Analisis Klasifikasi Kecelakaan Lalu Lintas di Kota Palembang Menggunakan Algoritma *Decision Tree*

Elsada Abelia
Fakultas Ilmu Komputer
Universitas Sriwijaya
Sumatera Selatan, Indonesia

Rossi Passarella Fakultas
Ilmu Komputer
Universitas Sriwijaya
Sumatera Selatan, Indonesia

***Abstract*—** Kecelakaan lalu lintas merupakan permasalahan serius dengan dampak sosial dan ekonomi yang signifikan. Penelitian ini bertujuan untuk melakukan klasifikasi jenis kecelakaan serta mengetahui pola-pola yang muncul berdasarkan karakteristik kecelakaan di Kota Palembang tahun 2021–2024 dengan beberapa pendekatan *machine learning*. Dataset berisi 2.637 baris data dengan fitur-fitur seperti tipe kecelakaan, kondisi cahaya, cuaca, fungsi jalan, kelas jalan, tipe jalan, bentuk geometri, dan kondisi permukaan jalan. Setelah tahap prapemrosesan dan seleksi fitur, dilakukan pemodelan menggunakan *Decision Tree*, *Random Forest*, *K-Nearest Neighbors (KNN)*, dan *XGBoost*. Hasil evaluasi menunjukkan bahwa model *Decision Tree* memiliki performa terbaik dengan akurasi 0,64. Analisis juga mengindikasikan bahwa kendaraan roda dua paling banyak terlibat dalam kecelakaan, terutama di jalan arteri saat cuaca cerah. Hasil penelitian ini diharapkan dapat menjadi bahan pertimbangan kebijakan peningkatan keselamatan lalu lintas di Palembang.

***Keywords*—** kecelakaan lalu lintas*, machine learning, decision tree,* Kota Palembang, klasifikasi

I. LATAR BELAKANG

Kecelakaan lalu lintas merupakan salah satu permasalahan utama dalam sistem transportasi yang berdampak luas terhadap aspek sosial dan ekonomi. Tingginya angka kecelakaan tidak hanya menyebabkan kerugian materi dan korban jiwa, tetapi juga menurunkan kualitas keselamatan dan kenyamanan dalam berlalu lintas. Di berbagai kota besar di Indonesia, termasuk Kota Palembang, fenomena ini menjadi perhatian serius karena melibatkan beragam faktor, seperti kepadatan lalu lintas, kondisi infrastruktur jalan, jenis kendaraan, dan perilaku pengendara. Dominasi kendaraan roda dua dalam populasi kendaraan menjadikan kelompok ini sebagai pihak yang paling rentan terhadap kecelakaan. Dalam upaya memahami pola-pola kecelakaan secara sistematis dan berbasis data, pendekatan *machine learning* (pembelajaran mesin) menjadi salah satu solusi yang dapat dimanfaatkan. Kecelakaan lalu lintas juga merupakan salah satu permasalahan kritis dalam sistem transportasi global yang terus meningkat, baik dari segi jumlah kejadian maupun dampaknya terhadap kehidupan masyarakat. Menurut laporan *World Health Organization (WHO)*, kecelakaan lalu lintas menyebabkan lebih dari 1,35 juta kematian setiap tahunnya, serta puluhan juta orang mengalami luka berat atau cacat permanen[1]. Di Indonesia, permasalahan ini semakin kompleks karena tingginya jumlah populasi kendaraan, terutama sepeda motor, yang secara signifikan berkontribusi terhadap proporsi kecelakaan fatal[2]. Kota-kota besar seperti Palembang menghadapi tantangan serupa, di mana pertumbuhan kendaraan yang tidak diimbangi dengan infrastruktur jalan yang memadai serta rendahnya disiplin lalu lintas menjadi penyebab utama tingginya tingkat kecelakaan.

Faktor penyebab kecelakaan bersifat multidimensional, mencakup karakteristik fisik jalan, kondisi cuaca, waktu kejadian, serta jenis kendaraan yang terlibat. Kompleksitas tersebut menyebabkan analisis konvensional berbasis statistik menjadi kurang efektif dalam mengungkap pola tersembunyi yang bersifat non-linear atau interaksi antar variabel. Oleh karena itu, pendekatan berbasis *machine learning* mulai banyak digunakan dalam analisis data kecelakaan lalu lintas karena kemampuannya untuk menangani volume data yang besar, mendeteksi hubungan yang kompleks, serta menghasilkan prediksi dan klasifikasi yang lebih akurat[3]. Dengan kemampuan tersebut, *machine learning* tidak hanya mampu memodelkan pola historis kecelakaan, tetapi juga mendukung pengambilan keputusan berbasis data (*data-driven decision making*) dalam konteks kebijakan keselamatan lalu lintas[4].

Beberapa studi terbaru menunjukkan bahwa penerapan algoritma seperti *decision tree* (pohon keputusan), *random forest* (hutan acak), dan *XGBoost* (penguatan gradien ekstrem) mampu meningkatkan akurasi klasifikasi tingkat keparahan kecelakaan hingga lebih dari 80%[5][6]. Misalnya, penelitian oleh Wu et al. (2023) memanfaatkan gabungan *deep learning* (pembelajaran mendalam) dan *random forest* untuk memprediksi tingkat cedera pengemudi dengan akurasi mencapai 92% berdasarkan data kecelakaan di Turki[7]. Selain itu, teknik *explainable AI (XAI)* seperti *SHAP* juga mulai diterapkan untuk meningkatkan transparansi model prediktif, memungkinkan peneliti dan pembuat kebijakan memahami kontribusi masing-masing fitur terhadap hasil klasifikasi[8]. Hal ini penting karena dalam konteks transportasi publik, interpretabilitas model sering kali sama pentingnya dengan akurasi. Lebih lanjut, perkembangan metode spasial-temporal seperti *Graph Neural Networks (GNN)* memperluas cakupan analisis kecelakaan dengan mempertimbangkan pengaruh spasial dan temporal antar ruas jalan.

Dalam studi yang dilakukan Nippani et al. (2023), model GNN berhasil mengidentifikasi risiko kecelakaan dengan nilai *area under curve (AUC)* sebesar 0,87, berdasarkan data jaringan jalan dan volume lalu lintas[9]. Selain itu, Shen dan Wei (2024) mengembangkan model STZITD-GNN yang mampu menangani ketidakpastian prediksi kecelakaan harian per segmen jalan, dengan peningkatan signifikan dalam akurasi dan keandalan model[10].

Berbagai inovasi ini menunjukkan bahwa pendekatan *machine learning* tidak hanya menawarkan keunggulan teknis dalam hal akurasi dan efisiensi, tetapi juga membuka peluang

baru dalam memahami dinamika kecelakaan secara menyeluruh.

Pemanfaatan model klasifikasi berbasis *machine learning* untuk menganalisis kecelakaan tidak hanya bertujuan memetakan tipe kejadian, tetapi juga mengidentifikasi pola dominan berdasarkan kondisi lingkungan, infrastruktur, dan waktu. Pengetahuan ini menjadi fondasi penting dalam mendukung intervensi kebijakan yang lebih tepat sasaran dan berbasis data.

Dalam penelitian ini, sejumlah algoritma klasifikasi seperti *decision tree*, *random forest*, *KNN*, dan *XGBoost* diuji untuk menganalisis kecelakaan lalu lintas di Kota Palembang. Berdasarkan hasil evaluasi, model *decision tree* menunjukkan performa terbaik dan menjadi fokus utama pembahasan lebih lanjut.

II. METODOLOGI PENELITIAN

2.1. Dataset dan Sumber Data

Penelitian ini menggunakan data sekunder kecelakaan lalu lintas dari Kepolisian Kota Palembang selama periode 2021 hingga 2024 dengan total 2.637 baris data dan 35 kolom. Data tersebut mencakup kolom seperti: No.Laka, Polres, Tanggal Kejadian, Waktu Kejadian, Tanggal Dilaporkan, Waktu Dilaporkan, Tingkat Kecelakaan, Jumlah Meninggal Dunia, Jumlah Luka Berat, Jumlah Luka Ringan, LAT, LONG, 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 Dilokasi, Kemiringan Jalan, Status Jalan, Penyebab, Perkiraan Nilai Rugi Material Kendaraan, Total Nilai Rugi Material non Kendaraan, Keterangan Kerugian.

Penelitian ini menambahkan kolom label baru yaitu ‘Jenis Kendaraan’ dan ‘Label Kecelakaan’, dimana jenis kendaraan berisi kendaraan yang mengalami kecelakaan. Jenis kendaraan diambil dari kolom keterangan kerugian karena kolom tersebut berisi informasi yang lain lain jadi dibuatlah kolom baru untuk label yaitu jenis kendaraan. Selanjutnya

kolom label kecelakaan berisi angka 1-14 dimana itu merupakan label untuk jenis kendaraan adapun penjelasan label 1-14 sebagai berikut: motor (1);motor dan motor (2);motor motor motor (3);motor dan mobil (4);motor dan pejalan kaki (5);motor dan sepeda (6);motor dan mobil truk (7);motor dan mobil tronton (8); kecelakaan yang melibatkan lebih dari 2 tetapi terlibat motor dalam kecelakaan tersebut (9);motor dan mobil trailer 10;motor dan mobil bus (11);motor dan mobil tangki (12);motor dan mobil box (13);motor dan mobil fuso (14);kecelakaan yang melibatkan lebih dari 2 kendaraan tetapi tidak ada motoryang terlibat (0);kecelakaan yg tidak melibatkan motor sama sekali (0). Pengelompokan ini bertujuan agar proses klasifikasi lebih fokus dan dapat menggambarkan pola kecelakaan dengan lebih jelas.

Penelitian ini tidak menggunakan metode otomatis seperti *f_classif* atau *selectKBest*, melainkan melakukan pemilihan fitur secara manual dan berbasis pemahaman domain. Pemilihan dilakukan dengan mempertimbangkan fitur-fitur yang memiliki hubungan logis dan potensial terhadap klasifikasi jenis kecelakaan, antara lain: Tipe Kecelakaan, Kondisi Cahaya, Cuaca, Fungsi Jalan, Kelas Jalan, Tipe Jalan, Bentuk Geometri, Kondisi Permukaan Jalan. Fitur-fitur tersebut dipilih karena dianggap representatif terhadap- karakteristik lingkungan dan kondisi saat kecelakaan terjadi.

Penelitian ini menggunakan pendekatan *supervised classification* (klasifikasi terawasi), dengan membandingkan beberapa algoritma *machine learning*, yaitu *decision tree*, *random forest*, *KNN*, dan *xgboost*[11]. Setiap model diuji untuk mengetahui performanya dalam mengklasifikasikan jenis kecelakaan berdasarkan fitur lingkungan dan infrastruktur jalan. Berdasarkan hasil evaluasi menggunakan metrik akurasi, *precision* (ketepatan), *recall* (daya tangkap), dan *f1-score* (rata-rata harmonik antara *precision* dan *recall)*, algoritma *decision tree* menunjukkan performa terbaik di antara model lainnya.

2.2. *DECISION TREE*

Decision tree adalah algoritma klasifikasi yang bekerja dengan membagi dataset ke dalam cabang-cabang berdasarkan nilai fitur, membentuk struktur menyerupai pohon. Setiap node internal mewakili suatu fitur, sedangkan setiap cabang menunjukkan kemungkinan nilai dari fitur tersebut, dan setiap daun merepresentasikan label output. Dalam konteks penelitian ini, *decision tree* digunakan untuk memetakan hubungan antara variabel-variabel seperti jenis jalan, kondisi permukaan, dan tipe kecelakaan dengan label kecelakaan yang telah dikategorikan sebelumnya[12]. Keunggulan *decision tree* terletak pada interpretabilitasnya yang tinggi, sehingga pola-pola penyebab kecelakaan dapat diidentifikasi secara jelas. Namun, kekurangan utama algoritma ini adalah kerentanannya terhadap *overfitting* (kelebihan pelatihan), terutama jika pohon tumbuh terlalu dalam tanpa pemangkasan.

Decision tree bekerja dengan membagi data ke dalam subset berdasarkan fitur tertentu melalui proses yang disebut pemilihan atribut optimal. Proses pemilihan ini menggunakan metrik seperti *entropy* (tingkat ketidakmurnian data) dan *information gain* (pengurangan ketidakpastian)[13]. *Entropy* digunakan untuk mengukur tingkat ketidakmurnian atau ketidakteraturan dalam data, dan dirumuskan sebagai:

$$Entropy(S) = \sum_{i=1}^{n} p_i \log_2(p_i) \quad (1)$$

di mana p_i merupakan probabilitas dari kelas ke-i. Kemudian, pemilihan fitur terbaik ditentukan menggunakan *information gain*, yaitu pengurangan *entropy* setelah pembagian data:

$$Informasi\ Gain(S,A) = Entry(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \quad (2)$$

Keterangan:

- A: Atribut fitur
- S_v: Subset data untuk nilai v

2.3. *RANDOM FOREST*

Random forest merupakan pengembangan dari *decision tree* yang menggabungkan banyak pohon keputusan *(ensemble)* untuk meningkatkan akurasi dan kestabilan model. Setiap pohon dalam *random forest* dilatih dengan subset acak dari data dan fitur, biasanya dengan pemungutan suara mayoritas (*majority voting*). Algoritma ini efektif untuk

dataset besar dengan banyak fitur kategorikal, seperti data kecelakaan lalu lintas[14]. Dalam penelitian ini, *random forest* membantu mengurangi *overfitting* dan meningkatkan generalisasi model saat memprediksi jenis kecelakaan berdasarkan kombinasi fitur lingkungan. Algoritma ini sangat efektif ketika digunakan pada dataset besar dengan banyak fitur kategorikal seperti yang digunakan dalam penelitian kecelakaan lalu lintas[15]. Hasil klasifikasi ditentukan berdasarkan mayoritas suara (*voting*) dari semua pohon:

$$y\hat{} = mode\{h_1(x), (h_2)(x), \dots, h_n(x)\} \quad (3)$$

Keterangan:

- $h_i(x)$: Output dari pohon ke-iii
- $y\hat{}$: Prediksi akhir
- n: Jumlah pohon

2.4. XGBOOST

XGBoost (extreme gradient boosting) adalah salah satu algoritma boosting yang bekerja dengan membangun model secara bertahap, di mana setiap model baru memperbaiki kesalahan dari model sebelumnya. *XGBoost* dikenal karena efisiensinya dalam pengolahan data dan kemampuannya menangani missing value serta data yang tidak seimbang[16]. Dalam penelitian ini, *XGBoost* digunakan untuk menangkap interaksi non-linear antar fitur seperti cuaca, kondisi cahaya, dan geometri jalan, yang mungkin memengaruhi probabilitas kecelakaan. Selain itu, kemampuannya melakukan regularisasi membantu mencegah *overfitting*, menjadikannya salah satu model unggulan dalam tugas klasifikasi. *XGBoost* merupakan algoritma *boosting* yang memperbaiki kesalahan prediksi dari model sebelumnya secara iteratif[17]. *XGBoost* mengoptimalkan fungsi objektif yang terdiri dari fungsi loss dan regularisasi:

a. Fungsi Objektif *(Objective Function)*:

$$L(\emptyset) = \sum_{i=1}^{n} l(y\, y\hat{}_{i,\ i}) + \sum_{k=1}^{K} \mathrm{fi}(f)_k \quad (4)$$

b. Regularisasi:

$$\mathrm{fi}(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{c} \omega_j^2 \quad (5)$$

Keterangan:

- l: Fungsi loss (misalnya *log loss* untuk klasifikasi)
- Ω: Fungsi regularisasi
- r: Jumlah daun
- ω_j: Skor daun ke-j
- γ,λ : Parameter regulasi

2.5. K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) adalah metode klasifikasi berbasis *instance-based learning* (pembelajaran berbasis contoh) yang mengklasifikasikan suatu titik data berdasarkan mayoritas label dari K tetangga terdekatnya dalam ruang fitur. KNN bekerja tanpa proses pelatihan eksplisit, tetapi menyimpan seluruh dataset sebagai basis referensi[18]. Dalam studi ini, KNN digunakan sebagai pembanding yang sederhana namun intuitif dalam mengelompokkan jenis kecelakaan berdasarkan kesamaan karakteristik, seperti lokasi, kondisi permukaan, dan cuaca. Walaupun mudah diterapkan, KNN cenderung kurang efisien dalam skala besar dan sangat dipengaruhi oleh distribusi data serta pemilihan nilai K. Proses pencarian tetangga menggunakan metrik jarak seperti *Euclidean distance* (jarak Euclidean):

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \quad (6)$$

KNN bersifat sederhana namun efektif, terutama ketika data memiliki distribusi yang baik, meskipun kinerjanya bisa menurun jika data tidak seimbang atau terlalu besar[19].

III. HASIL ANALISIS

3.1. Tahap Prapemrosesan

Tahapan prapemrosesan data dalam penelitian ini dilakukan secara menyeluruh untuk memastikan bahwa data yang digunakan dalam proses klasifikasi memiliki kualitas yang baik, bersih, dan siap untuk diolah oleh model *machine learning*[20]. Data yang digunakan merupakan data sekunder dari Kepolisian Kota Palembang, yang mencakup periode 2021 hingga 2024 dengan total sebanyak 2.637 entri kecelakaan. Langkah pertama dalam prapemrosesan adalah penanganan nilai kosong (*missing values*). Beberapa kolom ditemukan memiliki data yang hilang, khususnya pada kolom bertipe kategorikal seperti "Cuaca" dan "Fungsi Jalan". Untuk mengatasinya, dilakukan imputasi menggunakan modus (nilai yang paling sering muncul), yaitu nilai yang paling sering muncul pada kolom tersebut. Sedangkan untuk kolom numerik, seperti "Jumlah Korban", digunakan metode median untuk menghindari distorsi akibat *outlier*. Kolom-kolom yang memiliki terlalu banyak data kosong atau tidak memberikan kontribusi berarti terhadap proses klasifikasi, seperti nomor laporan dan keterangan bebas, dihapus dari dataset.

Langkah selanjutnya adalah penyamaan format data, khususnya pada kolom bertipe tanggal dan waktu, agar seragam dan dapat dikenali dalam analisis berbasis waktu. Kolom koordinat geografis seperti *latitude* dan *longitude* juga disesuaikan agar dapat digunakan untuk analisis spasial di tahap lanjutan. Setelah itu, dilakukan pembersihan data duplikat dan penyesuaian label kategori yang terlalu banyak (*high cardinality*) dengan cara menyatukan kategori-kategori minor ke dalam satu grup umum seperti "lainnya". Setelah data dinyatakan bersih dan konsisten, dilakukan proses pemilihan fitur (*feature selection*) secara manual berdasarkan eksplorasi awal dan pemahaman domain[21]. Metode otomatis seperti *f_classif* tidak digunakan pada tahap ini. Fitur yang dipilih untuk proses klasifikasi adalah fitur yang relevan secara kontekstual terhadap jenis kecelakaan yaitu "Tipe Kecelakaan", "Kondisi Cahaya", "Cuaca", "Fungsi Jalan", "Kelas Jalan", "Tipe Jalan", "Bentuk Geometri Jalan", dan "Kondisi Permukaan Jalan". Untuk memastikan pembaca memahami karakteristik dasar data, dilakukan analisis statistik deskriptif terhadap fitur-fitur utama. Tabel berikut menampilkan distribusi frekuensi kategori pada beberapa fitur penting yaitu fitur utama untuk memastikan pembaca memahami karakteristik dasar data yang dapat dilihat pada tabel 3.1.

Fitur	Kategori Dominan	Frekuensi	Persentase (%)

Tipe Kecelakaan	Kendaraan *Out of Control* keluar ke kiri jalan	468	17.75
Fungsi Jalan	Arteri	1,543	58.71
Cuaca	Cerah	2,613	99.35
Kondisi Cahaya	Terang / Jelas	2,603	98.75
Tipe Jalan	2/2 TB (2 Lajur / 2 Arah Tanpa Median)	1,100	41.76
Bentuk Geometri Jalan	Lurus	2,381	90.50
Kondisi Permukaan Jalan	Baik	2,570	97.90

Tabel 3.1. Statistik Deskriptif Fitur Utama

Dari tabel di atas dapat dilihat bahwa sebagian besar kecelakaan terjadi saat cuaca cerah, pada jalan arteri dengan permukaan jalan baik dan kondisi cahaya terang. Kondisi ini menunjukkan bahwa faktor perilaku pengendara memiliki peran penting dalam terjadinya kecelakaan, bukan semata- mata kondisi lingkungan yang buruk.

Agar fitur-fitur tersebut dapat diproses oleh algoritma klasifikasi, dilakukan encoding data kategorikal menggunakan teknik *label encoding* (pengkodean label), yaitu mengubah nilai kategorikal menjadi representasi numerik. Label target yang digunakan adalah "Label Kecelakaan", yaitu klasifikasi tipe kecelakaan berdasarkan pola kejadian, seperti kendaraan keluar jalur, tabrakan searah, atau kecelakaan di persimpangan. Seluruh data yang telah diproses kemudian dibagi menjadi dua bagian: data latih (*training data*) dan data uji (*testing data*) dengan perbandingan umum 80:20 dimana jumlah data latih adalah 2109 sedangkan jumlah data uji adalah 528. Pembagian ini bertujuan agar model dapat dilatih dengan cukup data untuk mengenali pola-pola kecelakaan, dan kemudian dievaluasi performanya secara objektif pada data uji yang tidak dikenalnya. Dengan tahapan prapemrosesan ini, dataset menjadi lebih siap untuk menjalani proses pelatihan model machine learning secara efektif.

3.2. Pelatihan dan Evaluasi Model

Setelah data dibersihkan, dipilih fiturnya, serta dibagi ke dalam data latih dan data uji, tahap selanjutnya adalah membangun dan melatih model klasifikasi. Dalam penelitian ini digunakan empat algoritma klasifikasi yang berbeda, yaitu *decision tree, random forest, KNN,* dan *XGBoost*. Pemilihan algoritma ini didasarkan pada popularitas, kekuatan masing- masing dalam menangani data kategorikal, serta keragaman pendekatan yang digunakan untuk membuat prediksi. Algoritma *decision tree* dikenal baik dalam mengekstraksi aturan berbasis pohon keputusan yang mudah diinterpretasikan. *Random forest* merupakan pengembangan dari *decision tree* dengan membentuk banyak pohon keputusan (*ensemble)* dan menggabungkan hasilnya, sehingga lebih tahan terhadap *overfitting*. Sementara itu, KNN bekerja dengan melihat kedekatan data baru terhadap data latih dalam ruang fitur dan memberikan label berdasarkan mayoritas tetangga terdekat. *XGBoost* adalah algoritma berbasis *boosting* yang kuat dalam mendeteksi pola *non-linear* dan menangani ketidakseimbangan data dengan baik.

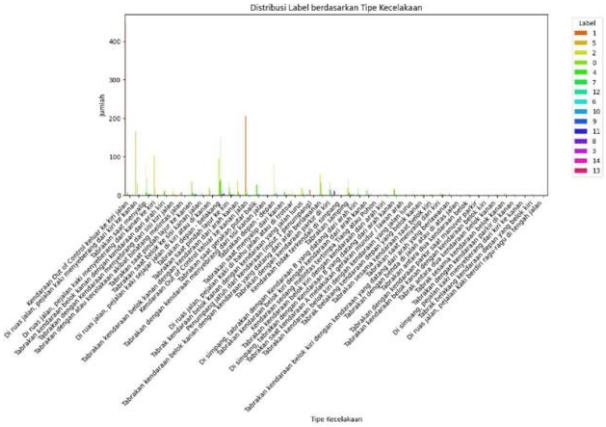
Proses pelatihan dilakukan menggunakan data latih yang telah disiapkan sebelumnya. Model kemudian dievaluasi menggunakan data uji untuk menilai performanya. Evaluasi dilakukan dengan menggunakan empat metrik utama: akurasi, *precision, recall,* dan *f1-score*[22]. Akurasi digunakan untuk mengukur proporsi prediksi yang benar terhadap keseluruhan data uji. *Precision* mengukur ketepatan model dalam memprediksi suatu kelas tertentu, *recall* mengukur kemampuan model dalam menemukan seluruh instance dari suatu kelas, sedangkan *f1-score* merupakan harmonisasi antara *precision* dan *recall*, yang berguna saat data memiliki distribusi label yang tidak seimbang.

3.3. Visualisasi Distribusi Fitur

Visualisasi distribusi fitur dilakukan untuk mengevaluasi hubungan antara fitur-fitur kategorikal dengan label target, yakni "Jenis Kendaraan" dan "Label Kecelakaan". Tujuan dari visualisasi ini adalah untuk mengetahui pola distribusi kecelakaan lalu lintas berdasarkan faktor-faktor lingkungan, infrastruktur jalan, dan kondisi cuaca, yang diharapkan dapat memberikan gambaran lebih jelas mengenai karakteristik kecelakaan di Kota Palembang. Setiap fitur divisualisasikan menggunakan diagram batang *(countplot)* dengan pewarnaan berdasarkan kategori label target.

3.3.1 Distribusi Berdasarkan Tipe Kecelakaan

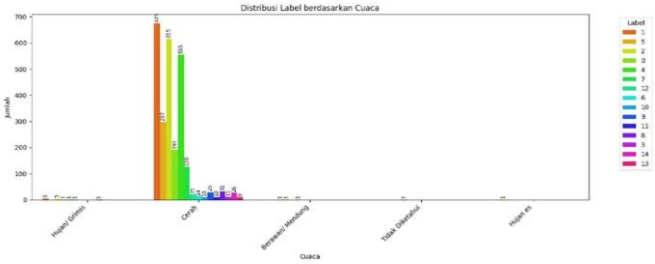
Gambar 3.3.1 menunjukkan bahwa kecelakaan paling banyak terjadi pada tipe kecelakaan Di ruas jalan, kendaraan *out of control* keluar ke kiri jalan, dengan label 1 sebagai kontributor utama sebanyak 448 kecelakaan. Hal ini mengindikasikan bahwa kehilangan kendali kendaraan merupakan penyebab dominan dalam dataset. Dalam tipe kecelakaan tersebut, kendaraan roda dua (label 1) mendominasi secara signifikan. Hal ini mencerminkan tingginya populasi sepeda motor di Kota Palembang serta kerentanannya terhadap insiden, khususnya dalam situasi kehilangan kendali atau interaksi langsung dengan kendaraan lain yang bergerak searah. Fenomena ini juga dapat dihubungkan dengan perilaku pengendara motor yang cenderung lebih lincah namun berisiko tinggi dalam manuver jalanan.



Gambar 3.3.1. Distribusi jenis kendaraan berdasarkan tipe kecelakaan.

3.3.2 Distribusi Berdasarkan Kondisi Cuaca

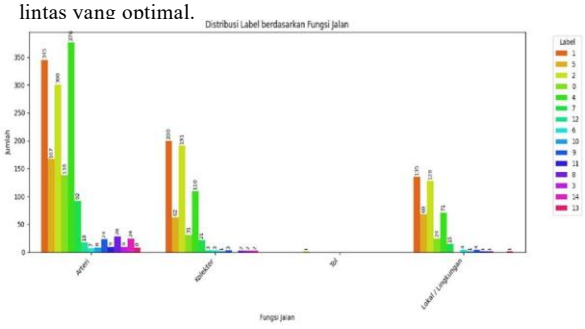
Gambar 3.3.2 memperlihatkan bahwa sebagian besar kecelakaan terjadi saat kondisi cuaca cerah. Data ini cukup menarik karena bertentangan dengan anggapan umum bahwa cuaca buruk seperti hujan atau kabut lebih berisiko. Tingginya kejadian kecelakaan saat cuaca cerah mengindikasikan bahwa faktor kecepatan berkendara dan kelalaian manusia menjadi penyebab utama. Pengendara mungkin merasa lebih aman dan cenderung meningkatkan kecepatan atau mengurangi kewaspadaan ketika cuaca baik. Sepeda motor tunggal kembali mendominasi jumlah kecelakaan dalam kondisi cuaca apapun, namun terlihat paling banyak pada kondisi cerah sebanyak 675.



Gambar 3.3.2. Distribusi kecelakaan berdasarkan kondisi cuaca.

3.3.3 Distribusi Berdasarkan Fungsi Jalan

Pada Gambar 3.3.3 ditunjukkan distribusi kecelakaan terhadap fungsi jalan, seperti jalan lokal, kolektor, tol, dan arteri. Jalan kolektor dan lokal memiliki jumlah kecelakaan yang cukup tinggi. Hal ini dapat dikaitkan dengan volume lalu lintas yang padat namun tidak diimbangi dengan pengawasan dan pengaturan lalu lintas yang optimal.

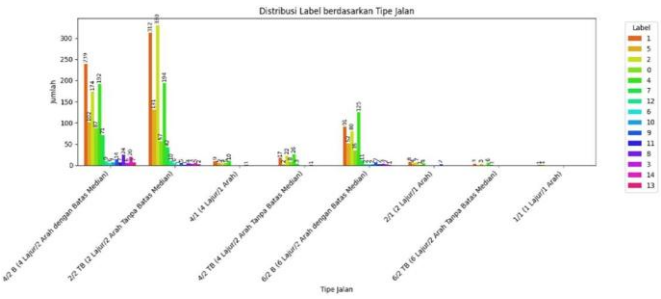


Gambar 3.3.3. Distribusi kecelakaan berdasarkan fungsi jalan.

Kendaraan roda dua menjadi jenis kendaraan yang paling sering terlibat dalam kecelakaan pada kedua jenis jalan tersebut, yang menunjukkan bahwa pengendara motor sering menggunakan jalur ini untuk aktivitas harian. Tetapi jumlah kecelakaan paling tinggi terjadi pada fungsi jalan arteri, jalan arteri merupakan jalur utama dengan volume lalu lintas tinggi dan kecepatan rata-rata yang lebih besar, sehingga memiliki risiko kecelakaan yang lebih besar dibandingkan dengan jalan kolektor atau lokal/lingkungan dimana posisi pertama kontributor adalah label 4 sebanyak 376 kecelakaan.

3.3.4 Distribusi Berdasarkan Tipe Jalan

Gambar 3.3.4 menunjukkan distribusi kecelakaan berdasarkan tipe jalan. Dari visualisasi terlihat bahwa tipe jalan 2/2 TB(dua lajur/dua arah tanpa batas median). Tipe jalan seperti ini memiliki potensi tabrakan lebih tinggi karena interaksi langsung antar kendaraan dari dua arah yang berlawanan, serta kurangnya perlindungan jika terjadi pergeseran jalur, label 2 menjadi yang paling banyak mendominasi kecelakaan dengan nilai 330 kecelakaan. Tidak adanya pembatas jalur fisik pada tipe jalan ini dapat meningkatkan risiko tabrakan frontal dan kecelakaan akibat pergerakan bebas antar lajur. Fenomena ini menegaskan pentingnya desain jalan yang aman dan pembatasan kecepatan pada jalan dua arah untuk mengurangi risiko kecelakaan.



Gambar 3.3.4. Distribusi kecelakaan berdasarkan tipe jalan.

Dari keempat visualisasi utama tersebut, dapat disimpulkan bahwa kecelakaan lalu lintas di Kota Palembang didominasi oleh kendaraan roda dua dan paling sering terjadi pada kondisi lingkungan yang dianggap ideal seperti cuaca cerah, jalan dua arah tanpa median, dan tipe kecelakaan non- kompleks seperti tabrakan searah. Hal ini menegaskan bahwa faktor perilaku pengendara menjadi penyebab utama kecelakaan dibandingkan kondisi jalan atau cuaca. Selain itu, distribusi berdasarkan tipe kecelakaan juga memperlihatkan kecenderungan umum bahwa insiden yang melibatkan kehilangan kendali dan tabrakan searah merupakan pola yang paling umum ditemukan dalam data. Fitur-fitur lain seperti kelas jalan, bentuk geometri, dan kondisi cahaya yang juga dianalisis memiliki pola distribusi serupa dan mendukung temuan utama, namun tidak ditampilkan secara visual dalam paper ini untuk menjaga fokus pembahasan. Analisis naratif terhadap fitur-fitur tersebut tetap dimasukkan dalam bagian interpretasi dan diskusi pola kecelakaan.

3.4. Evaluasi Model Klasifikasi

Evaluasi model klasifikasi dilakukan untuk mengetahui seberapa baik algoritma machine learning dalam mengelompokkan data kecelakaan ke dalam kategori yang sesuai berdasarkan label yang telah ditentukan. Dalam penelitian ini, digunakan empat algoritma klasifikasi populer, yaitu *decision tree, random forest, KNN,* dan *XGBoost*. Keempat model ini dipilih karena masing-masing memiliki karakteristik berbeda dalam menangani data kategorikal, linearitas, dan kompleksitas relasi antar fitur. Evaluasi dilakukan dengan membagi dataset menjadi data latih dan data uji menggunakan rasio 80:20, di mana 80% data digunakan untuk melatih model dan 20% sisanya digunakan

untuk menguji performa model. Untuk mengukur performa model, digunakan empat metrik evaluasi umum dalam *machine learning*, yaitu:

- Akurasi: proporsi prediksi yang benar terhadap seluruh jumlah prediksi.
- *Precision*: tingkat ketepatan prediksi positif yang benar dibandingkan dengan total prediksi positif.
- *Recall*: kemampuan model untuk menemukan semua instance positif yang sebenarnya.
- *F1-score*: rata-rata harmonik dari precision dan recall yang memberikan keseimbangan antara keduanya.

Berikut adalah hasil evaluasi singkat keempat model berdasarkan metrik utama:

Model	Akurasi	*Macro F1-Score*	*Weight F1-Score*
Decision Tree	0.64	0.20	0.59
Random Forest	0.61	0.18	0.56
XGBoost	0.61	0.17	0.55
K-Nearest Neighbors	0.59	0.15	0.51

Hasil evaluasi di atas menunjukkan bahwa model *decision tree* memiliki performa paling tinggi dibandingkan model lainnya, dengan akurasi 64% dan *weighted f1-score* sebesar 0.59. *Decision tree* bekerja dengan membangun struktur pohon dari data pelatihan, di mana setiap cabang merepresentasikan keputusan atas fitur tertentu, dan daun pohon mewakili label kelas akhir. Model ini efektif dalam mengidentifikasi pola *non-linear* dan menangani fitur kategorikal, sehingga cocok digunakan dalam studi ini.

Model *random forest*, yang merupakan kumpulan dari beberapa pohon keputusan, memiliki performa yang cukup baik dengan akurasi 61% dan *weighted f1-score* sebesar 0.56. *Random forest* memiliki keunggulan dalam mengurangi risiko *overfitting* yang umum terjadi pada *decision tree* tunggal karena menggabungkan prediksi dari banyak pohon (*ensemble*). Namun, model ini juga lebih kompleks dan memerlukan sumber daya komputasi yang lebih besar.

Model XGBoost, sebagai salah satu algoritma boosting yang kuat, juga menghasilkan akurasi sebesar 61%. Meski memiliki pendekatan iteratif yang mampu meningkatkan performa secara bertahap, *nilai macro* dan *weighted f1-score* dari XGBoost masih sedikit di bawah *random forest*. Hal ini kemungkinan dipengaruhi oleh distribusi data yang tidak seimbang antar kelas dan ukuran dataset yang relatif kecil.

Sementara itu, model KNN menunjukkan performa terendah di antara keempat model yang diuji, dengan akurasi hanya 59% dan *f1-score* yang juga paling rendah. KNN sangat bergantung pada distribusi data dan pemilihan parameter k yang tepat. Ketidakseimbangan data pada beberapa label menyebabkan model ini kesulitan dalam memberikan klasifikasi yang akurat, terutama untuk label yang jarang muncul (minoritas)[23].

Nilai *macro f1-score* yang rendah (0.20) menunjukkan adanya ketidakseimbangan data antar kelas. Beberapa label jenis kendaraan, seperti “motor dan motor” atau “motor dan mobil”, mendominasi dataset, sementara kategori lain seperti “motor dan pejalan kaki” memiliki jumlah sangat sedikit.

Ketidakseimbangan ini menyebabkan model cenderung lebih akurat pada kelas mayoritas, sementara kelas minor sulit dikenali dengan baik.

Untuk mengatasi kondisi tersebut, metode berikut dapat dipertimbangkan pada penelitian lanjutan:

1. SMOTE (Synthetic Minority Oversampling Technique) – menambah data sintetis pada kelas minor agar distribusi data lebih seimbang.
2. Class Weighting – memberi bobot lebih besar pada kelas minor agar model tidak bias terhadap kelas mayoritas.
3. Stratified Sampling – memastikan setiap kelas terwakili secara proporsional pada data latih dan data uji.

Pendekatan-pendekatan ini dapat meningkatkan nilai *macro f1-score* dan keadilan performa model antar kelas tanpa mengorbankan akurasi keseluruhan.

Nilai macro *f1-score* menunjukkan rata-rata performa model untuk semua kelas tanpa memperhatikan jumlah data per kelas, sedangkan *weighted f1-score* memberikan gambaran yang lebih adil karena memperhitungkan jumlah data pada setiap kelas. Dalam konteks data yang tidak seimbang seperti pada penelitian ini, *weighted f1-score* memberikan informasi yang lebih relevan mengenai performa keseluruhan model.

Dengan mempertimbangkan semua aspek tersebut, maka model *decision tree* dipilih sebagai model utama untuk dianalisis lebih lanjut. Keunggulannya dalam interpretasi yang jelas, efisiensi dalam pelatihan, serta kemampuannya dalam menangani variabel kategorikal membuatnya menjadi pilihan yang tepat untuk mendukung tujuan penelitian ini dalam mengidentifikasi dan memahami pola kecelakaan lalu lintas di Kota Palembang[24].

3.5. Interpretasi Pola Kecelakaan

Interpretasi pola kecelakaan dalam penelitian ini bertujuan untuk memahami hubungan antara karakteristik lingkungan dan kondisi jalan dengan kecenderungan jenis kecelakaan lalu lintas. Analisis dilakukan berdasarkan visualisasi distribusi fitur-fitur penting terhadap label kecelakaan yang telah dikategorikan. Fitur-fitur tersebut meliputi tipe kecelakaan, kondisi cahaya, cuaca, fungsi jalan, tipe jalan, bentuk geometri jalan, dan kondisi permukaan jalan. Hasil visualisasi memberikan gambaran bahwa terdapat pola-pola yang konsisten dalam keterkaitan antara kondisi lingkungan dan jenis kendaraan yang terlibat dalam kecelakaan.

Salah satu pola paling menonjol ditemukan pada fitur tipe kecelakaan, di mana tipe "Di ruas jalan, kendaraan out of control keluar ke kiri jalan" mencatat jumlah kecelakaan tertinggi. Label yang paling sering muncul dalam tipe kecelakaan ini adalah label 1, yang menunjukkan bahwa kendaraan roda dua sangat rentan terhadap kecelakaan jenis ini. Hal ini menunjukkan bahwa kehilangan kendali saat berkendara merupakan penyebab dominan dalam banyak insiden, khususnya yang melibatkan pengendara sepeda motor.

Selain itu, dari sisi fungsi jalan, kecelakaan paling banyak terjadi pada jalan arteri, yang merupakan jalur utama dengan volume lalu lintas tinggi dan kecepatan kendaraan yang

relatif cepat. Menariknya, jenis kendaraan yang paling banyak terlibat dalam kecelakaan pada jalan arteri adalah kendaraan dengan label 4, yang mengindikasikan kecelakaan antara sepeda motor dan mobil pribadi. Ini menunjukkan bahwa kendaraan roda dua tetap rentan meskipun berinteraksi dengan kendaraan pribadi di jalan utama.

Pada fitur tipe jalan, data menunjukkan bahwa jenis jalan 2/2 TB (dua lajur dua arah tanpa batas median) merupakan tipe jalan dengan frekuensi kecelakaan tertinggi. Pada jenis jalan ini, label 2 menjadi label kendaraan yang paling dominan terlibat dalam kecelakaan. Label ini menggambarkan kecelakaan antara motor dan motor, yang mengindikasikan bahwa kondisi jalan tanpa median berpotensi meningkatkan risiko tabrakan antar kendaraan roda dua, terutama ketika ada manuver yang salah arah atau pengendara berpindah jalur secara tiba-tiba.

Dari sisi kondisi cahaya dan cuaca, diketahui bahwa sebagian besar kecelakaan justru terjadi saat pencahayaan terang dan cuaca cerah. Ini menunjukkan bahwa kecelakaan tidak hanya dipicu oleh faktor eksternal yang sulit dikendalikan, seperti hujan atau kabut, melainkan sangat berkaitan dengan perilaku pengguna jalan saat kondisi terlihat aman. Cuaca cerah bisa membuat pengendara merasa lebih percaya diri untuk melaju cepat, yang secara tidak langsung meningkatkan risiko kecelakaan.

Pada fitur fungsi jalan lainnya seperti jalan kolektor dan lokal, kendaraan roda dua tetap menjadi jenis kendaraan yang paling banyak beroperasi di segmen ini, dengan risiko kecelakaan tinggi. Hal serupa juga terlihat pada fitur bentuk geometri jalan, di mana kecelakaan paling banyak ditemukan pada ruas jalan yang lurus. Meskipun secara teknis jalan lurus memberikan visibilitas lebih baik, kondisi ini sering disalahartikan oleh pengendara sebagai kesempatan untuk meningkatkan kecepatan.

Dari segi kondisi permukaan jalan, ditemukan bahwa sebagian besar kecelakaan terjadi pada permukaan jalan yang baik dan kering. Temuan ini mengonfirmasi bahwa penyebab utama kecelakaan bukanlah kondisi fisik jalan, melainkan perilaku pengguna jalan.

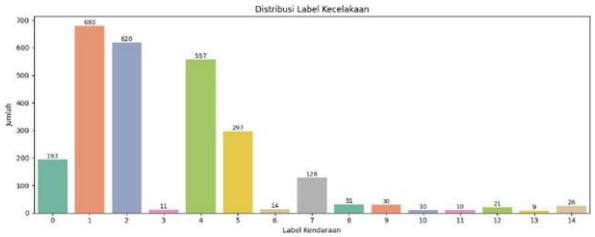
Secara keseluruhan, hasil interpretasi distribusi fitur menunjukkan bahwa kecelakaan lalu lintas di Kota Palembang didominasi oleh kendaraan roda dua dan sering terjadi pada kondisi lingkungan yang secara visual dianggap ideal, seperti cuaca cerah, permukaan jalan baik, dan geometri lurus. Pola ini memperkuat kesimpulan bahwa intervensi kebijakan keselamatan lalu lintas harus difokuskan pada edukasi dan penegakan hukum terhadap perilaku berkendara yang tidak aman, khususnya di wilayah dengan kepadatan lalu lintas tinggi dan pada waktu-waktu dengan lalu lintas padat. Visualisasi ini juga menjadi dasar penting dalam memprioritaskan lokasi rawan kecelakaan serta mengembangkan sistem peringatan dini berbasis data.

Dengan mengetahui distribusi dan pola dominan berdasarkan fitur-fitur yang dianalisis, penelitian ini memberikan arah yang jelas terhadap identifikasi titik-titik kritis dan kondisi yang membutuhkan perhatian lebih dalam upaya peningkatan keselamatan jalan.

3.6. Analisis Hasil Klasifikasi

Analisis hasil klasifikasi dilakukan untuk menilai sejauh mana model yang dibangun mampu mengidentifikasi dan mengelompokkan pola kecelakaan secara efektif. Dalam penelitian ini, model klasifikasi digunakan untuk menentukan jenis kecelakaan berdasarkan fitur-fitur yang telah dipilih. Berdasarkan hasil evaluasi yang telah dipaparkan sebelumnya, model *decision tree* dipilih sebagai model terbaik karena menunjukkan akurasi tertinggi dan *f1-score* yang relatif stabil dibandingkan model lainnya.

Hasil klasifikasi memperlihatkan bahwa model mampu mengenali pola kecelakaan yang melibatkan kendaraan roda dua secara signifikan, khususnya dalam label-label dominan seperti label 1 (kecelakaan tunggal sepeda motor) dan label 2 (motor vs motor). Hal ini sesuai dengan hasil interpretasi visualisasi yang menunjukkan bahwa kendaraan roda dua merupakan pihak yang paling banyak terlibat dalam kecelakaan lalu lintas. Model juga cukup baik dalam mengklasifikasikan kecelakaan berdasarkan tipe jalan dan kondisi lingkungan.



Gambar 3.6. Distribusi label kecelakaan.

Grafik "Distribusi Label Kecelakaan" menunjukkan jumlah kejadian kecelakaan berdasarkan label kendaraan yang mewakili kombinasi jenis kendaraan bermotor. Terlihat bahwa distribusi data tidak merata, di mana label 1 (680 kasus), label 2 (620 kasus), dan label 4 (557 kasus) mendominasi jumlah kecelakaan, sementara label lain seperti 3, 10, 11, dan 13 memiliki jumlah kasus yang sangat rendah. Namun demikian, model masih mengalami kesulitan dalam mengklasifikasikan beberapa label minor yang jumlah datanya sangat sedikit. Hal ini terlihat dari rendahnya nilai *recall* dan *f1-score* pada label-label yang tidak dominan. Ketidakseimbangan distribusi data antar label menyebabkan model lebih cenderung mempelajari pola dari kelas mayoritas. Meskipun demikian, secara keseluruhan, model telah memberikan hasil yang cukup baik untuk mengelompokkan data kecelakaan dalam kategori yang sesuai.

Analisis ini menunjukkan bahwa dengan pemilihan fitur yang tepat dan pendekatan klasifikasi yang sesuai, data kecelakaan lalu lintas dapat dikelompokkan secara lebih terstruktur. Hasil klasifikasi ini dapat digunakan untuk mendukung pengambilan keputusan dalam upaya peningkatan keselamatan jalan, dengan fokus pada jenis kecelakaan dan kondisi lingkungan yang memiliki risiko tinggi.

IV. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Penelitian ini menunjukkan bahwa pendekatan klasifikasi menggunakan algoritma *machine learning* dapat dimanfaatkan untuk menganalisis pola kecelakaan lalu lintas di Kota Palembang berdasarkan data kecelakaan historis tahun 2021 hingga 2024. Dengan memanfaatkan fitur-fitur relevan seperti tipe kecelakaan, fungsi jalan, kondisi cahaya,

cuaca, dan tipe jalan, model klasifikasi mampu mengelompokkan data ke dalam kategori jenis kecelakaan yang telah dilabelkan. Model *decision tree* terbukti menjadi model dengan performa terbaik di antara algoritma lain seperti *random forest*, *KNN*, dan *XGBoost*. Akurasi tertinggi dan interpretasi hasil yang mudah menjadi keunggulan utama dari *decision tree* dalam konteks data ini. Selain itu, visualisasi distribusi fitur memperlihatkan bahwa sebagian besar kecelakaan melibatkan kendaraan roda dua, terjadi di jalan lurus, permukaan baik, dan kondisi lingkungan cerah, serta didominasi oleh ruas jalan kelas II dan tipe jalan 2/2 TB. Temuan ini menegaskan bahwa faktor perilaku dan interaksi antar kendaraan masih menjadi penyebab utama kecelakaan, bukan semata-mata kondisi infrastruktur. Oleh karena itu, pendekatan analitik seperti ini dapat mendukung pengambilan keputusan yang lebih efektif bagi pemangku kebijakan dalam meningkatkan keselamatan lalu lintas.

4.2. Saran

Sebagai tindak lanjut, disarankan agar penelitian serupa di masa depan mengembangkan model klasifikasi dengan menyeimbangkan distribusi label untuk meningkatkan akurasi klasifikasi, terutama pada kategori yang jarang muncul. Integrasi data spasial seperti koordinat lokasi kecelakaan juga dapat memperkuat analisis, khususnya dalam memetakan titik rawan kecelakaan. Hasil dari penelitian ini sebaiknya dimanfaatkan oleh instansi terkait sebagai dasar perencanaan intervensi keselamatan lalu lintas yang lebih tepat sasaran. Selain itu, edukasi dan penegakan hukum terhadap pengendara roda dua sebagai kelompok paling rentan harus menjadi prioritas utama. Penambahan fitur perilaku pengendara serta waktu kejadian dalam dataset juga dapat menjadi langkah strategis untuk meningkatkan kualitas model di masa mendatang.

REFERENCES

[1] N. Casado-Sanz, B. Guirao, and M. Attard, "Analysis of the risk factors affecting the severity of traffic accidents on spanish crosstown roads: The driver's perspective," *Sustain.*, vol. 12, no. 6, 2020, doi: 10.3390/su12062237.

[2] S. Djalante, "Traffic Accident Characteristic Assessment to Enhance Sustainability in Road and Transportation Infrastructures in Indonesia," *OALib*, vol. 07, no. 10, pp. 1–12, 2020, doi: 10.4236/oalib.1106796.

[3] W. Zhou, "Traffic Accident Severity Prediction Based on Data Cleaning and Machine Learning (Random Forest / Xgboost)," *Highlights Sci. Eng. Technol.*, vol. 85, pp. 376–388, 2024, doi: 10.54097/h8cq6864.

[4] M. Megnidio-Tchoukouegno and J. A. Adedeji, "Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector," *Sustain.*, vol. 15, no. 3, 2023, doi: 10.3390/su15032014.

[5] A. Adefabi, S. Olisah, C. Obunadike, O. Oyetubo, E. Taiwo, and E. Tella, "Predicting Accident Severity: An Analysis of Factors Affecting Accident Severity Using Random Forest Model," *Int. J. Cybern. Informatics*, vol. 12, no. 6, pp. 107–121, 2023, doi: 10.5121/ijci.2023.120609.

[6] Ç. İ. Acı, G. Mutlu, M. Ozen, and M. Acı, "Enhanced Multi-Class Driver Injury Severity Prediction Using a Hybrid Deep Learning and Random Forest Approach," *Appl. Sci.*, vol. 15, no. 3, 2025, doi: 10.3390/app15031586.

[7] Z. Wu, A. Misra, and S. Bao, "Modeling Pedestrian Injury Severity: A Case Study of Using Extreme Gradient Boosting Vs Random Forest in Feature Selection," *Transp. Res. Rec.*, vol. 2678, no. 1, pp. 1–11, 2024, doi: 10.1177/03611981231170014.

[8] A. Nippani, D. Li, H. Ju, H. N. Koutsopoulos, and H. R. Zhang, "Graph Neural Networks for Road Safety Modeling: Datasets and Evaluations for Accident Analysis," *Adv. Neural Inf. Process. Syst.*, vol. 36, no. NeurIPS, 2023.

[9] X. Gao *et al.*, "Uncertainty-aware probabilistic graph neural networks for road-level traffic crash prediction," *Accid. Anal. Prev.*, vol. 208, no. October, p. 107801, 2024, doi: 10.1016/j.aap.2024.107801.

[10] H. Li and L. Chen, "Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories," *PLoS One*, vol. 20, no. 5 May, pp. 1–28, 2025, doi: 10.1371/journal.pone.0320656.

[11] J. Alotaibi, "Enhancing Traffic Accident Severity Prediction: Feature Identification Using Explainable AI," *Vehicles*, vol. 7, no. 2, p. 38, 2025, doi: 10.3390/vehicles7020038.

[12] P. Infante *et al.*, "Comparison of Statistical and Machine-Learning Models on Road Traffic Accident Severity Classification," *Computers*, vol. 11, no. 5, pp. 1–12, 2022, doi: 10.3390/computers11050080.

[13] E. Kuşkapan, M. Y. Çodur, and M. A. Sahraei, "Investigation of the Effect of Slope and Road Surface Conditions on Traffic Accidents Occurring in Winter Months: Spatial and Machine Learning Approaches," *Appl. Sci.*, vol. 14, no. 24, 2024, doi: 10.3390/app142411629.

[14] H. H. Pour *et al.*, "A Machine Learning Framework for Automated Accident Detection Based on Multimodal Sensors in Cars," *Sensors*, vol. 22, no. 10, pp. 1–21, 2022, doi: 10.3390/s22103634.

[15] R. Shafique, F. Rustam, S. Murtala, A. D. Jurcut, and G. S. Choi, "Advancing Autonomous Vehicle Safety: Machine Learning to Predict Sensor-Related Accident Severity," *IEEE Access*, vol. 12, no. January, pp. 25933–25948, 2024, doi: 10.1109/ACCESS.2024.3366990.

[16] I. Aldhari, M. Almoshaogeh, A. Jamal, F. Alharbi, M. Alinizzi, and H. Haider, "Severity Prediction of Highway Crashes in Saudi Arabia Using Machine Learning Techniques," *Appl. Sci.*, vol. 13, no. 1, 2023, doi: 10.3390/app13010233.

[17] B. Muktar and V. Fono, "Toward Safer Roads: Predicting the Severity of Traffic Accidents in Montreal Using Machine Learning," *Electron.*, vol. 13, no. 15, 2024, doi: 10.3390/electronics13153036.

[18] A. ÇELİK and O. SEVLİ, "Predicting Traffic Accident Severity Using Machine Learning Techniques," *Türk Doğa ve Fen Derg.*, vol. 11, no. 3, pp. 79–83, 2022, doi: 10.46810/tdfd.1136432.

[19] A. Sysoev, V. Klyavin, A. Dvurechenskaya, A. Mamedov, and V. Shushunov, "Applying Machine Learning Methods and Models to Explore the Structure of Traffic Accident Data," *Computation*, vol. 10, no. 4, pp. 1–12, 2022, doi: 10.3390/computation10040057.

[20] Z. C. Kuyumcu, H. Aslan, and N. Yurtay, "Casualty Analysis of the Drivers in Traffic Accidents in Turkey: A CHAID Decision Tree Model," *Appl. Sci.*, vol. 14, no. 24, 2024, doi: 10.3390/app142411693.

[21] M. A. K. Rifat, A. Kabir, and A. Huq, "An Explainable Machine Learning Approach to Traffic Accident Fatality Prediction," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 1905–1914, 2024, doi: 10.1016/j.procs.2024.09.704.

[22] J. Tang, Y. Huang, D. Liu, L. Xiong, and R. Bu, "Research on Traffic Accident Severity Level Prediction Model Based on Improved Machine Learning," *Systems*, vol. 13, no. 1, pp. 1–26, 2025, doi: 10.3390/systems13010031.

[23] A. K. N. Classifier, M. F. Febrianto, A. Priyatno, H. Adisty, and A. F. Saputri, "Prediksi Situasi Lalu Lintas Menggunakan Machine Learning Dengan," vol. 4221, no. April, pp. 28–34, 2024.

[24] K. Octavia, F. Yulianto, and T. A. Y. Siswa, "Penentuan jenis kecelakaan lalu linta di kota samarinda menggunakan metode," *Didakt. J. Ilm. PGSD FKIP Univ. Mandiri*, vol. 11, no. 1, pp. 212– 226, 2025.