JURNAL MEDIA INFORMASI TEKNOLOGI
Volume 2, No. 1, Februari 2025
DOI: https://doi.org/10.69616/mit.v2i1.198

Vehicle CO_2 Emission Predictive Analytics Using HistGradientBoosting Regression Algorithms

Bagus Al Qohar [1]*, Ahmad Ubai Dullah[1], Putri Utami[1], Jumanto[1]
[1] Universitas Negeri Semarang, Semarang, Indonesia.
* Corresponding Author. E-mail: bagusximipa6@students.unnes.ac.id

Keywords
Environmental sustainability; HistGradientBoosting; Machine learning; Predictive analytics; Vehicle CO_2 emissions

ABSTRACT

Vehicle CO_2 emissions are a significant contributor to climate change, so research on this subject is needed. Strong prediction models and data analysis techniques are required to obtain accurate results. This research aims to analyze and predict vehicle CO_2 emissions using machine learning algorithms. Given its efficiency in handling large datasets, the HistGradientBoosting Regression algorithm was selected for predicting vehicle CO_2 emissions. The process commenced with meticulous data preparation, which involved cleaning and feature engineering. Key factors such as engine size, fuel economy, and vehicle weight were analyzed to gain insights into their impact on emissions. The study utilized a dataset comprising vehicle specifications and emissions, training and testing the HistGradientBoosting. The model's performance was evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The findings indicate that this approach effectively identifies significant factors influencing emissions while achieving impressive prediction accuracy with MAE value 12.97, MSE value 2.40, RMSE value 3.60, and R^2 value 0.996. This research offers valuable insights for policymakers and manufacturers aiming to develop low-emission vehicles and promote sustainable transportation initiatives. The paper highlights the capability of machine learning to address environmental challenges.

PENDAHULUAN

Salah satu penyebab utama polusi udara yang berdampak pada kesehatan manusia dan lingkungan serta perubahan iklim adalah emisi CO2 dari kendaraan bermotor [1], [2], [3]. Khususnya di industri transportasi darat, urgensi untuk mengurangi emisi semakin meningkat mengingat peningkatan jumlah kendaraan setiap tahunnya [4]. Salah satu dari beberapa inisiatif yang diambil untuk menurunkan emisi adalah inovasi teknologi pada kendaraan [5]. Namun, keberhasilannya bergantung pada pemantauan dan analisis berbasis data yang menyeluruh untuk mendukung kebijakan yang ramah lingkungan.

Pembelajaran mesin memiliki potensi besar menurut penelitian sebelumnya [6], [7], [8]. Namun, sebagian besar penelitian berkonsentrasi pada teknik prediksi konvensional tanpa menggunakan algoritma yang dimaksudkan untuk meningkatkan akurasi dan efisiensi pada dataset yang besar, seperti *HistGradientBoosting* [9]. Metode ini terkenal untuk menangani data tabular berbasis fitur yang kompleks, sehingga menghasilkan prediksi yang lebih konsisten daripada metode sebelumnya.

Penelitian ini bertujuan untuk menemukan faktor-faktor utama yang mempengaruhi emisi CO_2 kendaraan dengan menggunakan algoritma *HistGradientBoosting* dan meramalkan emisi. Alasan algoritma ini dipilih karena menawarkan beberapa keunggulan. Keunggulan yang dimaksud merujuk

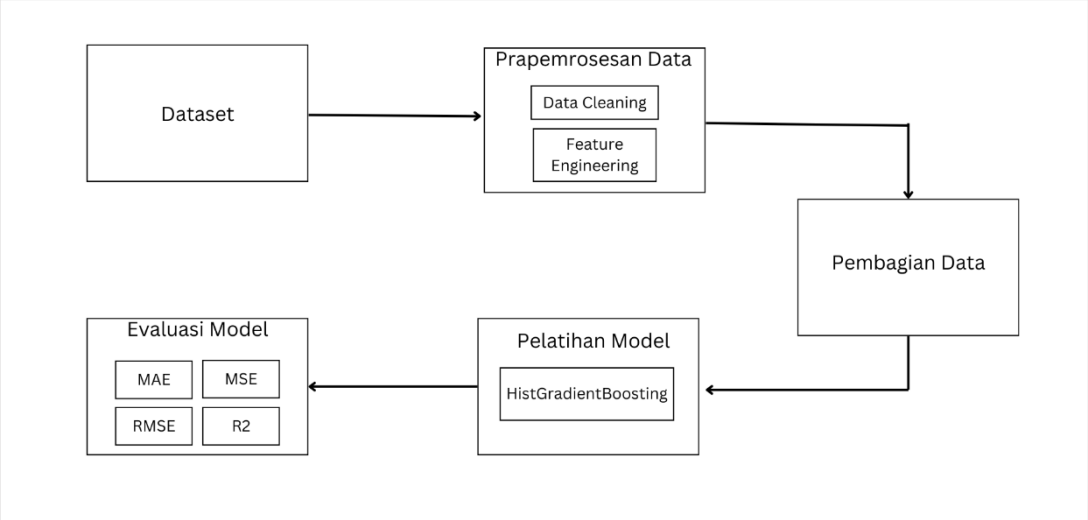
pada karakteristik teknis dan performa unik algoritma ini dibandingkan metode pembelajaran mesin lainnya, terutama dalam konteks menangani dataset besar dan kompleks. *HistGradientBoosting* dapat memproses dataset yang besar secara efisien dengan menggunakan pendekatan berbasis histogram. Dengan pendekatan berbasis histogram, algoritma ini menjadi lebih toleran terhadap outlier karena nilai ekstrim dapat dikelompokkan tanpa memengaruhi struktur model secara signifikan. Algoritma ini secara iteratif memperbaiki kesalahan dari model sebelumnya, dengan cara memberikan bobot lebih pada data yang sulit diprediksi, sehingga meningkatkan akurasi model secara keseluruhan.

Penelitian ini diharapkan dapat memberikan dasar ilmiah untuk pengambilan keputusan di bidang transportasi dan lingkungan serta metode baru yang lebih akurat dan efisien dalam memperkirakan emisi kendaraan. Penelitian ini mengisi kekosongan dalam studi prediksi emisi kendaraan dengan menggunakan algoritma *HistGradientBoosting*. Lebih jauh lagi, yang diteliti secara ekstensif dalam penelitian ini adalah elemen-elemen yang mempengaruhi emisi termasuk kapasitas mesin, konsumsi bahan bakar, dan berat kendaraan. Temuan dari penelitian ini akan sangat penting dalam membantu mendukung rencana transportasi yang berkelanjutan dan desain kendaraan yang ramah lingkungan.

Kebaruan dari penelitian ini terletak pada penggunaan algoritma *HistGradientBoosting* yang jarang diterapkan dalam prediksi emisi kendaraan. Meskipun metode ini telah digunakan dalam beberapa bidang lain, penelitian yang memanfaatkan *HistGradientBoosting* untuk memprediksi emisi kendaraan masih sangat terbatas. Dengan memanfaatkan algoritma ini, penelitian ini mengisi celah dalam literatur yang ada dan menawarkan pendekatan baru yang lebih efisien serta dapat memberikan hasil prediksi yang lebih akurat. Selain itu, penelitian ini juga menambah wawasan mengenai faktor-faktor yang berpengaruh terhadap emisi CO_2, yang sangat penting untuk pengembangan kebijakan transportasi yang lebih ramah lingkungan.

METODE

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan analitis menggunakan metode pembelajaran mesin untuk memprediksi emisi CO_2 kendaraan. Penelitian ini bertujuan untuk menganalisis hubungan antara fitur kendaraan dan emisi CO_2, serta memprediksi tingkat emisi berdasarkan data yang ada. Penelitian ini dilakukan pada bulan November 2024. Subjek penelitian ini adalah dataset kendaraan yang melibatkan lebih dari 7000 unit data kendaraan yang mencakup berbagai tipe dan model. Setiap data kendaraan memuat informasi mengenai tipe mesin, bobot kendaraan, konsumsi bahan bakar, dan emisi CO_2 yang dihasilkan. Gambar 1 menunjukkan prosedur penelitian yang telah dilakukan.



Gambar 1 Diagram Prosedur Penelitian

Prosedur penelitian dimulai dengan pengumpulan data kendaraan dari sumber yang dapat diakses publik. Data yang digunakan telah melalui tahap preprocessing, termasuk penanganan missing value dan normalisasi fitur untuk mempersiapkan data agar siap digunakan dalam model pembelajaran mesin. Setelah data dipersiapkan, model *HistGradientBoosting* diterapkan untuk menganalisis dan memprediksi emisi CO_2 berdasarkan input fitur kendaraan. Evaluasi model dilakukan menggunakan metrik seperti *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *Coefficient of Determination* (R2).

Dataset

Analisis dilakukan menggunakan dataset kendaraan yang bersumber dari basis data terbuka Kaggle [10] yang berfokus pada data jenis kendaraan dan karakteristiknya. Proses analisis dilakukan di laboratorium *Artificial Intelligence and Data Mining Center*, Universitas Negeri Semarang, Semarang, Indonesia. Target penelitian ini adalah kendaraan yang terdaftar dengan berbagai karakteristik, seperti kapasitas mesin, berat kendaraan, konsumsi bahan bakar, dan tipe bahan bakar yang digunakan. Sasaran penelitian adalah untuk memprediksi emisi CO_2 berdasarkan fitur-fitur kendaraan tersebut dan mengidentifikasi faktor-faktor yang paling berpengaruh terhadap emisi.

Prapemrosesan Data

Pada tahap pengolahan data, kualitas data yang digunakan sangat penting untuk memastikan hasil analisis yang valid dan akurat. Salah satu tantangan utama dalam pengolahan data adalah menangani masalah kualitas, seperti nilai yang hilang (*missing values*) dan duplikasi (*duplicates*). Kedua masalah ini dapat mengurangi kualitas model prediksi dan memengaruhi hasil analisis. Nilai yang hilang seringkali terjadi karena kesalahan dalam pengumpulan data atau data yang tidak tersedia untuk beberapa entri.

Beberapa teknik umum dapat digunakan untuk menangani nilai yang hilang [11]. Jika jumlah data yang hilang sangat kecil, baris yang mengandung nilai hilang dapat dihapus tanpa mempengaruhi analisis secara signifikan. Untuk data numerik, nilai yang hilang dapat diisi dengan rata-rata atau median dari kolom tersebut, sedangkan untuk data kategorikal, imputasi dapat dilakukan dengan modus (nilai yang paling sering muncul). Dalam beberapa kasus, teknik seperti regresi atau algoritma pembelajaran mesin lainnya digunakan untuk memperkirakan nilai yang hilang berdasarkan hubungan antara fitur lainnya dalam dataset. Penanganan yang tepat untuk nilai yang hilang sangat penting karena jika tidak ditangani dengan benar, dapat menyebabkan bias dalam model dan merendahkan akurasi prediksi. Beberapa studi yang relevan mengindikasikan bahwa imputasi yang tepat dapat meningkatkan kualitas hasil prediksi secara signifikan.

Duplikasi data terjadi ketika entri data yang sama tercatat lebih dari satu kali dalam dataset. Hal ini dapat mengakibatkan bias dalam model, karena duplikasi dapat memberikan bobot yang tidak proporsional pada data tertentu. Penanganan duplikasi biasanya melibatkan menemukan baris yang memiliki nilai identik pada semua kolom atau kolom kunci tertentu dan menghapusnya untuk memastikan hanya ada satu entri per kendaraan [12]. Dalam beberapa kasus, lebih baik untuk menggabungkan informasi dari entri duplikat. Penanganan duplikasi juga penting untuk menjaga integritas dataset dan mencegah hasil analisis yang bias. Teknik penghapusan duplikasi dapat diterapkan secara langsung menggunakan fungsi-fungsi di pustaka seperti Pandas pada Python, di mana *drop_duplicates()* digunakan untuk menghapus duplikasi pada data.

Pembagian Data

Dalam proses pembelajaran mesin, pembagian data merupakan langkah yang sangat penting untuk melatih dan menguji model. Berdasarkan penelitian analisis prediksi emisi CO_2 kendaraan menggunakan *HistGradientBoosting*. Langkah pertama dalam pembagian data yaitu memisahkan data menjadi fitur dan target. Fitur adalah variabel-variabel independen yang akan digunakan untuk memprediksi target, sedangkan target adalah variabel yang ingin diprediksi. Sebagai contoh dapat dilihat pada Tabel 1, fitur dapat mencakup atribut seperti kapasitas mesin, bobot kendaraan, dan konsumsi bahan bakar, sementara target adalah emisi CO_2 kendaraan.

Tabel 1. Pemisahan Data

Fitur	*Make, Vehicle Class, Engine Size(L), Cylinders, Transmission, Fuel Type, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg)*
Target	*CO2 Emissions (g/km)*

Setelah fitur dan target dipisahkan, langkah selanjutnya yaitu pembagian data menjadi dua bagian, set latih (*training set*) dan set uji (*test set*). Pembagian ini bertujuan untuk melatih model menggunakan set latih, dan menguji model dengan set uji yang belum pernah dilihat sebelumnya untuk mengevaluasi kinerjanya. Biasanya, pembagian data dilakukan dengan perbandingan tertentu, misalnya pada penelitian ini 80% data untuk set latih dan 20% untuk set uji. Pembagian data dapat dilakukan menggunakan fungsi *train_test_split* dari library *sklearn.model_selection*.

Pelatihan Model

Setelah data dibagi menjadi set latih (*training set*) dan set uji (*test set*), langkah selanjutnya adalah melatih model dengan menggunakan *HistGradientBoosting* pada set latih. Pada tahap ini, model pembelajaran mesin berusaha untuk belajar dari data latih, yaitu mengenali pola-pola yang ada antara fitur (*X_train*) dan target (*y_train*). *HistGradientBoosting* adalah model regresi berbasis *Gradient Boosting*, yang merupakan teknik *ensemble learning* [13]. *Gradient Boosting* bekerja dengan menggabungkan sejumlah model pembelajaran mesin yang lebih sederhana (misalnya, pohon keputusan) untuk membentuk model yang lebih kuat [14]. *HistGradientBoosting* adalah versi dari *Gradient Boosting* yang mengoptimalkan proses pelatihan dengan menggunakan *histogram-based techniques*, yang lebih efisien dalam memproses data besar dan mempercepat proses pelatihan dibandingkan dengan metode *gradient boosting* klasik [15]. Meskipun teknik *ensemble* seperti *XGBoost* dan *stacking* telah menunjukkan performa yang tinggi di berbagai aplikasi [16], [17], *HistGradientBoosting* menawarkan keunggulan yang lebih spesifik, termasuk kemampuannya dalam memproses set data yang besar secara efisien dengan menggunakan pendekatan berbasis histogram. Model ini secara iteratif memperbaiki kesalahan dari model sebelumnya, dengan cara memberikan bobot lebih pada data yang sulit diprediksi, sehingga meningkatkan akurasi model secara keseluruhan.

Selama pelatihan, *HistGradientBoosting* mencoba mempelajari hubungan antara fitur (*variabel input*) dan target (*variabel output*). Proses ini terdiri dari beberapa langkah kunci. Pada awalnya, model membuat prediksi yang sangat sederhana. Prediksi ini biasanya berupa rata-rata target pada data latih. Model kemudian akan mencoba meningkatkan prediksi ini dengan cara mengurangi *residual error* (kesalahan prediksi) pada setiap iterasi. Model bekerja dengan cara mengiterasi melalui beberapa tahap pembelajaran. Pada iterasi pertama, model mengukur kesalahan atau *residual* antara prediksi awal dan nilai target yang sebenarnya. Kemudian, model membangun pohon keputusan untuk memprediksi *residual* tersebut, memberikan bobot lebih besar pada data yang sulit diprediksi. Pada iterasi berikutnya, model memperbaiki kesalahan prediksi sebelumnya dengan menambahkan prediksi pohon keputusan baru ke dalam model secara bertahap. Proses ini terus berlanjut hingga jumlah iterasi yang telah ditentukan atau hingga model tidak dapat meningkatkan akurasi lebih lanjut.

Di setiap langkah iterasi, *HistGradientBoosting* mengurangi kesalahan dengan mempelajari kesalahan *residual* sebelumnya. Setiap pohon keputusan yang dibangun bertujuan untuk meminimalkan kesalahan prediksi pada langkah sebelumnya, meningkatkan performa model secara keseluruhan. Pada saat pelatihan, beberapa *hyperparameter* dari model *HistGradientBoosting* dapat diatur untuk mengoptimalkan kinerja model. Beberapa *hyperparameter* yang penting diantarnaya *n_estimators* atau jumlah iterasi atau jumlah pohon keputusan yang akan dibangun. *learning_rate* untuk mengatur kecepatan di mana model memperbarui parameter berdasarkan kesalahan yang diperoleh, dan *max_depth* atau kedalaman maksimal dari setiap pohon keputusan. Menetapkan nilai ini dapat membantu mencegah model menjadi terlalu kompleks dan *overfitting.* Dengan pengaturan

hyperparameter yang tepat, akan menghasilkan model yang mampu melakukan prediksi dengan akurasi yang tinggi.

Evaluasi Model

Setelah proses pelatihan selesai, model akan diuji pada set uji (*test set*) untuk mengevaluasi seberapa baik model dapat melakukan prediksi pada data yang tidak terlihat sebelumnya. Hasil prediksi dari model dibandingkan dengan nilai target yang sesungguhnya (*y_test*), dan metrik evaluasi seperti *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), atau *R-squared* (R^2) digunakan untuk mengukur kinerja model.

MAE mengukur besaran rata-rata kesalahan dalam satu set prediksi, tanpa mempertimbangkan arahnya [18], [19]. MAE dihitung sebagai rata-rata perbedaan absolut antara nilai prediksi dan nilai aktual seperti pada Rumus 1. MAE yang lebih rendah menunjukkan kinerja model yang lebih baik, karena ini menandakan bahwa prediksi lebih dekat dengan nilai aktual. Metrik ini sangat berguna karena mudah ditafsirkan dan memberikan ukuran langsung dari akurasi prediksi.

MSE menghitung rata-rata kuadrat kesalahan, yaitu rata-rata perbedaan kuadrat antara nilai prediksi dan nilai aktual ditunjukkan pada Rumus 2. Metrik RMSE juga memiliki definisi yang sama namun memberikan bobot lebih pada kesalahan yang lebih besar karena proses kuadrat ditunjukkan pada Rumus 3. MSE sensitif terhadap outlier, yang berarti bahwa beberapa kesalahan besar dapat secara signifikan meningkatkan nilai MSE [20], [21]. Oleh karena itu, meskipun memberikan ukuran akurasi prediksi secara keseluruhan, namun mungkin tidak selalu mencerminkan kinerja model secara seimbang.

R^2 adalah ukuran statistik yang mewakili proporsi varians untuk variabel dependen yang dijelaskan oleh variabel independen dalam model regresi ditunjukkan pada Rumus 4. Nilai R^2 sebesar 1 mengindikasikan bahwa model menjelaskan semua variabilitas data respons di sekitar rata-ratanya, sedangkan nilai 0 mengindikasikan bahwa model tidak menjelaskan variabilitas apapun [22], [23]. Nilai R^2 yang lebih tinggi umumnya menunjukkan kecocokan yang lebih baik antara model dengan data.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (1)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{RSS}{TSS} \quad (4)$$

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini berisi informasi terkait karakteristik kendaraan dan emisi CO_2 yang dihasilkan. Dataset terdiri dari beberapa fitur seperti kapasitas mesin, konsumsi bahan bakar, dan tipe bahan bakar, dengan target variabel berupa emisi CO_2 (dalam gram per kilometer). Dataset ini mencakup total sekitar 7000 data yang telah dikumpulkan dari berbagai jenis kendaraan. Dengan menggunakan teknik pembelajaran mesin yang canggih, khususnya *HistGradientBoosting*, penelitian ini menunjukkan peningkatan yang signifikan dalam akurasi

prediksi dibandingkan dengan metode konvensional. Temuan ini memiliki implikasi praktis untuk perencanaan transportasi yang berkelanjutan dan desain kendaraan yang ramah lingkungan. Dengan menawarkan kerangka kerja prediksi yang lebih tepat, penelitian ini mendukung upaya untuk mengurangi jejak karbon di sektor otomotif, membuka jalan bagi strategi berbasis data dalam mencapai tujuan lingkungan dan keberlanjutan.

Secara statistik deskriptif, rata-rata emisi rata-rata emisi CO_2 adalah sekitar 251.16 g/km dengan standar deviasi sebesar 59.29 g/km menunjukkan adanya variasi yang cukup besar antar kendaraan. Selain itu, distribusi nilai target memperlihatkan pola yang mendekati normal, dengan beberapa outlier yang merepresentasikan kendaraan dengan emisi sangat tinggi. Tabel 2 menunjukkan struktur data sebelum dan sesudah pemrosesan. Tabel ini sangat penting karena mengilustrasikan perubahan yang dilakukan pada dataset, yang sangat penting untuk memahami langkah-langkah analisis dan pemodelan selanjutnya.

Tabel 2. Tabel Struktur Data Sebelum dan Sesudah Prapemrosesan

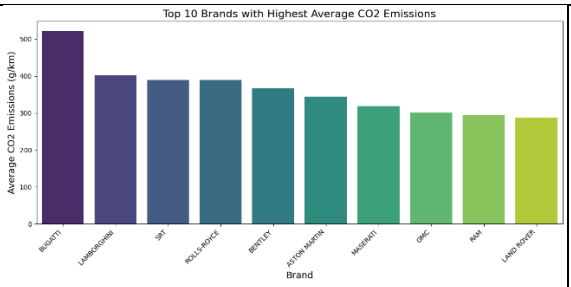
Sebelum			Sesudah		
Column	*Non-null Count*	*Dtype*	*Column*	*Non-null Count*	*Dtype*
Make	*7385*	*object*	*Make*	*6282*	*int64*
Model	*7385*	*object*	*Vehicle Class*	*6282*	*int64*
Vehicle Class	*7385*	*object*	*Engine Size (L)*	*6282*	*float64*
Engine Size (L)	*7385*	*float64*	*Cylinders*	*6282*	*int64*
Cylinders	*7385*	*int64*	*Transmission*	*6282*	*int64*
Transmission	*7385*	*object*	*Fuel Type*	*6282*	*int64*
Fuel Type	*7385*	*object*	*Fuel Consumption City (L/100 km)*	*6282*	*float64*
Fuel Consumption City (L/100 km)	*7385*	*float64*	*Fuel Consumption Hwy (L/100 km)*	*6282*	*float64*
Fuel Consumption Hwy (L/100 km)	*7385*	*float64*	*Fuel Consumption Comb (L/100 km)*	*6282*	*float64*
Fuel Consumption Comb (L/100 km)	*7385*	*float64*	*Fuel Consumption Comb (mpg)*	*6282*	*int64*
Fuel Consumption Comb (mpg)	*7385*	*int64*	*CO2 Emissions (g/km)*	*6282*	*int64*
CO2 Emissions (g/km)	7385	int64			

Tabel 2 menguraikan struktur awal dataset, termasuk berbagai fitur dan jenisnya, serta variabel target, yaitu emisi CO_2. Struktur awal ini penting untuk mengidentifikasi atribut yang akan digunakan dalam proses pemodelan prediktif. Tabel ini menyoroti langkah-langkah prapemrosesan yang diambil untuk mempersiapkan data untuk analisis. Hal ini termasuk menangani nilai yang hilang, menormalkan fitur, dan mungkin mengkodekan variabel kategorikal.

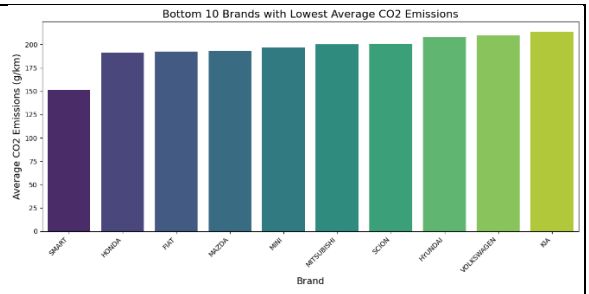
Langkah-langkah ini sangat penting karena memastikan bahwa data bersih dan sesuai untuk algoritma pembelajaran mesin yang digunakan dalam penelitian. Dengan membandingkan struktur data sebelum dan sesudah prapemrosesan memungkinkan pembaca untuk melihat dampak dari langkah-langkah ini pada set data. Sebagai contohnya yaitu menunjukkan berapa banyak nilai yang hilang yang telah ditangani atau bagaimana distribusi fitur tertentu berubah setelah normalisasi. Perbandingan ini sangat penting untuk memvalidasi keefektifan metode prapemrosesan yang diterapkan.

Gambar 2 dan Gambar 3 memberikan representasi visual yang melengkapi analisis data dan temuan terkait emisi CO_2 kendaraan. Representasi visual ini membantu mengidentifikasi pola dengan cepat, seperti jenis kendaraan mana yang cenderung memiliki emisi yang lebih tinggi atau emisi lebih rendah dan adanya nilai ekstrem yang dapat mempengaruhi analisis. Hal ini sangat penting untuk memahami variabilitas emisi dan dampak potensial dari karakteristik kendaraan tertentu terhadap

keluaran CO. Gambar-gambar ini sangat penting untuk memahami hasil dan hubungan antara berbagai faktor yang mempengaruhi emisi.

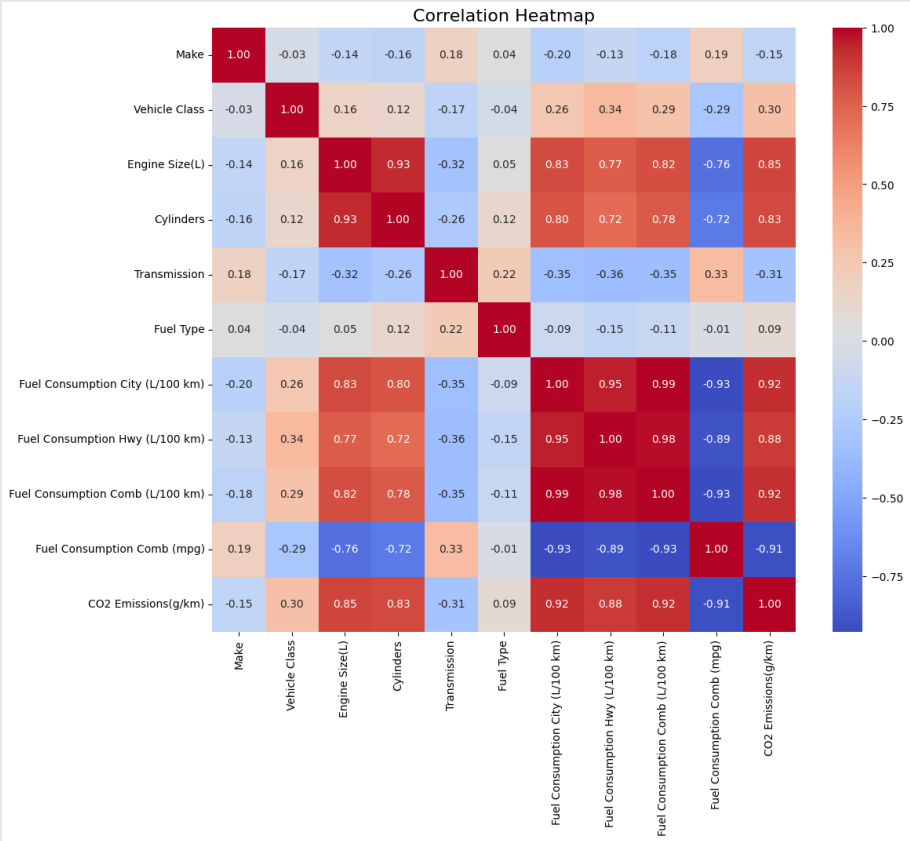


Gambar 2. *Brands* Dengan Rata-Rata Emisi CO2 Tertinggi



Gambar 3. *Brands* Dengan Rata-Rata Emisi CO2 Terendah

Visualisasi data memainkan peran krusial dalam analisis emisi CO_2 kendaraan. Dengan menggunakan representasi visual, peneliti dapat dengan cepat mengidentifikasi tren dan anomali dalam data, seperti jenis kendaraan yang memiliki emisi lebih tinggi atau lebih rendah. Hal ini sangat penting untuk mengarahkan perhatian pada faktor-faktor yang mungkin mempengaruhi emisi, serta untuk memvalidasi hasil analisis yang telah dilakukan sebelumnya.



Gambar 4. Grafik Matriks Korelasi

Gambar 4 menunjukkan korelasi antara berbagai fitur kendaraan dan emisi CO_2. Memahami korelasi ini sangat penting untuk mengidentifikasi faktor mana yang paling berdampak signifikan

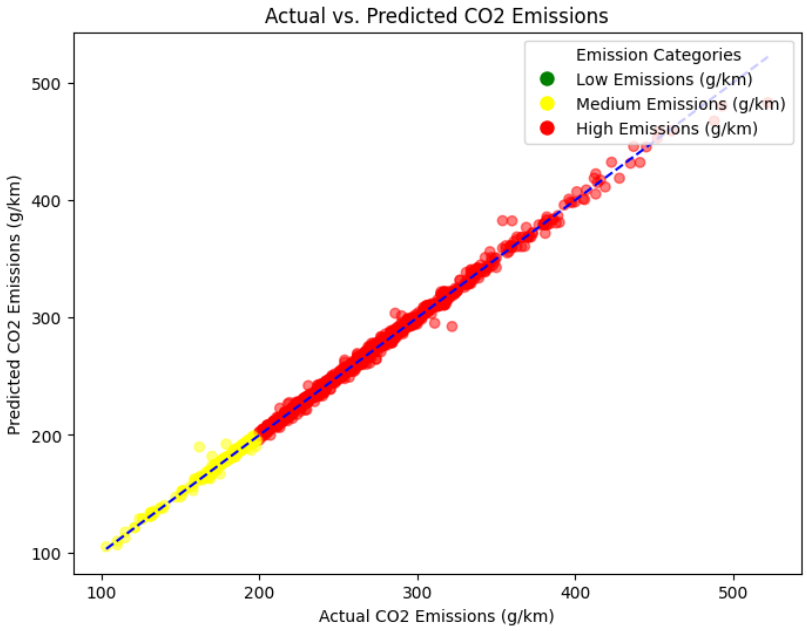
terhadap emisi. Grafik menampilkan koefisien korelasi, yang mengukur kekuatan dan arah hubungan antara fitur-fitur (seperti kapasitas mesin, berat, dan konsumsi bahan bakar) dan emisi CO_2. Korelasi positif menunjukkan bahwa ketika satu variabel meningkat, variabel lainnya juga meningkat, sementara korelasi negatif menunjukkan hubungan yang terbalik. Sebagai contoh, fitur-fitur seperti kapasitas mesin dan konsumsi bahan bakar diharapkan menunjukkan korelasi positif, yang berarti bahwa nilai yang lebih tinggi pada fitur-fitur ini menyebabkan peningkatan emisi. Hal ini sejalan dengan temuan dalam studi yang menekankan pentingnya faktor-faktor ini dalam memprediksi emisi. Hasil evaluasi prediksi emisi CO2 ditunjukkan pada Tabel 3.

Tabel 3. Hasil Evaluasi Prediksi

Metrik	Nilai
Mean Absolute Error (MAE)	12.972522
Mean Squared Error (MSE)	2.402790
Root Mean Squared Error (RMSE)	3.601739
R-squared (R^2)	0.996405

Metrik evaluasi untuk model prediktif menunjukkan hasil yang menjanjikan dalam hal akurasi dan keandalan. *Mean Squared Error* (MSE) sekitar 12,97 menunjukkan bahwa rata-rata perbedaan kuadrat antara emisi CO_2 yang diprediksi dan yang sebenarnya relatif rendah, yang menunjukkan bahwa prediksi model mendekati nilai aktual. *Mean Absolute Error* (MAE) sekitar 2,40 semakin memperkuat hal ini, karena mencerminkan perbedaan absolut rata-rata antara prediksi dan nilai aktual, yang mengindikasikan bahwa secara rata-rata, prediksi model meleset sekitar 2,40 unit emisi CO_2. *Root Mean Squared Error* (RMSE) sekitar 3,60, sebagai akar kuadrat dari MSE, memberikan ukuran kesalahan dalam unit yang sama dengan variabel target, yang juga cukup rendah, yang menunjukkan kinerja prediksi yang baik. Terakhir, nilai *R-squared* (R^2) sebesar 0,996 menunjukkan bahwa model tersebut menjelaskan sekitar 99,64% varians dalam data emisi CO_2, yang mengindikasikan kecocokan yang sangat baik. Secara keseluruhan, metrik-metrik ini menunjukkan bahwa model ini sangat efektif dalam memprediksi emisi CO_2 kendaraan, dengan kesalahan yang minimal dan kekuatan penjelasan yang kuat.

.



Gambar 5. Grafik Aktual vs Prediksi

Gambar 5 menunjukkan perbandingan nilai emisi CO_2 aktual kendaraan dengan nilai yang diprediksi oleh model *HistGradientBoosting*. Grafik ini sangat penting untuk mengevaluasi kinerja model prediksi. Grafik memiliki sumbu X yang mewakili nilai aktual emisi CO_2 (dalam gram per kilometer) dan sumbu Y yang menunjukkan nilai yang diprediksi oleh model. Ini memungkinkan visualisasi langsung dari seberapa baik model dapat memprediksi emisi berdasarkan data yang ada. Setiap titik pada grafik mewakili satu kendaraan dalam dataset. Titik-titik ini menunjukkan hubungan antara nilai aktual dan nilai prediksi. Jika model bekerja dengan baik, titik-titik ini akan terdistribusi dekat dengan garis diagonal yang menunjukkan kesetaraan antara nilai aktual dan prediksi.

Garis diagonal (y = x) dalam grafik berfungsi sebagai referensi. Titik-titik yang berada di sepanjang garis ini menunjukkan bahwa model memprediksi emisi dengan akurat. Sebaliknya, titik-titik yang jauh dari garis ini menunjukkan kesalahan prediksi yang signifikan. Jika grafik menunjukkan bahwa model memiliki banyak titik yang dekat dengan garis diagonal, ini menunjukkan bahwa model tersebut efektif dan dapat diandalkan untuk digunakan dalam analisis lebih lanjut dan perencanaan kebijakan. Gambar ini juga berfungsi sebagai alat visual yang membantu pembaca memahami kinerja model secara intuitif. Dengan melihat sebaran titik data, pembaca dapat dengan cepat menilai seberapa baik model dalam menangkap variabilitas emisi CO_2 berdasarkan fitur kendaraan yang dianalisis. Secara keseluruhan, Gambar 5 memberikan gambaran yang jelas tentang akurasi model prediksi emisi CO_2, yang merupakan aspek penting dalam penelitian ini untuk mendukung upaya pengurangan jejak karbon di sektor otomotif.

SIMPULAN

Penelitian ini berhasil menjawab tujuan untuk memprediksi emisi CO_2 kendaraan menggunakan algoritma *HistGradientBoosting*, yang menunjukkan keefektifannya dalam menangani data tabel berbasis fitur yang kompleks. Temuan menunjukkan bahwa faktor-faktor seperti kapasitas mesin, berat kendaraan, dan konsumsi bahan bakar secara signifikan memengaruhi emisi CO_2, ditunjukkan pada Gambar 4. Hal ini memberikan pemahaman yang lebih jelas tentang hubungan antara variabel-variabel dalam dataset. Selain itu model yang digunakan menunjukkan performa tinggi dengan hasil metrik evaluasi yaitu Mean Absolute Error (MAE) sebesar 12.97, Mean Squared Error (MSE) sebesar 2.40, Root Mean Squared Error (RMSE) sebesar 3.60, dan R-squared (R^2) sebesar 0.996 seperti yang ditunjukan pada Tabel 3. Hasil ini membuktikan bahwa model mampu memberikan prediksi yang sangat akurat. Para penulis menyimpulkan bahwa algoritma *HistGradientBoosting* menawarkan pendekatan yang lebih akurat dan efisien dibandingkan dengan metode prediksi tradisional, mengisi kesenjangan dalam literatur yang ada tentang prediksi emisi kendaraan. Berdasarkan hasil ini, disarankan agar para pembuat kebijakan dan produsen otomotif memanfaatkan wawasan ini untuk mengembangkan strategi yang bertujuan untuk mengurangi emisi, seperti mengoptimalkan desain kendaraan dan meningkatkan efisiensi bahan bakar. Selain itu, penelitian ini menganjurkan penerapan teknik pembelajaran mesin canggih yang berkelanjutan dalam penelitian lingkungan untuk mendorong solusi transportasi berkelanjutan dan menginformasikan kerangka kerja peraturan yang mempromosikan praktik ramah lingkungan di industri otomotif.

DAFTAR PUSTAKA

[1] M. Filonchyk, M. P. Peterson, L. Zhang, V. Hurynovich, and Y. He, “Greenhouse gases emissions and global climate change: Examining the influence of CO2, CH4, and N2O,” *Sci. Total Environ.*, vol. 935, p. 173359, Jul. 2024,

https://doi.org/10.1016/j.scitotenv.2024.173359

[2] C. Yin, J. Wu, X. Sun, Z. Meng, and C. Lee, "Road transportation emission prediction and policy formulation: Machine learning model analysis," *Transp. Res. Part D Transp. Environ.*, vol. 135, p. 104390, Oct. 2024, https://doi.org/10.1016/j.trd.2024.104390

[3] B. Zhu, S. Hu, X. (Michael) Chen, C. Roncoli, and D.-H. Lee, "Uncovering driving factors and spatiotemporal patterns of urban passenger car CO2 emissions: A case study in Hangzhou, China," *Appl. Energy*, vol. 375, p. 124094, Dec. 2024, https://doi.org/10.1016/j.apenergy.2024.124094

[4] W. Eko Cahyono, Parikesit, B. Joy, W. Setyawati, and R. Mahdi, "Projection of CO2 emissions in Indonesia," *Mater. Today Proc.*, vol. 63, pp. S438–S444, 2022, https://doi.org/10.1016/j.matpr.2022.04.091

[5] T. Thurner, K. Fursov, and A. Nefedova, "Early adopters of new transportation technologies: Attitudes of Russia's population towards car sharing, the electric car and autonomous driving," *Transp. Res. Part A Policy Pract.*, vol. 155, pp. 403–417, Jan. 2022, https://doi.org/10.1016/j.tra.2021.11.006

[6] S. Cesar de Lima Nogueira, S. H. Och, L. M. Moura, E. Domingues, L. dos S. Coelho, and V. C. Mariani, "Prediction of the NOx and CO2 emissions from an experimental dual fuel engine using optimized random forest combined with feature engineering," *Energy*, vol. 280, p. 128066, Oct. 2023, https://doi.org/10.1016/j.energy.2023.128066

[7] H. Fu, D. Yang, S. Wang, L. Wang, and D. Wang, "A novel online energy management strategy for fuel cell vehicles based on improved random forest regression in multi road modes," *Energy Convers. Manag.*, vol. 305, p. 118261, Apr. 2024, https://doi.org/10.1016/j.enconman.2024.118261

[8] G. Zhou, L. Mao, T. Bao, and F. Zhuang, "Machine learning-driven CO2 emission forecasting for light-duty vehicles in China," *Transp. Res. Part D Transp. Environ.*, vol. 137, p. 104502, Dec. 2024, https://doi.org/10.1016/j.trd.2024.104502

[9] T. A. Munshi, L. N. Jahan, M. F. Howladar, and M. Hashan, "Prediction of gross calorific value from coal analysis using decision tree-based bagging and boosting techniques," *Heliyon*, vol. 10, no. 1, p. e23395, Jan. 2024, https://doi.org/10.1016/j.heliyon.2023.e23395

[10] B. Sahan, "Vehicle CO2 Emissions Dataset." Accessed: Nov. 25, 2024. [Online]. Available: https://www.kaggle.com/datasets/brsahan/vehicle-co2-emissions-dataset/

[11] N. B. Shaik, K. Jongkittinarukorn, and K. Bingi, "XGBoost based enhanced predictive model for handling missing input parameters: A case study on gas turbine," *Case Stud. Chem. Environ. Eng.*, vol. 10, p. 100775, Dec. 2024, https://doi.org/10.1016/j.cscee.2024.100775

[12] H. Mesghali, B. Akhlaghi, N. Gozalpour, J. Mohammadpour, F. Salehi, and R. Abbassi, "Predicting maximum pitting corrosion depth in buried transmission pipelines: Insights from tree-based machine learning and identification of influential factors," *Process Saf. Environ. Prot.*, vol. 187, pp. 1269–1285, Jul. 2024, https://doi.org/10.1016/j.psep.2024.05.014

[13] R. N. Nashikkar, Y. N. Padhye, and R. Ingle, "Enhancing Malicious Domain Detection Using Advanced Machine Learning Techniques," in *2023 IEEE Pune Section International Conference (PuneCon)*, IEEE, Dec. 2023, pp. 1–6. https://doi.org/10.1109/PuneCon58714.2023.10450038

[14] E. A. Elhadjamor and H. Harbaoui, "A Comparison Analysis of Heart Disease Prediction Using Supervised Machine Learning Techniques," in *2024 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, Jun. 2024, pp. 1–6. https://doi.org/10.1109/ISCC61673.2024.10733656

[15] N. S. K. M. K. Tirumanadham, T. S, and S. M, "Evaluating Boosting Algorithms for

Academic Performance Prediction in E-Learning Environments," in *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, IEEE, Jan. 2024, pp. 1–8. https://doi.org/10.1109/IITCEE59897.2024.10467968

[16] D. A. A. Pertiwi, K. Ahmad, S. N. Salahudin, A. M. Annegrat, and M. A. Muslim, "Using Genetic Algorithm Feature Selection to Optimize XGBoost Performance in Australian Credit," *J. Soft Comput. Explor.*, vol. 5, no. 1, pp. 92–98, Apr. 2024, https://doi.org/10.52465/joscex.v5i1.302

[17] R. Rofik, R. Aulia, K. Musaadah, S. S. F. Ardyani, and A. A. Hakim, "Optimization of Credit Scoring Model Using Stacking Ensemble Learning and Oversampling Techniques," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, Dec. 2023, https://doi.org/10.52465/joiser.v2i1.203

[18] A. S. Mahajan and A. Shrivastav, "Short Term Load Forecasting based on Regression models," in *2023 International Conference for Advancement in Technology (ICONAT)*, IEEE, Jan. 2023, pp. 1–8. https://doi.org/10.1109/ICONAT57137.2023.10080359

[19] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression," *IEEE Signal Process. Lett.*, vol. 27, pp. 1485–1489, 2020, https://doi.org/10.1109/LSP.2020.3016837

[20] K. Rajesh and M. S. Saravanan, "Prediction of Customer Spending Score for the Shopping Mall using Gaussian Mixture Model comparing with Linear Spline Regression Algorithm to reduce Root Mean Square Error," in *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, May 2022, pp. 335–341. https://doi.org/10.1109/ICICCS53718.2022.9788162

[21] D. S. K. Karunasingha, "Root mean square error or mean absolute error? Use their ratio as well," *Inf. Sci. (Ny).*, vol. 585, pp. 609–629, Mar. 2022, https://doi.org/10.1016/j.ins.2021.11.036

[22] J. Rewilak, "The (non) determinants of Olympic success," *J. Sports Econom.*, vol. 22, no. 5, pp. 546–570, Jun. 2021, https://doi.org/10.1177/1527002521992833

[23] G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, "Prediction and Forecasting of Air Quality Index in Chennai using Regression and ARIMA time series models," *J. Eng. Res.*, vol. 9, Sep. 2021, https://doi.org/10.36909/jer.10253