

Evaluasi Model Machine Learning untuk Prediksi Harga Mobil dengan Perbandingan Ensemble dan Regresi Linear

Evaluation of Machine Learning Models for Car Price Prediction with Ensemble Methods and Linear Regression

Nur Oktavin Idris^{*1}, Fuad Pontooyo²

¹Teknologi Rekayasa Perangkat Lunak, Universitas Negeri Gorontalo

²Teknologi Rekayasa Energi Terbarukan, Universitas Negeri Gorontalo

E-mail: ¹nur.oktavin@ung.ac.id, ²fuad.pontooyo@ung.ac.id

Abstrak

Prediksi harga mobil merupakan tantangan utama dalam industri otomotif karena dipengaruhi oleh berbagai faktor, seperti spesifikasi teknis, jenis bahan bakar, dan sistem transmisi. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja model regresi linear serta metode ensemble learning, yaitu Random Forest dan Gradient Boosting, dalam memprediksi harga mobil. Dataset yang digunakan berasal dari Kaggle dengan 11.914 baris data dan 16 fitur. Proses penelitian meliputi tahap data understanding, data preparation, pemodelan, dan evaluasi menggunakan metrik Mean Squared Error (MSE) dan R-squared (R^2). Hasil penelitian menunjukkan bahwa model Gradient Boosting memiliki performa terbaik dengan nilai R^2 sebesar 0.963868 dan MSE terendah dibandingkan model lainnya, diikuti oleh Random Forest dengan R^2 sebesar 0.899657. Sebaliknya, regresi linear menunjukkan performa yang lebih rendah dengan R^2 sebesar 0.417905, menunjukkan keterbatasannya dalam menangani hubungan non-linear dalam data. Hasil prediksi dengan model terbaik menunjukkan estimasi harga yang cukup mendekati harga aktual, meskipun masih beberapa perlu perbaikan melalui optimasi hyperparameter. Penelitian ini menegaskan bahwa metode ensemble learning, khususnya Gradient Boosting, merupakan pendekatan yang lebih efektif dalam memprediksi harga mobil dibandingkan regresi linear. Model ini berpotensi diterapkan dalam industri otomotif untuk meningkatkan akurasi estimasi harga kendaraan bagi produsen, dealer, maupun konsumen.

Kata kunci: prediksi, machine learning, gradient boosting, random forest, regresi linear

Abstract

Car price prediction is a major challenge in the automotive industry because it is influenced by various factors, such as technical specifications, fuel type, and transmission system. This research aims to evaluate and compare the performance of linear regression models and ensemble learning methods, namely Random Forest and Gradient Boosting, in predicting car prices. The dataset used comes from Kaggle, with 11,914 rows of data and 16 features. The research process includes the stages of data understanding, data preparation, modeling, and evaluation using the Mean Squared Error (MSE) and R-squared (R^2) metrics. The research results show that the Gradient Boosting model has the best performance, with an R^2 value of 0.963868 and the lowest MSE compared to other models, followed by Random Forest with an R^2 of 0.899657. In contrast, linear regression showed lower performance, with an R^2 of 0.417905, indicating its limitations in handling non-linear relationships in the data. The best model's predictions show estimates close to actual prices, although some improvements are still needed through hyperparameter optimization. This research confirms that ensemble learning, especially Gradient Boosting, is a more effective approach in predicting car prices compared to linear regression. This model has the potential to be applied in the automotive industry to increase the accuracy of vehicle price estimates for producers, dealers, and consumers.

0.417905, indicating its limitations in handling non-linear relationships in the data. The prediction results from the best model show price estimates that are quite close to actual prices, although some improvements still need to be made through hyperparameter optimization. This research confirms that ensemble learning methods, especially Gradient Boosting, provide a more effective approach to predicting car prices than linear regression. This model has the potential to be applied in the automotive industry to improve the accuracy of vehicle price estimates for manufacturers, dealers, and consumers.

Keywords: prediction, machine learning, gradient boosting, random forest, linear regression

1. PENDAHULUAN

Prediksi harga mobil merupakan salah satu tantangan utama dalam industri otomotif karena dipengaruhi oleh berbagai faktor seperti spesifikasi teknis, merek, jenis bahan bakar, sistem transmisi, dan faktor lainnya. Kemampuan untuk memperkirakan harga jual kendaraan dengan akurat sangat penting bagi berbagai pemangku kepentingan, termasuk produsen, dealer, serta konsumen. Salah satu metode yang sering digunakan dalam prediksi harga mobil adalah regresi linear, yang mampu menangkap hubungan antara variabel independen dan harga jual mobil sebagai variabel dependen. Namun, metode ini memiliki keterbatasan dalam menangkap hubungan non-linear yang sering muncul dalam data kendaraan[1].

Seiring dengan berkembangnya teknik pembelajaran mesin (*machine learning*), metode *ensemble learning* seperti Random Forest dan Gradient Boosting semakin banyak digunakan untuk meningkatkan akurasi prediksi harga kendaraan. *Ensemble learning* bekerja dengan menggabungkan beberapa model untuk menghasilkan prediksi yang lebih stabil dan akurat dibandingkan pendekatan tunggal[2]. Studi oleh Kinadi et al. (2022) menunjukkan bahwa Random Forest Regressor memiliki tingkat akurasi yang lebih tinggi dibandingkan regresi linear dalam memprediksi harga kendaraan[3]. Sementara itu, Alghifari et al. (2022) menemukan bahwa Extreme Gradient Boosting (XGBoost) dengan Bayesian Optimization menghasilkan nilai R^2 sebesar 0.9827, menjadikannya model dengan performa terbaik dalam memprediksi harga mobil[2].

Beberapa penelitian sebelumnya telah mengeksplorasi berbagai metode dalam prediksi harga mobil. Ahmad et al. (2024) menggunakan Lasso Regression dan Ridge Regression dan memperoleh akurasi mendekati 90% dalam prediksi harga mobil[4]. Sementara itu, Chuyang Jin (2021) membandingkan berbagai metode regresi seperti linear regression, polynomial regression, decision tree, dan random forest, dengan hasil terbaik diperoleh pada model random forest dengan R^2 sebesar 0.90416[5]. Hankar et al. (2022) juga menemukan bahwa Gradient Boosting Regressor (GBR) memiliki performa yang lebih baik dibandingkan metode lain dalam memprediksi harga kendaraan[6]. Di sisi lain, penelitian oleh Yashi et al. (2022) menemukan bahwa metode LightGBM memiliki keunggulan dibandingkan Random Forest, dengan MSE sebesar 0.0373 dan R^2 sebesar 0.936, menunjukkan keunggulan LightGBM dalam menangani dataset yang besar dan kompleks[7].

Penelitian yang telah dilakukan sebelumnya juga memprediksi harga mobil BMW menggunakan tools matlab dan algoritma Feed-Forward Backpropagation, dengan hasil yang menunjukkan bahwa metode ini memiliki tingkat kesalahan prediksi sekitar 11.46% dibandingkan harga aktual[8]. Hal ini menunjukkan bahwa pendekatan berbasis *machine learning* dapat meningkatkan efisiensi dan akurasi dalam estimasi harga mobil.

Meskipun berbagai pendekatan telah dilakukan, masih terdapat kesenjangan dalam penelitian sebelumnya terkait evaluasi perbandingan *ensemble learning* dengan regresi linear dalam skala yang lebih luas, seperti penelitiannya[8] hanya berfokus pada satu model tertentu tanpa mempertimbangkan perbandingan langsung pada dataset yang lebih luas dan bervariasi dengan berbagai merek dan model mobil. Selain itu, masih sedikit penelitian yang mengevaluasi kinerja model dalam konteks dataset dataset yang lebih representatif terhadap kondisi pasar sebenarnya[9]

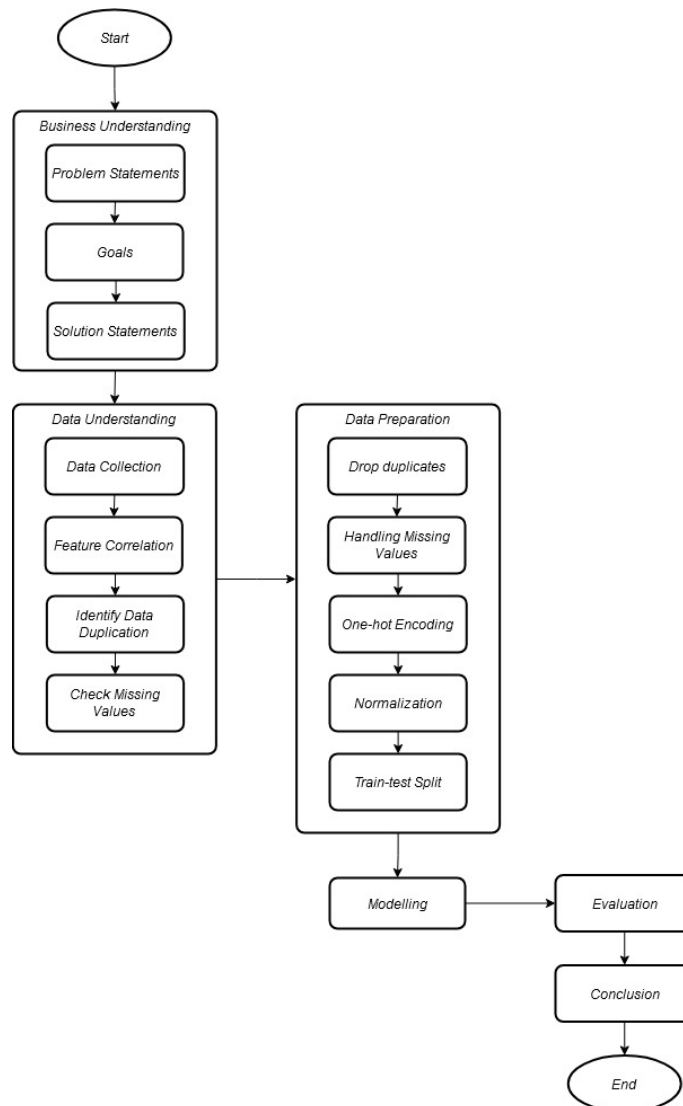
Oleh karena itu, penelitian ini melakukan evaluasi terhadap regresi linear serta metode *ensemble learning* (Random Forest dan Gradient Boosting) dalam memprediksi harga mobil, serta mengidentifikasi fitur kendaraan yang paling berpengaruh terhadap harga jual. Selain itu penelitian ini mengevaluasi performa model berdasarkan metrik Mean Squared Error (MSE) dan R-squared (R^2). Dengan demikian, hasil penelitian ini diharapkan dapat menjadi rekomendasi bagi industri otomotif dalam memilih model prediksi harga kendaraan yang lebih akurat dan efektif untuk digunakan oleh dealer, produsen, maupun konsumen.

2. METODOLOGI PENELITIAN

Penelitian ini dilakukan untuk evaluasi model *machine learning* dalam memprediksi harga mobil yaitu menggunakan algoritma Random Forest, Gradient Boosting, dan Regresi Linear. Pemilihan algoritma ini didasarkan pada efektivitasnya dalam analisis prediktif, di mana Gradient Boosting diharapkan memberikan kinerja terbaik karena kemampuannya dalam mengoreksi kesalahan secara iteratif[10]. Algoritma Random Forest dipilih karena kemampuannya dalam menangani data yang kompleks dan meningkatkan akurasi prediksi melalui teknik *ensemble learning*. Dalam konteks prediksi harga mobil bekas, Random Forest telah menunjukkan kinerja yang baik dalam menangani berbagai variabel yang mempengaruhi harga, seperti tahun pembuatan, jarak tempuh, dan kondisi kendaraan[11]. Sementara itu, Linear Regression dipilih karena kesederhanaannya dan kemampuannya dalam menunjukkan hubungan linear antara variabel independen dan dependen. Dalam penelitian sebelumnya, algoritma regresi linear telah digunakan untuk memprediksi harga dengan hasil yang cukup akurat, terutama ketika hubungan antara variabel bersifat linear[12]. Dengan menggabungkan ketiga algoritma ini, diharapkan dapat diperoleh model prediksi harga mobil yang akurat dan andal, dengan mempertimbangkan keunggulan masing-masing algoritma dalam menangani data dan variabel yang berbeda.

Tools yang digunakan adalah bahasa pemrograman python dan google colab. *Library* atau pustaka yang digunakan diantaranya adalah panda, NumPy, matplotlib, seaborn, LinearRegression, RandomForestRegressor, GradientBoostingRegressor. Pandas digunakan untuk membaca data dan mengkonversi data ke *dataframe*. NumPy digunakan untuk pembersihan data dan eksplorasi data, matplotlib dan seaborn digunakan untuk visualisasi data. LinearRegression, RandomForestRegressor, GradientBoostingRegressor digunakan untuk tahap pemodelan.

Sebuah kerangka sistematika penelitian yang terstruktur diperlukan untuk memastikan pelaksanaan yang tepat. Sistematika proses penelitian ini seperti pada gambar 1 menunjukkan langkah-langkah yang dilakukan untuk menyelesaikan permasalahan berdasarkan metodologi.



Gambar 1. Proses Penelitian

2.1. Data Understanding

Dataset yang digunakan bersumber dari Kaggle (Car Features and MSRP) yang dikembangkan CooperUnion dan Jeffrey Shih. Dataset ini berisi informasi mengenai spesifikasi dan harga mobil. Dataset ini memiliki 11914 baris data dan 16 kolom/fitur. Gambar 2 menunjukkan 5 baris pertama dataset.

	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity	MSRP
0	BMW	Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135
1	BMW	Series 1	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916	40650
2	BMW	Series 1	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350
3	BMW	Series 1	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	29450
4	BMW	Series 1	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916	34500

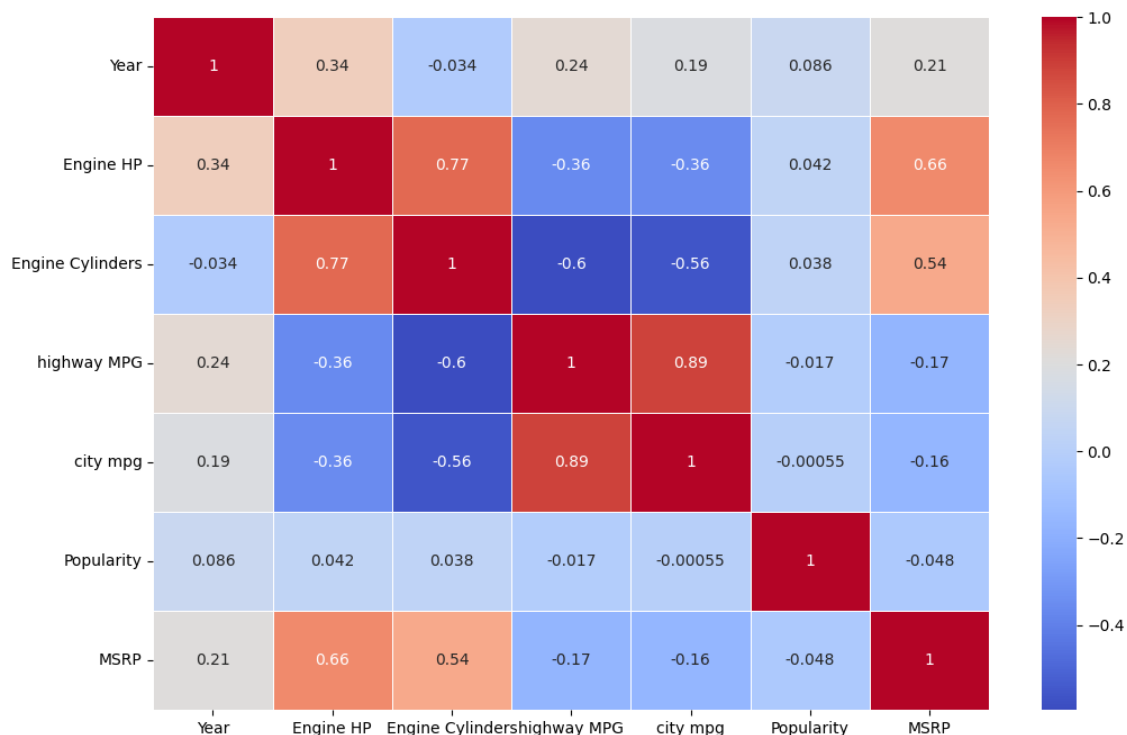
Gambar 2. Informasi Dataset

Berikut Tabel 1 merupakan deskripsi dari kolom pada dataset.

Tabel 1. Deskripsi Kolom atau Fitur

Nama Kolom	Deskripsi	Tipe Data
Make	Merek Mobil	Object
Model	Model Mobil	Object
Year	Tahun Produksi	Integer
Engine Fuel Type	Jenis bahan bakar mesin	Object
Engine HP	Tenaga mesin	Float
Engine Cylinders	Jumlah silinder pada mesin	Float
Transmission Type	Jenis transmisi (misalnya: manual atau otomatis)	Object
Driven_wheels	Tipe roda penggerak (misalnya: penggerak roda depan atau belakang)	Object
Number of Doors	Jumlah pintu mobil	Float
Market Category	Kategori pasar mobil (seperti luxury, crossover, dll.)	Object
Vehicle Size	Ukuran kendaraan (misalnya: <i>compact</i> , <i>midsize</i> , <i>large</i>)	Object
Vehicle Style	Gaya kendaraan (misalnya: sedan, SUV)	Object
Highway MPG	Efisiensi bahan bakar (Konsumsi bahan bakar di jalan raya dalam miles per gallon)	Integer
City mpg	Efisiensi bahan bakar di dalam kota dalm MPG	Integer
Popularity	Indikator popularitas mobil	Integer
MSRP	Harga mobil yang disarankan (<i>Manufacturer's Suggested Retail Price</i>)	Integer

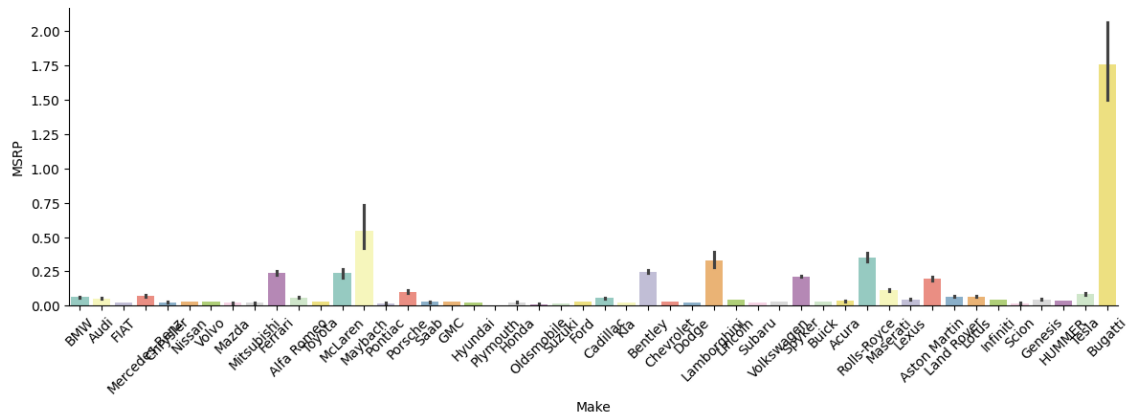
Secara keseluruhan dataset ini memiliki 3 tipe data yaitu *object*, *integer* dan *float*. *Object* merupakan tipe data yang berisi teks atau string. Matriks korelasi fitur numerik ditunjukkan pada gambar 3.



Gambar 3. Matriks Korelasi untuk Fitur Numerik

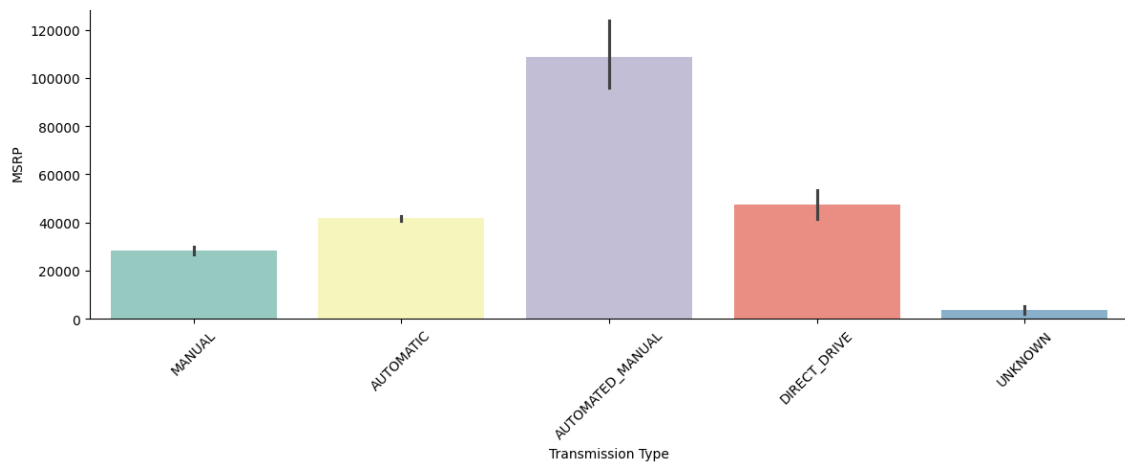
Matriks korelasi menggambarkan hubungan antar fitur numerik dalam dataset, membantu mengidentifikasi fitur yang relevan untuk analisis atau model prediktif. Gambar 3 menunjukkan korelasi tinggi antara Engine HP dan Engine Cylinders (0.77), serta City MPG dan Highway MPG (0.89), yang menunjukkan bahwa daya mesin lebih besar cenderung meningkatkan harga mobil. Sebaliknya, Highway MPG dan Engine Cylinders memiliki korelasi negatif (-0.6), menunjukkan bahwa mobil dengan lebih banyak silinder biasanya kurang efisien bahan bakarnya.

Untuk korelasi fitur kategorik atau object terdiri dari beberapa fitur seperti yang ditunjukkan pada gambar 4 untuk fitur Make, gambar 5 untuk fitur transmisi, gambar 6 fitur driven_wheels, gambar 7 fitur vehicle size dan gambar 8 fitur vehicle style.



Gambar 4. Rata-rata MSRP Relatif Terhadap Fitur Make

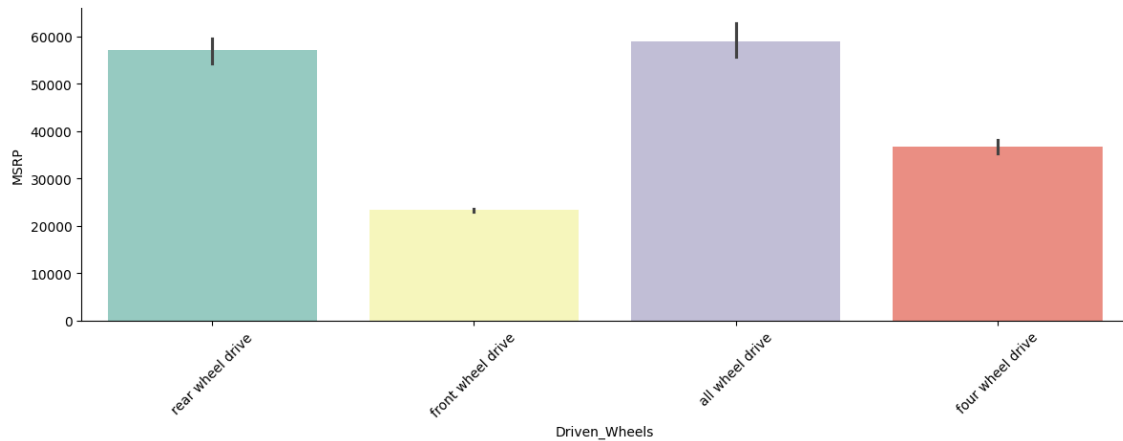
Grafik pada Gambar 4 menunjukkan rata-rata harga (MSRP) berdasarkan merek mobil, yang mengungkapkan pengaruh merek terhadap harga jual. Merek seperti Bugatti dan Maybach memiliki MSRP jauh lebih tinggi, mencerminkan status mereka sebagai produsen mobil mewah. Sebaliknya, merek seperti Hyundai, Honda, dan Chevrolet memiliki MSRP lebih rendah, menargetkan pasar yang lebih terjangkau. Variasi harga yang signifikan terlihat pada merek seperti McLaren dan Rolls-Royce, yang melayani berbagai segmen harga. Grafik ini penting untuk memahami perbedaan pasar dan kontribusi merek dalam penetapan harga.



Gambar 5. Rata-rata MSRP Relatif Terhadap Fitur Transmission Type

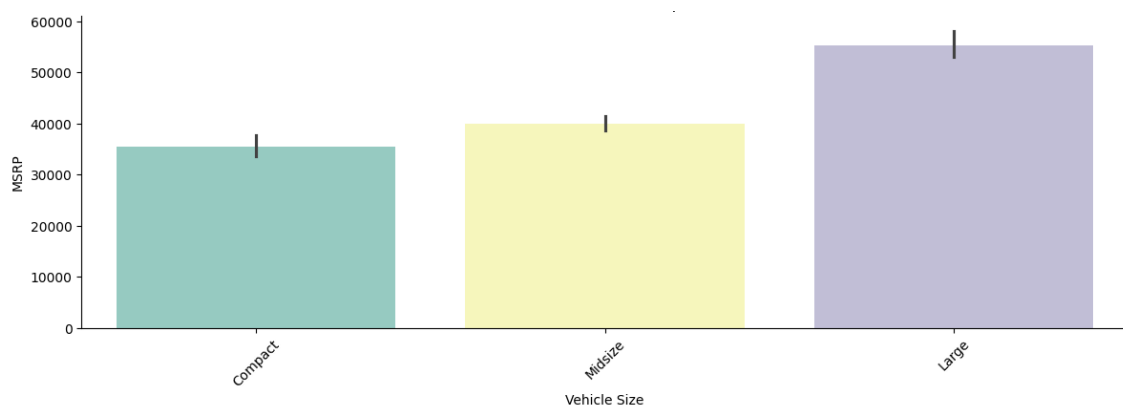
Grafik gambar 5 menunjukkan rata-rata harga (MSRP) berdasarkan jenis transmisi mobil, yang mengungkapkan pengaruh tipe transmisi terhadap harga pasar kendaraan. Mobil dengan transmisi Automated Manual memiliki MSRP tertinggi, umumnya ditemukan pada mobil premium atau performa tinggi. Sementara transmisi Manual dan Automatic memiliki MSRP lebih rendah, cocok dengan mobil konvensional. Transmisi Direct Drive, yang ada pada mobil listrik, memiliki harga lebih tinggi dibandingkan Manual atau Automatic, namun masih di bawah Automated Manual. Kategori Unknown memiliki MSRP terendah,

kemungkinan karena data yang tidak lengkap. Grafik ini menggambarkan bagaimana tipe transmisi mempengaruhi harga jual kendaraan.



Gambar 6. Rata-rata MSRP Relatif Terhadap Fitur Driven_Wheels

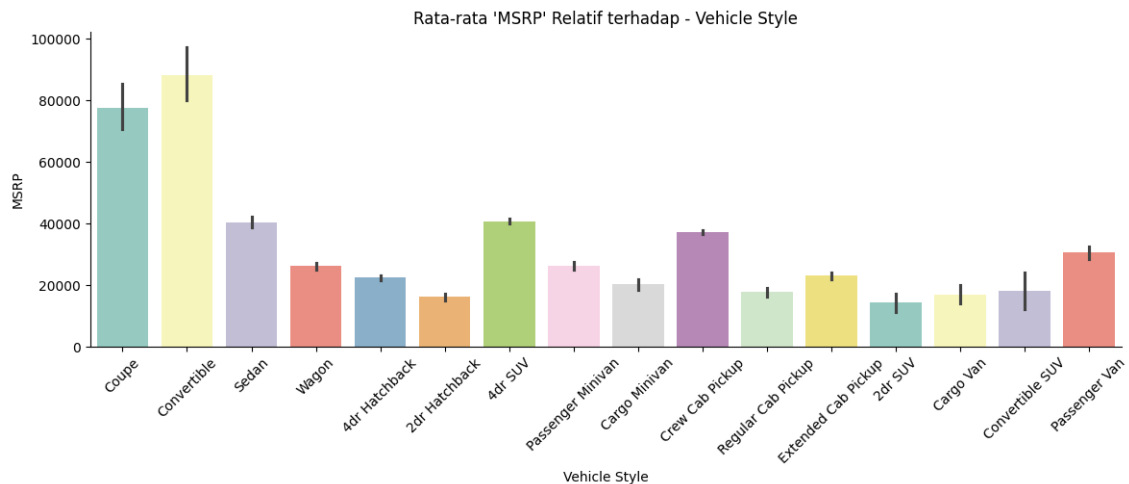
Gambar 6 menunjukkan rata-rata MSRP berdasarkan jenis sistem penggerak roda (Driven Wheels), yang mempengaruhi harga pasar kendaraan. Mobil dengan All Wheel Drive memiliki MSRP tertinggi, biasanya ditemukan pada kendaraan premium atau performa tinggi seperti SUV. Rear Wheel Drive juga memiliki MSRP tinggi, sering dikaitkan dengan mobil sport atau sedan mewah. Sebaliknya, Front Wheel Drive memiliki MSRP lebih rendah, mencerminkan fokus pada efisiensi dan keterjangkauan. Four Wheel Drive sedikit lebih tinggi dari Front Wheel Drive, digunakan pada kendaraan yang membutuhkan kapabilitas di medan sulit. Grafik ini menggambarkan pengaruh sistem penggerak terhadap nilai pasar kendaraan.



Gambar 7. Rata-rata MSRP Relatif Terhadap Fitur Vehicle Size

Gambar 7 menunjukkan rata-rata MSRP berdasarkan ukuran kendaraan, yang mengungkapkan bagaimana dimensi kendaraan memengaruhi harga. Kendaraan berukuran Large memiliki MSRP tertinggi, mencerminkan harga mobil besar seperti SUV mewah. Midsize memiliki harga rata-rata menengah, seringkali mewakili

kendaraan keluarga dengan keseimbangan performa dan efisiensi. Compact memiliki MSRP terendah, sesuai dengan karakteristik mobil kecil dan efisien untuk segmen pasar ekonomis. Grafik ini penting untuk memahami segmentasi pasar otomotif berdasarkan ukuran kendaraan.



Gambar 8. Rata-rata MSRP Relatif Terhadap Fitur Vehicle Style

Gambar 8 menunjukkan rata-rata MSRP berdasarkan desain kendaraan, mengungkapkan bagaimana desain memengaruhi harga pasar. Kendaraan dengan gaya Convertible dan Coupe memiliki MSRP tertinggi, mencerminkan mobil premium atau sport dengan desain eksklusif. 4dr SUV juga memiliki MSRP tinggi, menandakan popularitas SUV mewah. Sebaliknya, gaya seperti 4dr Hatchback dan Cargo Minivan memiliki MSRP lebih rendah, sesuai dengan kendaraan fungsional atau ekonomis. Variasi harga ini menggambarkan bagaimana desain, fungsi, dan target pasar memengaruhi nilai kendaraan, serta memberikan panduan penting bagi konsumen dan produsen.

Selanjutnya dilakukan pemeriksaan duplikasi data pada dataset. Serta melakukan identifikasi *missing value* untuk menghitung jumlah *missing values* pada setiap kolom dataset. Beberapa kolom terdeteksi *missing values* yaitu Engine Fuel Type memiliki 3 nilai, Engine HP sebanyak 69 nilai, Engine Cylinders sebanyak 30 nilai, Number of Doors memiliki 6 nilai, dan Market Category sebanyak 3376 nilai. Identifikasi *missing values* ini membantu mengidentifikasi kolom mana yang memerlukan penanganan, seperti imputasi atau penghapusan baris/kolom.

2.2. Data Preparation

Data Preparation merupakan proses menyiapkan data dengan cara membersihkan, mengorganisasikan dan memanipulasi data menggunakan bahasa python sehingga data siap untuk dianalisis pada pemodelan regresi[13]. Dalam tahap persiapan data ini dilakukan pembersihan data dengan menghapus data duplikat yang terdiri dari 715 baris duplikat.

Kemudian dilakukan penanganan *missing values* dengan menggunakan metode imputasi berbasis statistik. Metode imputasi ini menjaga informasi dalam data sebanyak mungkin tanpa menghapus baris yang mengandung nilai kosong. Fungsi median digunakan untuk mengisi nilai kosong pada kolom numerik dengan nilai tengah (median) dari kolom tersebut. Median digunakan untuk mengurangi dampak *outlier* pada data, sehingga lebih representatif dibandingkan rata-rata. Kemudian fungsi mode digunakan untuk mengisi nilai kosong pada kolom non-numerik (kategorik) dengan nilai yang paling sering muncul (modus) di kolom tersebut. Mode dipilih karena mode merepresentasikan kategori yang paling umum. Strategi ini memastikan dataset tetap utuh dan representatif.

Selanjutnya menghapus kolom yang tidak relevan atau tidak memiliki pengaruh signifikan terhadap target prediksi harga (MSRP) seperti *Market Category*, dan *Number of Doors*. Hal ini bertujuan mengurangi kompleksitas dataset dengan hanya menyertakan fitur yang relevan untuk prediksi harga, selain itu menghindari memasukkan kolom yang dapat menyebabkan noise atau mengurangi akurasi model. Langkah ini merupakan bagian penting dari proses seleksi fitur, memastikan hanya fitur yang relevan digunakan untuk pemodelan.

Berikutnya pemisahan fitur yang relevan dan target. Fitur atau kolom yang relevan adalah variabel yang dipilih berdasarkan signifikansi statistik dan hubungan yang jelas dengan target. Fitur relevan ini yang mempengaruhi harga jual kendaraan (MSRP) sebagai target. Proses pemisahan antara fitur (X) dan target (y) dilakukan untuk memastikan model hanya mempelajari hubungan dengan target, menghindari noise, dan mengoptimalkan akurasi prediksi. Dengan demikian, target dipisahkan sebagai *output* yang diprediksi, sementara fitur relevan disiapkan untuk dilatih dalam model *machine learning*. Pendekatan ini mendukung analisis yang terstruktur dan memungkinkan interpretasi yang lebih baik terhadap hasil prediksi.

Selanjutnya teknik *one-hot encoding* yang merupakan teknik untuk mengubah data kategorik menjadi data numerik dimana setiap kategori menjadi kolom baru dengan nilai 0 atau 1. Fitur atau kolom yang akan diubah menjadi numerik adalah *make*, *model*, *engine fuel type*, *transmission type*, *driven_wheels*, *vehicle size*, dan *vehicle type*.

Setelah itu dilakukan normalisasi yang bertujuan untuk menyelaraskan skala data agar algoritma *machine learning* memiliki performa yang lebih baik dan bekerja optimal. Teknik normalisasi yang digunakan adalah Standarisasi dengan *StandardScaler* dari pustaka *sklearn*.

Selanjutnya *train-test split* yaitu membagi data menjadi data latih (*train*) dan data uji (*test*). Data latih akan digunakan untuk membangun model, sedangkan data uji akan digunakan untuk menguji performa model. Setelah melalui *preprocessing* menjadi sebanyak 11199 baris data yang kemudian dibagi menjadi 80% untuk data latih yaitu 8959 dan untuk data uji menjadi 20% yaitu 2240.

2.2. Modelling

Pada tahap ini, melatih model menggunakan pendekatan Random Forest, Gradient Boosting dan regresi linear dalam memprediksi harga mobil (MSRP).

Random Forest

Random Forest adalah metode *ensemble* yang menggabungkan beberapa pohon keputusan untuk menghasilkan prediksi yang lebih stabil dan akurat[3]. Menggunakan pustaka `sklearn.ensemble.RandomForestRegressor` dengan memasukkan `X_train` dan `y_train` dalam membangun model. Parameter yang digunakan adalah `random_state=42`, `n_estimators=100`, `max_depth=20`.

Gradient Boosting

Gradient Boosting merupakan algoritma machine learning berbasis *ensemble*, yang bekerja dengan menggabungkan beberapa model pohon keputusan secara iteratif untuk mengurangi kesalahan prediksi, sehingga lebih kuat dibandingkan metode pohon tunggal[2]. Menggunakan pustaka `sklearn.ensemble.GradientBoostingRegressor` dan parameter `n_estimators=200`, `max_depth=5`, `learning_rate=0.2`.

Regresi Linear

Regresi linear adalah metode statistik yang digunakan untuk memprediksi nilai kontinu berdasarkan hubungan antara variabel dependen (target) dan satu atau lebih variabel independen (fitur). Menggunakan pustaka `sklearn.linear_model.LinearRegression` dengan parameter `fit_intercept=True`, `n_jobs=-1`.

2.4. Evaluation

Mean Square Error

Mean Square Error (MSE) adalah salah satu metrik evaluasi yang umum digunakan dalam statistik dan *machine learning*. Fungsi dari MSE adalah untuk mengukur seberapa baik suatu model dalam memetakan nilai prediksi ke nilai sebenarnya dengan menggunakan kuadrat kesalahan sekecil mungkin sebagai dasar perhitungan. Berikut rumus yang digunakan untuk menghitung MSE[7].

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

Keterangan:

n adalah jumlah total observasi atau sampel.

y_i adalah nilai aktual atau nilai sebenarnya dari observasi ke- i .

$\hat{y}_i$ adalah nilai prediksi model untuk observasi ke- i .

R Squared

R Squared (R^2) adalah ukuran statistik yang menunjukkan seberapa besar variasi suatu variabel dependen dapat dijelaskan oleh variabel independen dalam suatu

model regresi. Dengan kata lain, R^2 menunjukkan seberapa baik suatu model regresi (variabel independen) memprediksi hasil data observasi (variabel dependen). R^2 merupakan angka yang berkisar antara 0 sampai 1 yang mengindikasikan besarnya kombinasi variabel independen secara bersama-sama mempengaruhi nilai variabel dependen. Perhitungan R Squared menggunakan rumus berikut[2]:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Keterangan:

y_i = Nilai aktual pada data ke-i

$\hat{y}_i$ = Nilai prediksi dari model pada data ke-i

$\bar{y}_i$ = Rata-rata dari semua nilai y_i dalam dataset

3. HASIL DAN PEMBAHASAN

3.1 Hasil Evaluasi

Hasil evaluasi menggunakan metrik evaluasi MSE dan R^2 dengan parameter yang telah ditentukan pada Random Forest, Gradient Boosting dan Regresi Linear ditunjukkan pada tabel 2.

Tabel 2. Hasil Evaluasi Metrik

Model	MSE	R^2
Random Forest	486169848.307284	0.899657
Gradient Boosting	175063625.591782	0.963868
Regresi Linear	2820305019.141654	0.417905

Tabel 2 menunjukkan evaluasi kinerja Random Forest, Gradient Boosting, dan Regresi linear berdasarkan metrik MSE (Mean Squared Error) dan R^2 . Gradient Boosting menunjukkan performa terbaik dengan R^2 sebesar 0.963868 dan MSE terendah, menandakan bahwa model ini memiliki kemampuan prediksi yang sangat baik dengan error yang kecil. Gradient Boosting memiliki nilai R^2 mendekati 1, yang berarti model ini mampu menunjukkan variabilitas data dengan sangat baik. Random Forest juga memiliki performa yang cukup baik dengan R^2 sebesar 0.899657, meskipun MSE-nya lebih tinggi dibanding Gradient Boosting. Sebaliknya, Regresi Linear memiliki performa yang jauh lebih rendah, dengan R^2 hanya 0.417905 dan MSE yang tinggi, menunjukkan bahwa model ini kurang cocok untuk data yang lebih kompleks. Jadi Model Gradient Boosting sebagai pilihan terbaik untuk memaksimalkan akurasi prediksi pada penelitian ini.

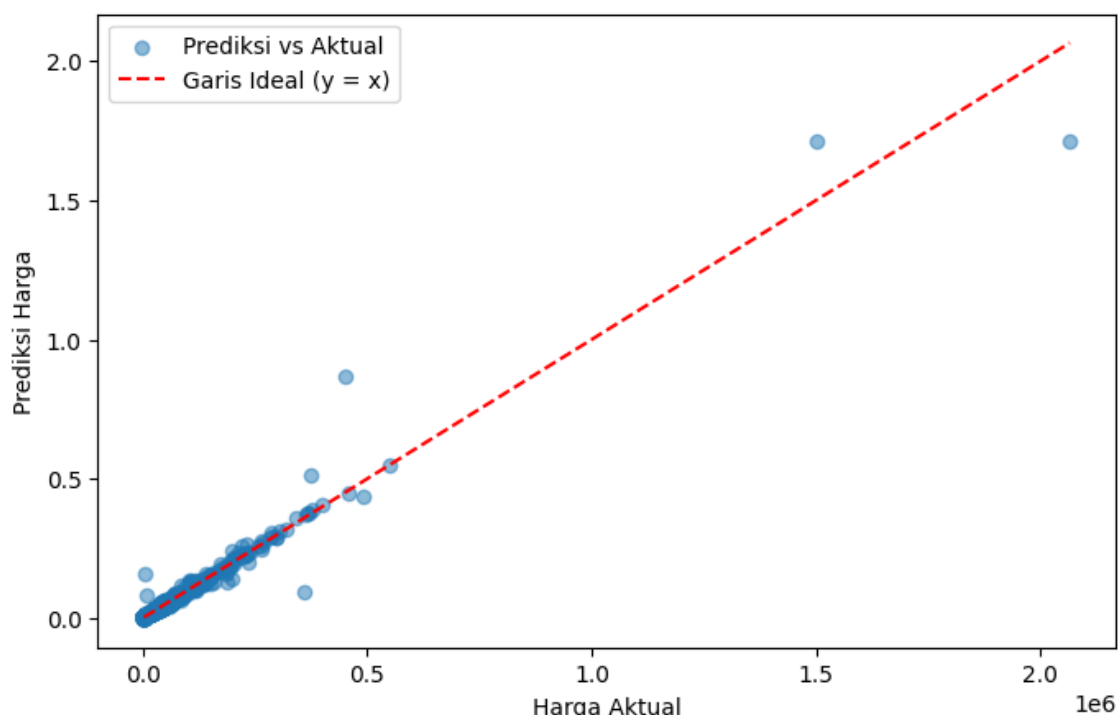
3.2 Hasil Prediksi Harga dengan Pengujian

Tabel 3 menampilkan lima hasil perbandingan antara harga aktual (MSRP) dan harga prediksi dalam satuan dolar.

Tabel 3. Perbandingan Harga Aktual dan Prediksi

Merek	Model	MSRP/Harga	Prediksi
Ford	E-Series Wagon	34160	33861
Honda	CR-V	33395	28815
Suzuki	Equator	19299	19420
Dodge	Journey	25195	23691
Chevrolet	Silverado 1500	54925	47279

Berdasarkan tabel 3, menunjukkan bahwa nilai harga yang diprediksi oleh model cukup mendekati harga aktual. Seperti mobil Ford E-Series Wagon memiliki harga aktual sebesar 34160, sementara model memprediksi sebesar 33861, yang menunjukkan prediksi cukup akurat, begitupun mobil Dodge Journey. Suzuki Equator prediksinya sangat akurat. Mobil lainnya seperti Honda CR-V dan Chevrolet Silverado 1500 memiliki selisih prediksi yang lebih rendah dibandingkan harga asli. Meskipun masih terdapat selisih yang lebih rendah antara harga aktual dan prediksi, hasil ini menunjukkan bahwa model telah berhasil menangkap pola harga kendaraan dengan cukup baik. Visualisasi hasil prediksi ditunjukkan pada gambar 8.



Gambar 9. Visualisasi Prediksi Harga

Berdasarkan Gambar 8 dengan grafik scatter plot menunjukkan hubungan antara harga aktual (sumbu X) dan harga prediksi (sumbu Y). Titik-titik biru merepresentasikan perbandingan antara harga aktual dan hasil prediksi, sementara garis merah putus-putus adalah garis ideal ($y = x$) yang menunjukkan prediksi

sempurna. Mayoritas titik berada dekat dengan garis ideal, menandakan bahwa model mampu melakukan prediksi dengan cukup baik, meskipun terdapat beberapa titik yang menyimpang dari garis tersebut.

Beberapa titik jauh di atas atau di bawah garis menunjukkan bahwa model masih mengalami sedikit kesalahan prediksi pada beberapa sampel, terutama pada harga yang lebih tinggi. Hal ini bisa disebabkan oleh kurangnya fitur yang menangkap faktor harga tertentu atau adanya variasi harga yang sulit diprediksi oleh model. Secara keseluruhan, distribusi titik yang mendekati garis ideal mengindikasikan bahwa model Gradient Boosting memiliki performa yang cukup baik dalam memprediksi harga mobil.

4. KESIMPULAN

Berdasarkan hasil evaluasi, Gradient Boosting menunjukkan performa terbaik dengan nilai R^2 sebesar 0.963868 dan MSE terendah, model ini memiliki kemampuan prediksi yang sangat baik dengan error yang kecil. Sementara itu, Random Forest memiliki akurasi cukup baik dengan R^2 sebesar 0.899657, sedangkan Regresi Linear menunjukkan performa yang jauh lebih rendah dengan R^2 hanya 0.417905, menandakan keterbatasannya dalam menangani data yang kompleks. Sehingga dapat disimpulkan bahwa model Gradient Boosting memiliki kemampuan prediksi yang sangat baik dengan error yang kecil yang mampu menghasilkan estimasi yang cukup mendekati harga aktual. Visualisasi hasil prediksi menunjukkan bahwa sebagian besar data berada dekat dengan garis ideal, menandakan bahwa model dapat menangkap pola harga kendaraan dengan cukup baik. Namun, masih terdapat beberapa hal yang dapat diperbaiki melalui optimasi hyperparameter, fitur atau peningkatan kualitas data. Secara keseluruhan, penelitian ini menegaskan bahwa Gradient Boosting merupakan model terbaik dalam memprediksi harga mobil, dengan potensi penerapan lebih lanjut dalam industri otomotif untuk meningkatkan akurasi estimasi harga kendaraan bagi produsen, dealer, maupun konsumen. Penelitian ini dapat dikembangkan lebih lanjut dengan mengintegrasikan metode *machine learning* lainnya, memperkaya dataset, atau mengoptimalkan fitur yang digunakan untuk meningkatkan keakuratan model prediksi melalui optimasi hyperparameter, atau penggunaan model ensemble tambahan.

DAFTAR PUSTAKA

- [1] V. Viswanatha, A. C. Ramachandra, B. D. Parameshachari, H. V. Vachan, and S. S. Shetty, "Predicting the Price of used Cars using Machine Learning," *2023 Int. Conf. Evol. Algorithms Soft Comput. Tech. EASCT 2023*, pp. 1–6, 2023, doi: 10.1109/EASCT59475.2023.10393486.
- [2] F. A. Alghifari, R. Andreswari, and E. Sutoyo, "Used Cars Price Prediction In DKI Jakarta Using Extreme Gradient Boosting And Bayesian Optimization Algorithm," *Proc. - Int. Conf. Adv. Data Sci. E-Learning Inf. Syst. ICADEIS 2022*,

- pp. 1–5, 2022, doi: 10.1109/ICADEIS56544.2022.10037301.
- [3] A. F. Kinadi, R. Andreswari, E. Sutoyo, R. Nugraha, and A. A. B. Kamil, "Used Car Price Prediction in Surabaya Using Random Forest Regressor Algorithms," *Proc. - Int. Conf. Adv. Data Sci. E-Learning Inf. Syst. ICADEIS 2022*, pp. 1–4, 2022, doi: 10.1109/ICADEIS56544.2022.10037526.
- [4] M. Ahmad *et al.*, "Car Price Prediction using Machine Learning," *2024 IEEE 9th Int. Conf. Conver. Technol. I2CT 2024*, pp. 3–7, 2024, doi: 10.1109/I2CT61223.2024.10544124.
- [5] C. Jin, "Price Prediction of Used Cars Using Machine Learning," *Proc. 2021 IEEE Int. Conf. Emerg. Sci. Inf. Technol. ICESIT 2021*, pp. 223–230, 2021, doi: 10.1109/ICESIT53460.2021.9696839.
- [6] M. Hankar, M. Birjali, and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," *11th Int. Symp. Signal, Image, Video Commun. ISIVC 2022 - Conf. Proc.*, pp. 1–4, 2022, doi: 10.1109/ISIVC54825.2022.9800719.
- [7] Y. Li, Y. Li, and Y. Liu, "Research on Used Car Price Prediction Based on Random Forest and LightGBM," *2022 IEEE 2nd Int. Conf. Data Sci. Comput. Appl. ICDSCA 2022*, pp. 539–543, 2022, doi: 10.1109/ICDSCA56264.2022.9988116.
- [8] N. O. Idris, A. Achban, S. A. Utiahman, J. Karim, and F. Pontooyo, "Predicting the selling price of cars using business intelligence with the feed-forward backpropagation algorithms," *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, 2020, doi: 10.1109/ICIC50835.2020.9288594.
- [9] S. Shaprapawad, P. Borugadda, and N. Koshika, "Car Price Prediction: An Application of Machine Learning," *6th Int. Conf. Inven. Comput. Technol. ICICT 2023 - Proc.*, no. Icict, pp. 242–248, 2023, doi: 10.1109/ICICT57646.2023.10134142.
- [10] L. P. Nasyuli, I. Lubis, and A. M. Elhanafi, "Penerapan Model Machine Learning Algoritma Gradient Boosting dan Linear Regression Melakukan Prediksi Harga Kendaraan Bekas," *J. Ilmu Komput. dan Sist. Informasi (JIRSI)*, vol. 2, no. 2, pp. 299–310, 2023, doi: 10.70340/jirsi.v2i2.56.
- [11] P. H. Putra, A. Azanuddin, B. Purba, and Y. A. Dalimunthe, "Random Forest and Decision Tree Algorithms for Car Price Prediction," *J. Mat. Dan Ilmu Pengetah. Alam LLDikti Wil. 1*, vol. 4, no. 1, pp. 81–89, 2023, doi: 10.54076/jumpa.v3i2.305.
- [12] D. Miftahul Huda, G. Dwilestari, A. Rizki Rinaldi, and Lin, "Prediksi Harga Mobil Bekas Menggunakan Algoritma Regresi Linear Berganda," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 150–157, 2024, doi: 10.36499/jinrpl.v6i1.10266.
- [13] A. Testas, "Multiple Linear Regression with Pandas, Scikit-Learn, and PySpark," in *Distributed Machine Learning with PySpark: Migrating Effortlessly from Pandas and Scikit-Learn*, Berkeley, CA: Apress, 2023, pp. 53–74. doi: 10.1007/978-1-4842-9751-3_3.