Clustering Untuk Pemetaan Daerah Rawan Kecelakaan Lalu Lintas Di Kota Malang Berbasis Website. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *5*(2), 836–844. https://doi.org/10.36040/JATI.V5I2.3791

Yudhistira, A., & Andika, R. (2023). Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering. *Journal of Artificial Intelligence and Technology Information*, *1*(1), 20–28. https://doi.org/10.58602/JAITI.V1I1.22

Yusuf Haafidh Nur Siddiq, D. M. (2024). *Polda Jatim Catat 31.973 Kecelakaan Terjadi Sepanjang 2023, Tertinggi di Sidoarjo*. 4 Januari 2024. https://jatim.viva.co.id/kabar/9836-polda-jatim-catat-31973-kecelakaan-terjadi-sepanjang-2023-tertinggi-di-sidoarjo