PREDIKSI KECELAKAAN LALU LINTAS POLANDIA DENGAN XGBOOST, CATBOOST DAN RANDOM FOREST

Christabella Jocelynne Chandra [1] Carlouis Fernando Hariyadi [2] Novandry Aprilian [3] Lekrey Jacob Jerel Laipiopa [4]

[1] [2] [3] [4] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman St No.1, RT.6/RW.16, Tomang, Grogol Petamburan, Jakarta Barat Indonesia
email : christabella.535220166@stu.untar.id [1] carlouis.535220156@stu.untar.ac.id [2]
novandry.535220165@stu.untar.ac.id [3] lekrey.535220134@stu.untar.ac.id [4]

ABSTRAK

Kecelakaan lalu lintas tetap menjadi perhatian publik yang menonjol, dipengaruhi oleh kondisi sosial-ekonomi maupun lingkungan seperti cuaca. Studi ini bertujuan untuk memprediksi jumlah kecelakaan lalu lintas di Polandia berdasarkan faktor cuaca seperti kelembapan, suhu, dan curah hujan, serta variabel sosial-ekonomi seperti kepadatan penduduk, jumlah mobil penumpang, dan kepadatan jalan beraspal. Tiga algoritma ensemble learning, yaitu XGBoost, CatBoost, dan Random Forest, digunakan untuk mengevaluasi kinerja prediksi masing-masing. Dataset dibagi menggunakan Time Series Cross Validation, dan akurasi model dievaluasi menggunakan Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), serta koefisien determinasi (R^2). Hasil penelitian menunjukkan bahwa ketiga model memiliki performa yang baik, dengan Random Forest menghasilkan kinerja terbaik, diikuti oleh XGBoost dan CatBoost.

Key words

CatBoost, Ensemble Learning, Random Forest, Split, XGBoost,

1. Pendahuluan

Kecelakaan lalu lintas merupakan salah satu penyebab kematian pada publik. *World Health Organization* (WHO) memperkirakan bahwa 1.9 milyar orang meninggal karena kecelakaan lalu lintas [1]. Tingginya jumlah korban jiwa, cedera, serta dampak ekonomi yang muncul akibat kecelakaan lalu lintas dapat menjadi beban bagi sistem kesehatan dan sosial negara. Kompleksitas faktor penyebab kecelakaan membuat prediksi jumlah kecelakaan menjadi tantangan yang sulit, terutama karena dipengaruhi oleh faktor lingkungan, perilaku, kondisi cuaca, serta aspek sosial-ekonomi.

Salah satu kelompok faktor penting yang mempengaruhi kecelakaan lalu lintas adalah kondisi cuaca. Cuaca berpengaruh terhadap visibilitas, kondisi permukaan jalan, serta perilaku pengemudi [2]. Dalam penelitian ini, variabel cuaca yang digunakan meliputi kelembapan udara, suhu udara, dan curah hujan. Ketiga variabel ini merepresentasikan kondisi atmosfer yang dapat mempengaruhi jarak pandang, gesekan ban terhadap permukaan jalan, serta tingkat konsentrasi pengemudi. Misalnya, kelembapan tinggi dan hujan deras dapat menyebabkan licinnya permukaan jalan, sedangkan suhu ekstrem dapat mempengaruhi kondisi fisik pengemudi dan kendaraan [3].

Selain faktor cuaca, variabel sosial-ekonomi juga dapat menjadi peranan penting dalam menentukan tingkat risiko kecelakaan. Faktor-faktor tersebut mencerminkan kondisi sosial, kepadatan aktivitas, dan ketersediaan infrastruktur transportasi di suatu wilayah. Dalam penelitian ini, digunakan tiga indikator sosial-ekonomi, yaitu kepadatan penduduk, jumlah kendaraan penumpang, dan kepadatan jalan beraspal. Populasi yang padat cenderung meningkatkan interaksi antar pengguna jalan dan potensi tabrakan. Jumlah kendaraan penumpang mencerminkan tingkat mobilitas masyarakat, sedangkan kepadatan jalan beraspal menunjukkan ketersediaan infrastruktur transportasi yang dapat berpengaruh pada sebaran dan intensitas lalu lintas [4].

Telah ada banyak penelitian yang mengkaji pengaruh cuaca dan kondisi sosial-ekonomi terhadap kecelakaan lalu lintas seperti penelitian prediksi "*Bayesian Modeling of Traffic Accident Rates in Poland Based on Weather Conditions*" oleh Adam et al. yang dilakukan pada dataset sama dengan *Bayessian Modelling* yang menunjukkan hasil korelasi tinggi antar variabel cuaca dan sosio-ekonomi [5]. Penelitian ini akan menggunakan metode pembelajaran mesin (*Machine Learning*), seperti *Extreme Gradient Boosting* (XGBoost), *Categorical Boosting*

(CatBoost), dan Random Forest dan akan melakukan perbandingan kinerja model.

Penelitian mengenai "Analisis Pengaruh Kondisi Cuaca dan Jalan Terhadap Kecelakaan Lalu Lintas Menggunakan K-Means" oleh Muhammad et al. merupakan penelitian yang menganalisis pengaruh kondisi cuaca dan kondisi jalan terhadap terjadinya kecelakaan lalu lintas [2]. Pada penelitian ini, faktor cuaca akan ditambah dengan faktor sosio-ekonomi untuk memprediksi jumlah kecelakaan lalu lintas di Polandia.

Penelitian "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python" oleh Dwi et al. merupakan penelitian yang menggunakan metode XGBoost untuk melakukan prediksi menggunakan pola data historis kecelakaan lalu lintas [6]. Pada penelitian ini, faktor cuaca dan sosio-ekonomi telah ditambahkan agar dapat memprediksi jumlah kecelakaan lalu lintas lebih akurat dengan mempertimbangkan variabel yang dapat mempengaruhi jumlah terjadinya kecelakaan lalu lintas.

Pada penelitian "Jenis Kendaraan Sebagai Prediksi Tingkat Keparahan Kecelakaan Di Kabupaten Sleman Yogyakarta" oleh Eoudia et al. melakukan analisis mengenai faktor risiko individu yang mempengaruhi tingkat kekerasan kecelakaan menggunakan regresi logistik [7]. Pada penelitian ini, akan dilakukan regresi untuk memprediksi jumlah kecelakaan lalu lintas di Polandia berdasarkan faktor sosio-ekonomi keseluruhan dan bukan individu.

Pada penelitian "Pengembangan Model Prediksi Kecelakaan Lalu Lintas Pada Jalan Tol Purbaleunyi" oleh Lucky et al. melakukan prediksi kecelakaan lalu lintas berdasarkan arus lalu lintas dan faktor lingkungan lain pada satu entitas jalan [8]. Penelitian ini memprediksi jumlah kecelakaan lalu lintas pada keseluruhan negara Polandia yang dipecah menjadi 16 *voivodeship* dengan faktor cuaca dan sosio-ekonomi.

Metode pembelajaran mesin XGBoost, CatBoost, dan Random Forest dipilih karena memiliki kemampuan tinggi dalam memodelkan hubungan non-linear, melakukan seleksi fitur, dan meningkatkan akurasi prediksi. Penelitian ini dilakukan dengan tujuan membandingkan metode-metode untuk mengetahui metode terbaik untuk prediksi jumlah kecelakaan lalu lintas.

2. Metode Penelitian

Bagian ini akan menjelaskan mengenai metode yang digunakan dalam penelitian ini yakni metode algoritma pembelajaran ensembel (*Ensemble Learning Algorithm*) XGBoost, Random Forest dan CatBoost. Metode-metode tersebut dipilih karena mampu menangani hubungan non-linear antar variabel. Sehingga, metode tersebut dapat menemukan hubungan antar faktor-faktor yang ditetapkan pada penelitian ini. Metode tersebut juga relatif tahan terhadap *outlier* dan *noise* yang umum dijumpai dalam data cuaca dan sosio-ekonomi [9].

Dengan menggunakan metode ini, penelitian ini akan memprediksi jumlah kecelakaan lalu lintas di negara Polandia. Penilaian hasil kinerja ketiga metode akan dievaluasi menggunakan metrik evaluasi MSE, RMSE, MAE, dan R^2.

2.1 *Pre-processing* Dataset

Dataset akan pertama di agregasi menjadi *weekly* per *voivodeship*. Ada sebanyak 16 *voivodeship* sehingga jumlah data berupa 2.894 baris. Dataset lalu akan digabungkan, dengan data kategorikal seperti *voivodeship*, yakni divisi administratif di Polandia, dilakukan *Label Encoding* di mana setiap kategori digantikan dengan nomor koresponden. Tabel *voivodeship* dengan *Label* nya dapat dilihat pada Tabel 2.1.

Tabel 2.1 Tabel *Voivodeship*

Nama *Voivodeship*	*Label Encoding*
Dolnośląskie	0
Kujawsko-pomorskie	1
Lubelskie	2
Lubelskie	3
Mazowieckie	4
Małopolskie	5
Opolskie	6
Podkarpackie	7
Podlaskie	8
Pomorskie	9
Warmińsko-mazurskie	10
Wielkopolskie	11
Zachodniopomorskie	12
Łódzkie	13
Śląskie	14
Świętokrzyskie	15

Setelah *Label Encoding*, dataset akan dilakukan fitur *lag* sebanyak 1 minggu dan 4 minggu agar model dapat mempelajari pola tren lebih baik dan karena faktor cuaca dan sosio-ekonomi dapat memiliki dampak untuk masa depannya. Dataset juga akan ditambahkan *Rolling Mean* yakni teknik untuk menghaluskan data dengan cara menghitung nilai rata-rata dari sekumpulan data dalam sebuah jendela ukuran tertentu yang kemudian digeser secara berurutan di sepanjang data. *Rolling Mean* ini dilakukan per 3 minggu dan 7 minggu.

Dataset lalu akan dipecah menggunakan *Time Series Cross Validation* yakni teknik pembagian data yang dirancang untuk data berurutan waktu (*Time Series*). Hal ini dilakukan agar tiap model dapat dievaluasi sebanyak 5 *split* di berbagai titik waktu. Pada pemecahan berikut, per potongan berukuran sekitar 483 baris. *Split* pertama akan menggunakan potongan 1 sebagai data latih dan potongan 2 sebagai data uji sedangkan potongan lain tidak digunakan, *split* kedua akan menggunakan potongan 1 dan 2 sebagai data latih dengan potongan 3 sebagai data uji dan proses ini akan diulang hingga terdapat *split* 5 di

mana potongan 1, 2, 3, 4 akan digunakan sebagai data latih dan potongan 5 akan digunakan sebagai data uji.

2.2 XGBoost

XGBoost adalah algoritma pembelajaran ensembel yang berdasarkan pembelajaran mesin. Algoritma ini digunakan untuk prediksi yang memerlukan akurasi tinggi dan dikembangkan oleh Tianqi Chen et al. sebagai perluasan dari *Gradient Boosting Decision Tree* (GBDT). Tujuan algoritma ini adalah untuk meningkatkan kinerja dan efisiensi GBDT dengan menambahkan secara berurutan serangkaian pohon keputusan (*Weak Learner*). Setiap pohon baru dibangun untuk memperbaiki kesalahan pohon sebelumnya sehingga prediksi akhir merupakan gabungan dari semua pohon [10]. XGBoost menggunakan teknik penyusutan yaitu penyesuaian laju pembelajaran dan *Subsampling* kolom untuk membatasi *Overfitting*. Selain itu, program ini juga dibangun dengan skala besar, dilengkapi dengan optimasi *cache*, algoritma yang sadar kepadatan rendah, dan dukungan pemrosesan paralel, sehingga lebih cepat dibandingkan implementasi GBDT reguler [11].

XGBoost umumnya bekerja dengan menambahkan pohon keputusan secara berulang. Pada setiap iterasi, model akan menambahkan pohon baru untuk meminimalkan fungsi objektif keseluruhan. Prinsip *Gradient Boosting* memanfaatkan algoritma *Gradient Descent* untuk meminimalkan arah penurunan tercuram dari fungsi kerugian, sehingga model secara bertahap meningkatkan prediksinya dibandingkan dengan prediksi sebelumnya [10]. XGBoost memiliki beberapa peningkatan dibandingkan dengan algoritma *Boosting* konvensional, yang meliputi hal-hal berikut:

1. Penggunaan gradien orde dua.
   Selain gradien pertama, XGBoost menggunakan gradien orde dua (hessian) dalam estimasi *Update*, sehingga perhitungan lebih akurat dan konvergen lebih cepat.
2. Regularisasi L1 dan L2.
   Untuk mengendalikan kompleksitas model, XGBoost menambahkan *Lasso Penalty* (L1) dan *Ridge* (L2) pada struktur pohon dan bobot daun [11]. Regularisasi ini mencegah *Overfitting* dan membantu generalisasi.
3. *Shrinkage* dan *Subsampling*.
   Setelah setiap iterasi, XGBoost mengalikan bobot pohon baru dengan faktor η (*learning rate)* untuk mengecilkan pengaruh masing-masing pohon dan memberi ruang bagi pohon berikutnya. Selain itu, pemilihan subset fitur atau sampel meingkatkan diversifikasi pohon dan efisiensi komputasi [11].

Model prediksi XGBoost dapat dinyatakan sebagai penjumlahan K pohon keputusan :

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Dengan setiap f_k merupakan fungsi dari pohon regresi dalam ruang F (Himpunan semua pohon). Tujuan pelatihan adalah meminimalkan *Regularized Objective Function.* Tianqi Chen et al. mendefinisikan fungsi objektif sebagai [11].

$$L(\phi) = \sum_{i} l(\widehat{y_i}, y_i) + \sum_{i} \Omega(f_k) \tag{2}$$

Di mana $l(\widehat{y_i}, y_i)$ adalah fungsi *Loss* diferensiabel, $\widehat{y_i}$ adalah prediksi, adalah target, dan $\Omega(f_k)$ adalah fungsi regularisasi untuk pohon ke-k. Regularisasi diukur melalui jumlah daun T dan norma L2 bobot daun ω:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda|\omega|^2 \tag{3}$$

Dengan γ sebagai penalti jumlah daun dan λ sebagai parameter regularisasi L2 [11]. Ketika parameter regularisasi diatur ke nol, formulasi ini kembali ke metode *Gradient Boosting* tradisional.

2.3 Random Forest

Random Forest merupakan algoritma yang tergolong dalam metode pembelajaran ensembel yang menggunakan teknik *Bagging* (*Bootstrap Aggregating*). Metode ini akan membangun sejumlah besar pohon keputusan (*Decision Trees*) yang dilatih secara bersamaan, sehingga secara signifikan meningkatkan kekuatan dan stabilitas prediksi [12].

Metode ini dapat digunakan untuk melakukan regresi untuk memperkirakan nilai numerik dan klasifikasi untuk mengelompokkan data ke dalam kategori tertentu. Dalam proses pelatihan, setiap pohon dibuat dari *subset* data pelatihan yang diambil secara acak, dan pada setiap pemisahan simpul, hanya *subset* fitur yang dipilih secara acak untuk dipertimbangkan, sebuah proses yang menambahkan keragaman dan membantu mengatasi kelemahan utama dari pohon keputusan tunggal, yaitu *Overfitting* [13].

Ketika digunakan untuk klasifikasi, Random Forest menghasilkan prediksi berdasarkan suara mayoritas (*Majority Vote*) dari seluruh pohon yang dibangun. Sementara itu, untuk tugas regresi, hasil akhirnya diperoleh dengan mengambil rata-rata dari prediksi yang dihasilkan oleh setiap pohon [14]. Secara matematis, prediksi regresi dapat ditulis sebagai:

$$Y = \frac{1}{N}\sum_{n=1}^{N} h_i\,(X) \tag{4}$$

Di mana Y adalah prediksi akhir, N adalah jumlah pohon dalam hutan, dan $h_i(X)$ adalah prediksi dari pohon ke-i [13]. Meskipun Random Forest mungkin memiliki akurasi yang lebih rendah dibandingkan beberapa pendekatan seperti *Extreme Gradient Boosting Trees*, keunggulannya terletak pada ketahanannya terhadap perbedaan skala fitur, kemampuannya menangani fitur yang tidak relevan, dan interpretasi model yang relatif mudah [14].

2.4 CatBoost

CatBoost adalah algoritma *Gradient Boosting* berbasis pohon keputusan yang dikembangkan oleh Yandex pada tahun 2017 [15]. Algoritma ini bertujuan untuk mengatasi kelemahan *Boosting* konvensional, khususnya dalam menangani fitur kategorial dan mengurangi bias prediksi [16]. Berbeda dengan algoritma lain yang membutuhkan *One-hot Encoding*, CatBoost memperkenalkan *Ordered Target Statistics* dan *Ordered Boosting* untuk mencegah *Target Leakage* serta mengurangi *Prediction Shift*, sehingga menghasilkan model yang lebih stabil dan mampu melakukan generalisasi lebih baik [16] [17]. Selain itu, CatBoost menggunakan pohon simetris (*Oblivious Decision Trees*), yang menjadikan proses inferensi lebih efisien dan mendukung paralelisasi pada CPU maupun GPU [18]. Secara matematis, prediksi CatBoost dinyatakan sebagai penjumlahan sejumlah pohon keputusan:

$$y = \sum_{k=1}^{k} f_k(x_i) \ , \ f_k \epsilon F \qquad (5)$$

Pada rumus di atas, F adalah himpunan semua pohon regresi. Fungsi objektif yang diminimalkan adalah:

$$L(\phi) = \sum_{n=1}^{n} \ell(yi, \widehat{yi}) + \sum_{k=1}^{k} \Omega(f_k) \qquad (6)$$

Fungsi *loss* yang digunakan bersifat diferensiabel seperti log *loss*, sedangkan regularisasi dipakai untuk mengendalikan kompleksitas pohon. CatBoost juga memanfaatkan gradien orde pertama dan Hessian orde kedua untuk mempercepat konvergensi pelatihan.

$$g_i = \frac{\vartheta \ell(y_i \ , \ \widehat{y_i})}{\vartheta \widehat{y_i}} \ , \ h_i = \frac{\vartheta^2 \ell(y_i \ , \ \widehat{y_i})}{\vartheta \widehat{y_i}} \qquad (7)$$

Agar dapat memperbarui bobot pada iterasi berikutnya, regularisasi model akan dilakukan melalui penalti jumlah daun T dan bobot daun *wi* , yang dituliskan sebagai:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \qquad (8)$$

Dengan γ sebagai penalti jumlah daun dan λ koefisien regularisasi L2 [18]. Dengan kombinasi inovasi tersebut, CatBoost terbukti unggul pada dataset tabular berskala besar dengan banyak fitur kategorial, dan telah digunakan secara luas pada berbagai penelitian mulai dari klasifikasi jaringan sungai [19], prediksi kanker serviks [20], hingga deteksi penyakit ginjal kronis [21].

2.5 Metrik Evaluasi

Untuk menilai kinerja para model, akan digunakan metrik evaluasi berikut yakni:

1. MSE (*Mean Squared Error*)
   MAE berupa metode pengukuran rata-rata dari kuadrat selisih antara nilai sebenarnya dan hasil nilai prediksi model. MAE dapat digunakan untuk menunjukkan rata-rata kesalahan absolut antar hasil prediksi dan nilai asli. Rumus MAE dapat dihitung dengan persamaan berikut [22]:

   $$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - x| \qquad (9)$$

   Di mana n berupa jumlah data, x_i adalah nilai hasil prediksi, dan x merupakan nilai sebenarnya.
2. RMSE (*Root Mean Square Error*)
   RMSE berupa metode yang merupakan akar kuadrat dari MSE untuk mengembalikan kesalahan ke satuan asli data. Rumus dari RMSE dapat dilihat sebagai berikut [22]:

   $$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{x_i})^2} \qquad (10)$$

   Di mana x_i merupakan nilai sebenarnya dan $\widehat{x_i}$ merupakan nilai prediksi.
3. MAE (*Mean Absolute Error*)
   MAE merupakan metode yang mengukur nilai rata-rata selisih absolut antara nilai aktual dan nilai prediksi dengan tujuan menunjukkan seberapa besar kesalahan rata-rata. Rumus MAE dapat dilihat sebagai berikut [23]:

   $$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - \widehat{x_i}| \qquad (11)$$

   Di mana x_i merupakan nilai sebenarnya dan $\widehat{x_i}$ merupakan nilai prediksi.
4. R^2 (Koefisien Determinasi)
   R^2 merupakan metode yang mengukur seberapa baik model menjelaskan variasi data aktual. Nilainya berkisar di antara 0 – 1. Rumus R^2 dapat dilihat sebagai berikut [23]:

$$R^2 = 1 - \frac{SSres}{SStot} \quad (12)$$

Di mana $SSres$ merupakan jumlah kuadrat residual, yaitu perbedaan nilai aktual dengan nilai yang diprediksi dan SStot merupakan jumlah kuadrat total, yaitu perbedaan nilai aktual dengan nilai rata-rata aktual.

3. Hasil Percobaan

Bagian ini akan menunjukkan hasil pengujian tiap model dengan dataset yang sudah melalui *pre-processing* dan telah di *split* sebanyak 5 kali. Model akan diuji dan mengeluarkan hasil evaluasi untuk tiap *split* lalu hasil evaluasi tiap *split* akan di hitung rata-ratanya untuk menentukan hasil kinerja model secara keseluruhan.

3.1 Dataset yang Digunakan

Penelitian ini menggunakan dataset yang dikumpulkan pada penelitian Ł. Faruga et al. [24] Pada penelitian tersebut, dataset dikumpulkan dengan tiga faktor utama yaitu data kecelakaan lalu lintas, data cuaca dan data sosio-ekonomik. Dataset kecelakaan dikumpulkan dari SEWIK (*Accident and Collision Recording System*) yang merupakan database nasional Polandia yang di supervisi oleh kepolisian nasional Polandia, dataset cuaca diambil dari IMGW (*Institute of Metereology and Water Management*) yang merupakan sumber observasi cuaca Polandia yang dioperasikan oleh kementerian infrastruktur Polandia dan dataset sosio-ekonomik dikumpulkan melalui GUS (*Local Data Bank of Statistics Poland*) yang merupakan sumber primer untuk data nasional Polandia.

Pada penelitian ini, ketiga dataset akan digunakan dengan variabel-variabel berikut dipilih:

1. Kelembapan Udara
   Kelembapan udara menunjukkan jumlah uap air yang terkandung di atmosfer pada suatu wilayah dan waktu tertentu yang dalam bentuk persentase (%). Variabel ini dipilih karena kelembapan udara dapat menyebabkan jalan licin dan cuaca berkabut atau hujan yang dapat mengganggu pengemudi jalan.
2. Suhu Udara
   Suhu udara adalah ukuran tingkat panas atau dingin suatu wilayah, pada dataset ini, metrik pengukurannya menggunakan celsius. Variabel ini dipilih karena dapat mempengaruhi kondisi permukaan jalan, pelaku pengemudi, serta volume lalu lintas.
3. Curah Hujan
   Curah hujan menunjukkan jumlah air yang turun dari atmosfer dalam bentuk hujan yang dinyatakan dalam milimeter (mm) per satuan waktu. Variabel ini dipilih karena hujan dapat mengganggu visibilitas pengemudi lalu lintas, berkurangnya traksi ban pada jalan, serta peningkatan jarak pengereman.
4. Kepadatan Penduduk
   Kepadatan penduduk mengukur jumlah penduduk yang tinggal di suatu area yang dinyatakan sebagai jumlah penduduk per kilometer (jiwa/km^2). Variabel ini dipilih karena dapat mencerminkan intensitas aktivitas manusia dan tingkat urbanisasi di suatu wilayah tertentu.
5. Jumlah Kendaraan Penumpang
   Jumlah kendaraan penumpang menggambarkan total kendaraan pribadi yang beroperasi dalam suatu wilayah. Variabel ini dipilih karena menunjukkan hubungan dengan interaksi antar kendaraan di jalan yang dapat menyebabkan kecelakaan lalu lintas.
6. Kepadatan Jalan Beraspal
   Kepadatan jalan beraspal menunjukkan panjang jalan beraspal per 100 kilometer (km/100km^2). Variabel ini dipilih karena menunjukkan ketersediaan infrastruktur darat di suatu wilayah yang dapat mempengaruhi frekuensi terjadinya kecelakaan lalu lintas.
7. *Voivodeship*
   Voivodeship merupakan unit wilayah administratif tertinggi di Polandia, setara dengan provinsi di Indonesia. Terdapat 16 *voivodeship* di Polandia yang memiliki karakteristik geografis, ekonomi, dan sosial yang berbeda-beda. Variabel *voivodeship* digunakan sebagai variabel kategorikal yang dapat membedakan setiap wilayah analisis.

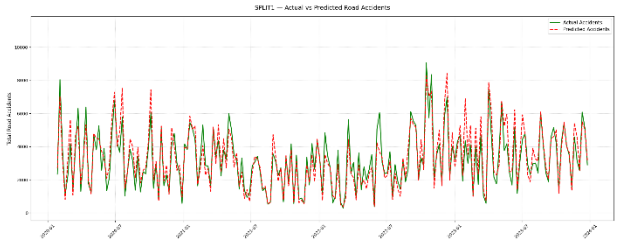
3.2 Hasil XGBoost

XGBoost merupakan model yang memiliki kinerja kedua terbaik jika dibandingkan dengan hasil model Random Forest dan CatBoost. Hasil XGBoost menunjukkan bahwa model cocok digunakan untuk prediksi jumlah kecelakaan lalu lintas di Polandia. dengan hasil evaluasi per *split* dapat di lihat pada Tabel 3.1 berikut:

Tabel 3.1 Hasil XGBoost

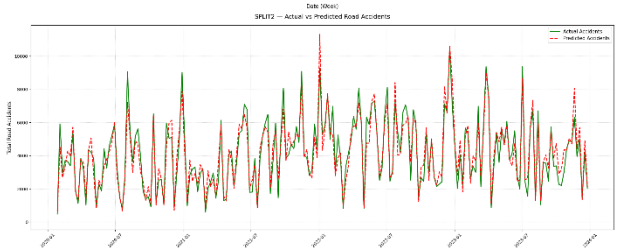
Split	MSE	RMSE	MAE	R^2
1	171642.80	414.30	269.60	0.8467
2	167806.04	409.64	286.73	0.8957
3	136211.76	369.07	233.08	0.8873
4	106153.38	325.81	204.20	0.8808
5	131673.70	362.87	234.25	0.9102
Rata2	142697.54	376.34	245.57	0.88

Pada hasil evaluasi di atas, dapat dilihat bahwa pada *split* pertama, model sudah cukup baik dan dapat memprediksi variasi sekitar 84.67% jumlah kecelakaan dengan kesalahan rata-rata sekitar 269. Grafik dari *split* berikut dapat dilihat di gambar di bawah:



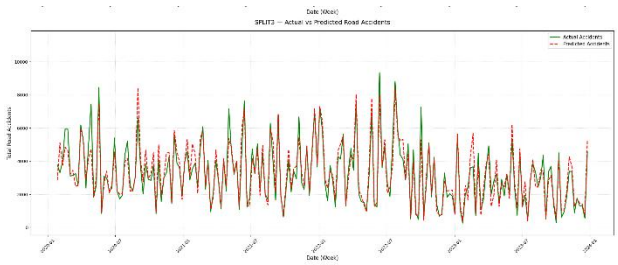
Gambar 3.1 *Split* 1 XGBoost

Split kedua menunjukkan bahwa kinerja model meningkat dengan bertambahnya data uji menjadi 89.57% tetapi MAE mengalami peningkatan sedikit. Grafik dari *split* tersebut dapat dilihat di gambar di bawah:



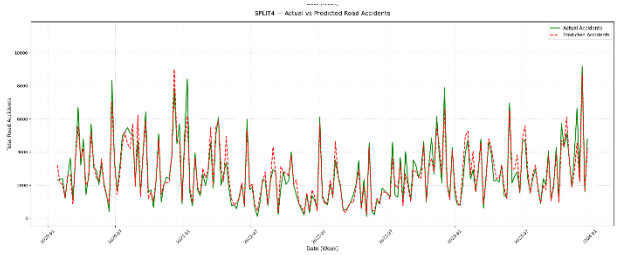
Gambar 3.2 *Split* 2 XGBoost

Split ketiga menunjukkan kinerja yang naik signifikan dibanding *split* pertama dan kedua. Kesalahan model menurun menjadi 233 dan nilai R^2 sekitar 88.73%. Grafik dari *split* tersebut dapat dilihat di gambar di bawah:



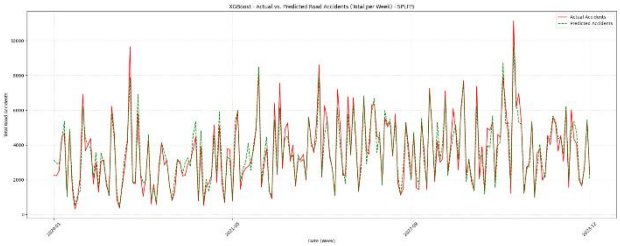
Gambar 3.3 *Split* 3 XGBoost

Split keempat merupakan *split* dengan kesalahan terkecil sehingga menunjukkan nilai prediksi mendekati nilai aktual meskipun R^2 sedikit menurun. Grafik dari *split* tersebut dapat dilihat di gambar di bawah:



Gambar 3.4 *Split* 4 XGBoost

Split terakhir merupakan *split* dengan hasil terbaik. Grafik untuk *split* ini dapat dilihat sebagai berikut:



Gambar 3.5 *Split* 5 XGBoost

Model mampu memprediksi 91.2% variasi jumlah kecelakaan lalu lintas dan dengan tingkat kesalahan prediksi rata-rata hanya sekitar 234. Dengan keseluruhan, model XGBoost dapat menunjukkan ketahanan terhadap variasi faktor cuaca dan sosio-ekonomi. Pada rata-rata dari keseluruhan evaluasi *split*, model XGBoost mencapai tingkat akurasi tinggi dengan nilai R^2 sebesar 0.88 dan rata-rata kesalahan sekitar 245 kasus. Model telah dapat menangkap pola kompleks antara faktor cuaca dan sosial-ekonomi terhadap jumlah kecelakaan lalu lintas. Hal ini dapat terjadi karena mekanisme *Boosting* yang memperbaiki kesalahan prediksi secara bertahap.

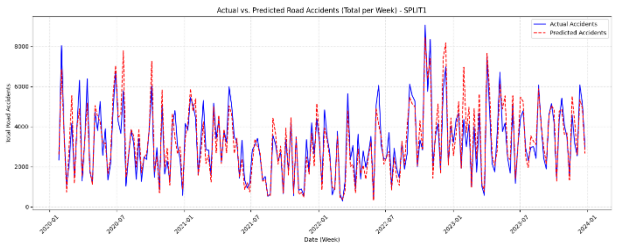
3.3 Hasil Random Forest

Random Forest memiliki hasil yang lebih stabil di seluruh variasi kondisi dibanding XGBoost dan merupakan model dengan kinerja terbaik antara XGBoost dan CatBoost. Hasil dari evaluasi *per split* dapat dilihat pada tabel berikut:

Tabel 3.2 Hasil Random Forest

Split	MSE	RMSE	MAE	R^2
1	151757.97	389.56	252.06	0.8644
2	148648.87	385.55	273.35	0.9076
3	125596.74	354.40	214.21	0.8961
4	104444.37	323.18	199.71	0.8827
5	125257.00	353.92	225.54	0.9146
Rata2	131140.99	361.32	232.97	0.89

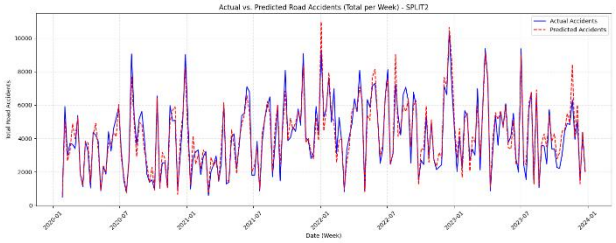
Pada *split* pertama, model mampu menjelaskan sebanyak 86.44% variasi jumlah kecelakaan dengan kecelakaan rata-rata sebanyak 252. Grafik untuk *split* ini dapat dilihat di gambar berikut:



Gambar 3.6 *Split* 1 Random Forest

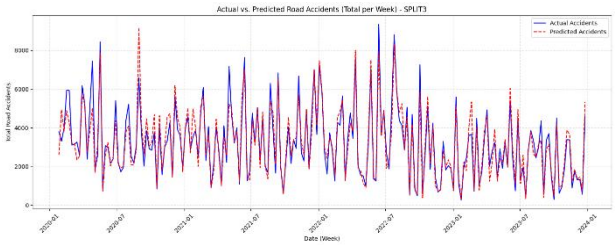
Split kedua memiliki kinerja signifikan lebih tinggi sebanyak 90.76% tetapi nilai MAE sedikit naik yang dapat menandakan adanya beberapa titik ekstrem

(*Outlier*). Grafik untuk *split* ini dapat dilihat pada grafik berikut:



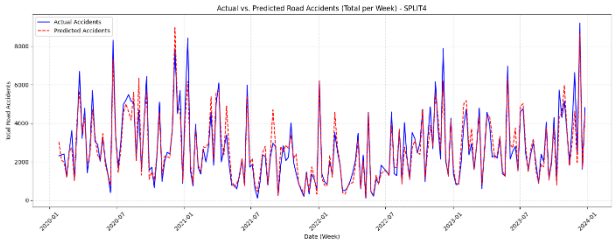
Gambar 3.7 *Split* 2 Random Forest

Split ketiga menunjukkan penurunan *error* yang cukup signifikan yakni 214 dengan nilai R^2 yang cukup tinggi. Hal ini menunjukkan model menangkap hubungan non-linear antar variabel dengan baik. Grafik untuk *split* ini dapat dilihat pada grafik berikut:



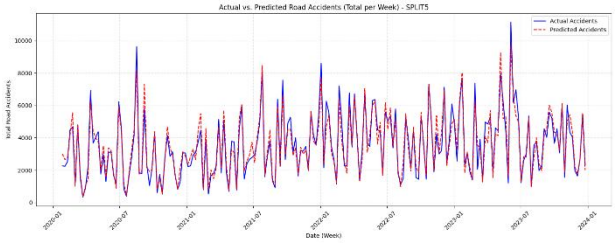
Gambar 3.8 *Split* 3 Random Forest

Split keempat menunjukkan tingkat kesalahan yang terendah. Hal ini sama dengan hasil XGBoost pada *split* keempat yang dapat menandakan distribusi data yang lebih konsisten. Grafik untuk memvisualisasikan *split* ini dapat dilihat di gambar berikut:



Gambar 3.9 *Split* 4 Random Forest

Split kelima memiliki hasil terbaik dari antara semua *split* dengan grafiknya dapat dilihat sebagai berikut:



Gambar 3.10 *Split* terakhir Random Forest

Split ini menghasilkan R^2 tertinggi yakni 91.46% yang menandakan model Random Forest memiliki kemampuan generalisasi yang baik dan tangguh terhadap variasi antar subset data. Rata-rata dari keseluruhan evaluasi *split* menunjukkan bahwa model Random Forest menunjukkan performa prediksi yang kuat dan stabil dengan kesalahan rata-rata sekitar 233 kasus dan kemampuan menjelaskan variasi data sebesar 89%. Hal ini menegaskan bahwa metode ensembel berbasis *Bagging* ini efektif dalam memodelkan hubungan kompleks antara faktor cuaca dan sosial-ekonomi terhadap tingkat kecelakaan lalu lintas di Polandia.

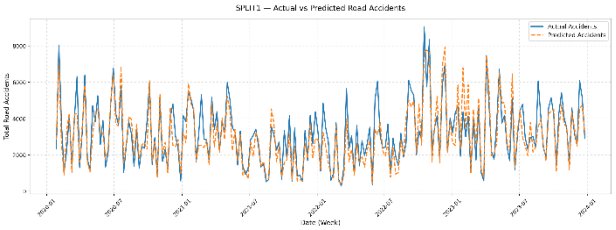
3.4 Hasil CatBoost

CatBoost memiliki hasil yang baik tetapi tidak sebaik model-model lainnya seperti XGBoost dan Random Forest. Hasil evaluasi model ini per *split* dapat dilihat pada tabel berikut:

Tabel 3.3 Hasil CatBoost

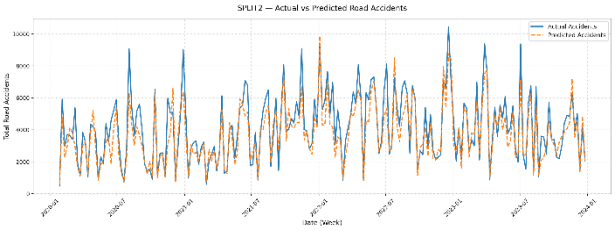
Split	MSE	RMSE	MAE	R^2
1	191674.64	437.81	290.78	0.8288
2	181755.49	426.33	298.08	0.8870
3	141240.47	375.82	243.67	0.8832
4	99459.85	315.37	203.39	0.8883
5	126740.85	356.01	228.37	0.9135
Rata2	148174.26	382.27	252.86	0.88

Pada *split* pertama, model menunjukkan performa terendah dibanding *split* lainnya dengan kesalahan prediksi yang relatif tinggi. Hal ini dapat terjadi karena CatBoost sensitif terhadap subset data dengan ketidakseimbangan distribusi fitur atau karena jumlah data yang tidak banyak (~400) pada *split* pertama. Visualisasi untuk *split* ini dapat dilihat pada gambar berikut:



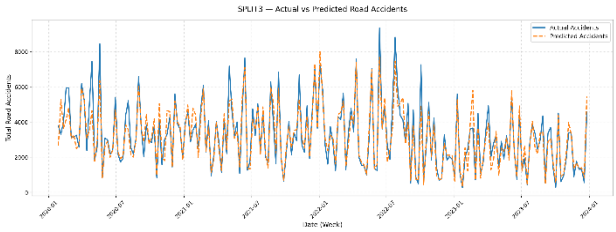
Gambar 3.10 *Split* 1 CatBoost

Pada *split* kedua, performa meningkat dan menunjukkan bahwa model lebih mampu mengenali pola antar variabel meskipun frekuensi terjadinya kesalahan sedikit tinggi. Grafik dari *split* tersebut dapat dilihat pada gambar berikut:



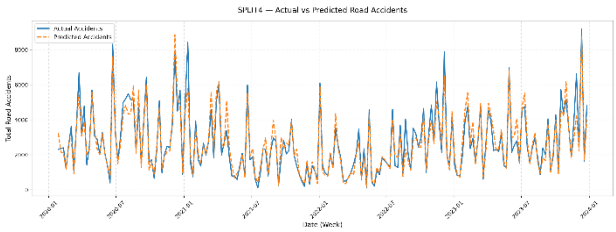
Gambar 3.11 *Split* 2 CatBoost

Pada *split* ketiga, kesalahan semakin kecil dan nilai R^2 sebanyak 88.32% yang dapat menunjukkan bahwa model efektif dalam menangkap hubungan non-linear. Grafik untuk *split* ini dapat dilihat pada gambar berikut:



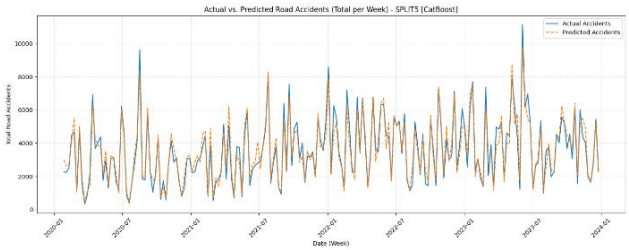
Gambar 3.12 *Split* 3 CatBoost

Pada *split* keempat, performa model sudah lebih baik dibanding *split* sebelumnya. Hal ini dapat menunjukkan bahwa CatBoost membutuhkan data latih lebih banyak agar dapat lebih menangkap hubungan antar variabel. Jumlah kesalahan pada *split* ini juga paling sedikit yang seragam dengan model-model lain pada *split* keempat. Grafik dari *split* tersebut dapat dilihat pada gambar berikut:



Gambar 3.13 *Split* 4 CatBoost

Split kelima menunjukkan hasil terbaik dengan R^2 tertinggi yakni sebanyak 91.35% dan nilai MAE terendah sebanyak 228.37. Grafik dari *split* ini dapat dilihat di gambar di bawah:



Gambar 3.14 *Split* 5 CatBoost

Hasil ini menunjukkan bahwa CatBoost dapat menyesuaikan diri dengan pola kompleks dan stabil sehingga menghasilkan prediksi paling akurat dibanding model lain. Dengan keseluruhan, nilai R^2 yang tinggi dan kesalahan relatif rendah menunjukkan bahwa model CatBoost dapat berkerja baik untuk prediksi dataset ini akan tetapi tidak sebaik model lainnya jika jumlah data tidak banyak.

3.4 Perbandingan Kinerja Model

Hasil evaluasi menunjukkan bahwa ketiga model memiliki performa yang tinggi dengan nilai R^2 sekitar 88%. Model dengan kinerja terbaik adalah model Random Forest dengan rata-rata nilai R^2 sebesar 89%, nilai MSE sebesar 131,140.99, RMSE sebesar 361.32, MAE sebesar 232.97. Model XGBoost menunjukkan kinerja yang kedua terbaik dengan rata-rata nilai R^2 sebesar 88%, nilai MSE sebesar 142,697.54, RMSE sebesar 376.34, dan MAE sebesar 245.57. Dengan model CatBoost memiliki hasil terburuk antar ketiga model dengan rata-rata nilai R^2 sebesar 88%, nilai MSE sebesar 148,174.26, RMSE sebesar 382.27, dan MAE sebesar 252.86.

4. Kesimpulan

Penelitian ini bertujuan untuk memprediksi jumlah kecelakaan lalu lintas di Polandia berdasarkan faktor kondisi cuaca (humiditas, temperatur, dan presipitasi) dan faktor sosial-ekonomi (kepadatan penduduk, jumlah kendaraan penumpang, dan kepadatan jalan beraspal) menggunakan tiga algoritma pembelajaran ensembel, yaitu XGBoost, CatBoost, dan Random Forest. Ketiga model dievaluasi menggunakan metrik MSE, RMSE, MAE, dan koefisien determinasi (R^2) untuk menilai tingkat akurasi dan kemampuan generalisasi terhadap data.

Hasil dari ketiga model menunjukkan bahwa meskipun nilai CatBoost terburuk, ketiga model menunjukkan perbedaan performa yang cukup kecil. Hal ini menandakan bahwa seluruh metode dapat melakukan prediksi dengan akurasi yang baik. Secara keseluruhan, model Random Forest merupakan model yang ter-optimal digunakan untuk prediksi kecelakaan lalu lintas karena memiliki prediksi tertinggi dan ketahanan yang baik terhadap *noise* maupun *outlier* yang dapat muncul pada dataset cuaca dan sosio-ekonomi. Akan tetapi, XGBoost dan CatBoost memberikan hasil yang cukup kompetitif.

Untuk meningkatkan kinerja model serta memperlihatkan perbedaan performa antar algoritma secara lebih signifikan, penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan mencakup variabel sosial-ekonomi yang lebih beragam. Penambahan faktor seperti tingkat pendidikan, pendapatan rata-rata, serta kepadatan lalu lintas di tiap wilayah diharapkan dapat memperkaya kompleksitas data sehingga model dapat mempelajari pola hubungan yang lebih representatif terhadap kondisi nyata.

REFERENSI

[1] WHO, “Road traffic injuries,” World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] M. Romadon and R. Passarella, “ANALISIS PENGARUH KONDISI CUACA DAN JALAN TERHADAP KECELAKAAN LALU LINTAS MENGGUNAKAN K-MEANS,” Dec. 2024, Accessed:

Oct. 07, 2025. [Online]. Available: http://repository.unsri.ac.id/id/eprint/162335

[3] I. J. Effendi, A. N. N. Rizki, D. O. R. Rahman, and B. Fadel, "Pendekatan Descriptive Analysis Berbasis Data Untuk Mengevaluasi Kecelakaan Lalu Lintas Di Indonesia," *J. Inform. Ilmu Komput. dan Sist. Inf.*, vol. 2, no. 3, 2024, Accessed: Oct. 07, 2025. [Online]. Available: https://animator.uho.ac.id/index.php/journal/article/view/1241

[4] K. Zhang, S. Wang, C. Song, S. Zhang, and X. Liu, "Spatiotemporal Heterogeneity Analysis of Provincial Road Traffic Accidents and Its Influencing Factors in China," *Sustain. 2024, Vol. 16, Page 7348*, vol. 16, no. 17, p. 7348, Aug. 2024, doi: 10.3390/SU16177348.

[5] A. Filapek, Ł. Faruga, and J. Baranowski, "Bayesian Modeling of Traffic Accident Rates in Poland Based on Weather Conditions," *Appl. Sci. 2025, Vol. 15, Page 7332*, vol. 15, no. 13, p. 7332, Jun. 2025, doi: 10.3390/APP15137332.

[6] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, "Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, p. 188, Oct. 2020, doi: 10.24843/JIM.2020.V08.I03.P04.

[7] B. L. Fauzan, T. Agustin, A. Musthofiah, and H. Mahmudah, "Prediksi Klasifikasi Kecelakaan Lalu Lintas di Kota Surakarta dengan Menggunakan Metode Regresi Logistik Multinomial," *Sustain. Civ. Build. Manag. Eng.*, vol. 1, no. 4, pp. 9–9, Aug. 2024, doi: 10.47134/SCBMEJ.V1I4.3159.

[8] L. A. (Lucky) Rakhmat, A. (Aine) Kusumawati, R. B. (Russ) Frazila, and S. (Sri) Hendarto, "Pengembangan Model Prediksi Kecelakaan Lalu Lintas Pada Jalan Tol Purbaleunyi," *J. Tek. Sipil ITB*, vol. 19, no. 3, pp. 277–288, Dec. 2012, doi: 10.5614/JTS.2012.19.3.8.

[9] S. B. Prakash, M. Raj, and A. P, "EARTHQUAKE PREDICTION USING RANDOM FOREST TREE WITH XGBOOST AND CATBOOST," *I C N K A I – 2 K 2 5 Int. Lev. Conf. Nexus Knowl. Using AI IoT*, Jun. 2025, Accessed: Oct. 07, 2025. [Online]. Available: https://secp.prabodhanamfoundation.org/csn/index.php/csn-secp/article/view/91

[10] S. Markovic *et al.*, "Application of XGBoost model for in-situ water saturation determination in Canadian oil-sands by LF-NMR and density data," *Sci. Rep.*, vol. 12, no. 1, pp. 1–14, Dec. 2022, doi: 10.1038/S41598-022-17886-6;SUBJMETA.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

[12] N. K. Dewi, U. D. Syafitri, and S. Y. Mulyadi, "PENERAPAN METODE RANDOM FOREST DALAM DRIVER ANALYSIS," *FORUM Stat. DAN KOMPUTASI*, vol. 16, no. 1, pp. 35–43, 2011, Accessed: Oct. 07, 2025. [Online]. Available: https://journal.ipb.ac.id/statistika/article/view/5443

[13] M. S. Efendi, Sarwido, and A. K. Zyen, "Penerapan Algoritma Random Forest Untuk Prediksi Penjualan Dan Sistem Persediaan Produk," *Resolusi Rekayasa Tek. Inform. dan Inf.*, vol. 5, no. 1, pp. 12–20, Sep. 2024, doi: 10.30865/RESOLUSI.V5I1.2149.

[14] D. Aqsha, "PERBANDINGAN KINERJA ALGORITMA EXTREME GRADIENT BOOSTING DAN RANDOM FOREST UNTUK PREDIKSI HARGA RUMAH DI JABODETABEK," *J. Ilmu Komput. dan Sist. Inf.*, vol. 13, no. 1, Jan. 2025, doi: 10.24912/JIKSI.V13I1.32863.

[15] Yandex, "CatBoost — Yandex Technologies." Accessed: Oct. 07, 2025. [Online]. Available: https://yandex.com/dev/catboost/

[16] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," Oct. 2018, Accessed: Oct. 07, 2025. [Online]. Available: https://arxiv.org/pdf/1810.11363

[17] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, pp. 1–45, Dec. 2020, doi: 10.1186/S40537-020-00369-8/FIGURES/9.

[18] A. Mironov and I. Khuziev, "Optimization of Oblivious Decision Tree Ensembles Evaluation for CPU," Nov. 2022, Accessed: Oct. 07, 2025. [Online]. Available: https://arxiv.org/pdf/2211.00391

[19] W. Kainz, M. A. Brovelli, D. Wang, and H. Qian, "CatBoost-Based Automatic Classification Study of River Network," *ISPRS Int. J. Geo-Information 2023, Vol. 12, Page 416*, vol. 12, no. 10, p. 416, Oct. 2023, doi: 10.3390/IJGI12100416.

[20] S. Geeitha, K. Ravishankar, J. Cho, and S. V. Easwaramoorthy, "Integrating cat boost algorithm with triangulating feature importance to predict survival outcome in recurrent cervical cancer," *Sci. Rep.*, vol. 14, no. 1, pp. 1–19, Dec. 2024, doi: 10.1038/S41598-024-67562-0;SUBJMETA.

[21] M. E. Haque *et al.*, "StackLiverNet: A Novel Stacked Ensemble Model for Accurate and Interpretable Liver Disease Detection," Jul. 2025, Accessed: Oct. 07, 2025. [Online]. Available: https://arxiv.org/pdf/2508.00117

[22] L. Benedict, "Prediksi Tingkat Kematian Covid-19 di Indonesia dengan menggunakan Metode Linear Regression," 2022.

[23] R. Ritonga, M. Ibnu Rasyid, J. Angga Putra, and M. Masrizal, *Optimalisasi Kinerja Pegawai Pertanian (Studi Kasus Penggunaan Algoritma Regresi Linear)*. Literasi Nusantara, 2024. Accessed: Oct. 07, 2025. [Online]. Available: https://penerbitlitnus.co.id/portfolio/optimalisasi-kinerja-pegawai-pertanian/

[24] Ł. Faruga, A. Filapek, M. Kraszewska, and J. Baranowski, "Dataset for Traffic Accident Analysis in Poland: Integrating Weather Data and Sociodemographic Factors," *Appl. Sci. 2025, Vol. 15, Page 7362*, vol. 15, no. 13, p. 7362, Jun. 2025, doi: 10.3390/APP15137362.