

Analisis Pengelompokan dan Prediksi Harga Kendaraan Listrik Menggunakan *K-Means Clustering* dan Regresi Linier Berganda

Kharla Septiana ¹, Afni Cahaya Putri ², Nadya Monalisa ³, Dewi Astuti ⁴, Etis Sunandi ⁵, Wina Ayu Lestari ⁶

^{1,2,3,4,5,6} Statistika, Universitas Bengkulu

INFORMASI ARTIKEL

Diterima 17 Mei 2026
Direvisi 26 Juli 2026
Diterbitkan 27 Juli 2026

Kata kunci:

K-Means Clustering;
Regresi Linier Berganda;
Kendaraan Listrik;
Metrik Evaluasi;
Cross-Validation

ABSTRAK

Perkembangan industri kendaraan listrik telah meningkatkan kebutuhan akan model prediksi harga yang akurat untuk mendukung pengambilan keputusan berbasis data. Penelitian sebelumnya umumnya menerapkan metode pengelompokan dan regresi secara terpisah sehingga kontribusi hasil pengelompokan sebagai variabel tambahan dalam model prediksi belum banyak dievaluasi. Selain itu, potensi kebocoran informasi target akibat pelibatan variabel harga dalam proses pembentukan klaster juga belum banyak dikaji. Penelitian ini menerapkan metode *K-Means Clustering* untuk mengelompokkan kendaraan listrik berdasarkan karakteristik teknis sebelum membangun model prediksi harga. Penelitian ini berkontribusi melalui evaluasi empiris terhadap penggunaan hasil pengelompokan sebagai bentuk rekayasa fitur serta menunjukkan dampak kebocoran informasi target terhadap performa model. Dataset yang digunakan terdiri atas dua ribu observasi yang memuat variabel harga kendaraan, tahun produksi, kapasitas baterai, jarak tempuh, dan kecepatan pengisian daya. Jumlah klaster optimal ditentukan menggunakan Metode Siku. Hasil pengelompokan kemudian ditambahkan sebagai variabel kategorik pada model regresi linier berganda, sedangkan performa model dievaluasi menggunakan validasi silang sepuluh lipatan dengan metrik Akar Rata-Rata Kuadrat Galat, Rata-Rata Galat Mutlak, dan koefisien determinasi. Hasil penelitian menunjukkan bahwa penambahan variabel klaster yang dibentuk tanpa melibatkan variabel harga hanya memberikan perubahan performa yang sangat kecil. Nilai Akar Rata-Rata Kuadrat Galat meningkat dari 33.353,66 menjadi 33.357,02, sedangkan Rata-Rata Galat Mutlak menurun dari 26.200,81 menjadi 26.140,91. Koefisien determinasi hanya meningkat dari 0,1106 menjadi 0,1109 sehingga peningkatan kemampuan prediksi tidak bermakna secara praktis. Sebaliknya, pelibatan variabel harga dalam pembentukan klaster menghasilkan peningkatan performa yang semu akibat kebocoran informasi target. Dengan demikian, hasil pengelompokan tidak selalu meningkatkan kemampuan prediksi apabila dibentuk sesuai kaidah metodologis tanpa melibatkan variabel target.

Analysis of Electric Vehicle Clustering and Price Prediction Using K-Means Clustering and Multiple Linear Regression

ARTICLE INFO

Received May 17, 2026
Revised July 26, 2026
Published July 27, 2026

Keyword:

K-Means Clustering;
Multiple Linear Regression;
Electric Vehicles;
Evaluation Metrics;
Cross-Validation

ABSTRACT

The rapid growth of the electric vehicle industry increased the need for accurate price prediction models to support data-driven decision-making. Previous studies generally applied clustering and regression methods separately, so the contribution of clustering results as additional features in prediction models had not been widely evaluated. Furthermore, the potential for target leakage caused by including the price variable in the clustering process had received limited attention. This study applied the K-Means Clustering method to group electric vehicles based on their technical characteristics before developing a price prediction model. The study contributed by empirically evaluating the use of clustering results as a feature engineering approach and by demonstrating the impact of target leakage on model performance. The dataset consisted of 2,000 observations containing vehicle price, production year, battery capacity, driving range, and charging speed. The optimal number of clusters was determined using the Elbow Method. The clustering results were then incorporated as a categorical variable into a multiple linear regression model, and model performance was evaluated using ten-fold cross-validation with the Root Mean Squared Error, Mean Absolute Error, and coefficient of determination metrics. The results showed that adding cluster labels generated without the price variable produced only marginal changes in model performance. The Root Mean Squared Error increased slightly from 33,353.66 to 33,357.02, whereas the Mean Absolute Error decreased from 26,200.81 to 26,140.91. The coefficient of determination also increased only marginally from 0.1106 to 0.1109, indicating no practically meaningful improvement in predictive performance. In contrast, including the price variable in the clustering process produced an artificial improvement in model performance due to target leakage. Therefore, clustering results did not necessarily improve prediction performance when the clusters were generated using a methodologically sound approach that excluded the target variable.

This work is licensed under a [Creative Commons Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)



Corresponding Author:

Etis Sunandi, Statistika, Universitas Bengkulu
Email: esunandi@unib.ac.id

1. PENDAHULUAN

Perkembangan teknologi kendaraan listrik (*electric vehicle*/EV) dalam beberapa tahun terakhir mengalami peningkatan yang pesat seiring meningkatnya kebutuhan akan transportasi ramah lingkungan dan efisiensi energi. Kendaraan listrik menjadi salah satu solusi strategis dalam mengurangi emisi karbon dan ketergantungan terhadap bahan bakar fosil di berbagai negara [5]. Hossain et al. [5] menyatakan bahwa penggunaan EV secara luas berperan signifikan dalam mengurangi konsumsi bahan bakar fosil dan emisi gas rumah kaca pada sektor transportasi darat, khususnya apabila didukung oleh infrastruktur energi terbarukan yang memadai. Hal ini sejalan dengan temuan Gan et al. [4] yang menekankan bahwa besarnya pengurangan emisi gas rumah kaca dari penggunaan EV sangat bergantung pada sumber energi listrik yang digunakan.

Pertumbuhan penjualan kendaraan listrik secara global juga menunjukkan tren yang signifikan, didorong oleh berbagai kebijakan insentif di tingkat nasional maupun regional [11,19]. Sanguesa et

al. [11] mencatat bahwa EV semakin diminati karena penurunan harga dan meningkatnya kesadaran lingkungan, yang turut mendorong perkembangan teknologi baterai dan infrastruktur pengisian daya. Penelitian Sonar et al. [14] menegaskan bahwa harga, jarak tempuh, dan waktu pengisian daya merupakan faktor paling krusial yang memengaruhi keputusan konsumen dalam membeli kendaraan listrik. Kendaraan dengan kapasitas baterai yang lebih besar umumnya memiliki harga lebih tinggi karena mampu memberikan performa dan efisiensi energi yang lebih baik. Oleh karena itu, diperlukan pendekatan analisis berbasis data yang mampu membantu memahami karakteristik kendaraan sekaligus menghasilkan prediksi harga yang lebih baik.

Dalam analisis kendaraan listrik, metode *machine learning* dan *data mining* banyak digunakan untuk mendukung berbagai proses analisis berbasis data, termasuk untuk memprediksi penetrasi pasar dan mendukung pengambilan keputusan berbasis data [1]. Salah satu metode yang sering digunakan dalam proses pengelompokan data adalah K-Means Clustering. Sinaga dan Yang [13] membuktikan bahwa algoritma K-Means mampu mengelompokkan data secara efisien berdasarkan kemiripan karakteristik tanpa memerlukan label awal. Choi dan Jiao [2] berhasil menerapkan analisis kluster K-Means untuk mengungkap disparitas kepemilikan kendaraan listrik berdasarkan faktor demografis dan spasial, menunjukkan potensi besar metode ini dalam segmentasi pasar EV. Di sisi lain, prediksi harga kendaraan listrik juga menjadi topik penelitian yang penting. Khan et al. [8] menunjukkan bahwa pendekatan multi-model berbasis *machine learning*, termasuk regresi linier berganda, efektif digunakan untuk memprediksi spesifikasi dan konsumsi energi kendaraan listrik berdasarkan sejumlah variabel input. Selain itu, Uy et al. [16] menunjukkan bahwa metode K-Means Clustering efektif digunakan untuk mensegmentasi konsumen kendaraan listrik ke dalam kelompok-kelompok yang berbeda berdasarkan karakteristik demografis dan preferensi kepemilikan, sehingga mendukung analisis berbasis data yang lebih terstruktur.

Dalam membangun model prediksi, evaluasi performa merupakan tahapan yang tidak dapat diabaikan. Beberapa penelitian terdahulu menunjukkan bahwa evaluasi menggunakan metrik *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi (R^2) merupakan pendekatan yang umum digunakan dalam penelitian *machine learning* pada data kendaraan listrik, sebagaimana ditunjukkan oleh Hussain et al. [6]. Selain itu, penggunaan *cross-validation* diperlukan untuk mengukur performa model agar hasil yang diperoleh lebih akurat dan stabil, sebagaimana diterapkan oleh Ullah et al. [15] serta Pokharel et al. [10].

Meskipun penelitian terkait *clustering* dan prediksi kendaraan listrik telah banyak dilakukan, sebagian besar penelitian sebelumnya masih menerapkan kedua metode tersebut secara terpisah. Misalnya, Dubey et al. [3] hanya membandingkan performa model regresi linier, *decision tree*, dan *random forest* untuk memprediksi harga EV tanpa melibatkan proses segmentasi data. Sementara itu, Simsek et al. [12] dan Mishra et al. [9] berfokus pada peramalan permintaan dan penjualan EV menggunakan pendekatan *deep learning* tanpa memanfaatkan informasi kluster sebagai variabel prediktor. Di sisi lain, Zhang dan Kang [18] menerapkan *clustering* hanya untuk memprofilkan segmen pengguna EV secara deskriptif tanpa mengintegrasikan hasilnya ke dalam model prediksi harga. Kondisi tersebut menunjukkan bahwa kontribusi hasil *clustering* sebagai variabel tambahan dalam model prediksi masih belum banyak dievaluasi. Selain itu, penelitian mengenai integrasi *clustering* dan regresi juga masih menyisakan potensi permasalahan metodologis, seperti kebocoran informasi target (*target leakage*) apabila variabel target turut dilibatkan dalam proses pembentukan kluster.

Berdasarkan kesenjangan tersebut, penelitian ini menawarkan pendekatan yang berbeda dengan mengevaluasi secara sistematis penggunaan hasil K-Means Clustering sebagai variabel tambahan dalam model regresi linier berganda. Pendekatan ini merupakan salah satu bentuk *feature engineering* yang memungkinkan model memanfaatkan informasi segmentasi kendaraan berdasarkan kemiripan karakteristik teknis, selain hubungan numerik antarvariabel. Melalui pendekatan tersebut, penelitian ini tidak hanya mengevaluasi apakah hasil *clustering* mampu meningkatkan performa model prediksi harga kendaraan listrik, tetapi juga mengidentifikasi potensi keterbatasan metodologis yang dapat muncul apabila proses pembentukan kluster tidak dirancang secara tepat.

Kebaruan penelitian ini terletak pada evaluasi empiris penggunaan hasil *K-Means Clustering* sebagai bentuk *feature engineering* dalam model regresi linier berganda, sekaligus analisis dampak *target leakage* terhadap performa model prediksi harga kendaraan listrik.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis pengelompokan kendaraan listrik menggunakan metode K-Means Clustering, membangun model prediksi harga kendaraan listrik menggunakan regresi linier berganda, serta mengevaluasi kontribusi hasil *clustering* sebagai variabel tambahan terhadap performa model prediksi. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan analisis data berbasis *machine learning* serta menjadi referensi bagi penelitian maupun pengambilan keputusan pada industri kendaraan listrik. Secara metodologis, penelitian ini juga memberikan ilustrasi empiris mengenai penggunaan hasil *clustering* sebagai variabel kategorik tambahan dalam regresi linier berganda sekaligus menunjukkan pentingnya menghindari kebocoran informasi target dalam pengembangan model prediksi berbasis *machine learning*.

2. METODE

Penelitian ini menggunakan kombinasi pendekatan *unsupervised learning* dan *supervised learning* untuk menganalisis pengelompokan serta memprediksi harga kendaraan listrik berbasis data. Tahapan penelitian meliputi deskripsi dataset, pra-pemrosesan data, segmentasi menggunakan algoritma K-Means Clustering, pembangunan model prediksi menggunakan regresi linier berganda, pengujian asumsi klasik, dan evaluasi performa model. Khan et al. [8] menunjukkan bahwa pendekatan multi-model berbasis *machine learning* efektif dalam memprediksi berbagai karakteristik dan konsumsi energi kendaraan listrik, yang menjadi landasan pemilihan metode regresi dalam penelitian ini. Selanjutnya, performa model dievaluasi menggunakan metode *10-fold cross-validation* menggunakan metrik *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi (R^2).

Regresi linier berganda dipilih sebagai model dasar (*baseline*) dalam penelitian ini karena memiliki struktur yang sederhana dan koefisien yang mudah diinterpretasikan, sehingga pengaruh penambahan variabel hasil *clustering* terhadap performa model dapat diamati secara langsung. Perlu ditegaskan bahwa fokus penelitian ini bukan untuk mencari algoritma prediksi dengan akurasi tertinggi, melainkan untuk mengevaluasi sejauh mana informasi hasil pengelompokan tersebut memberikan kontribusi terhadap performa model prediksi harga kendaraan listrik. Selain itu, regresi linier berganda umum digunakan sebagai model dasar (*baseline*) sebelum dibandingkan dengan pendekatan *machine learning* yang lebih kompleks, sebagaimana pendekatan komparatif berbasis *machine learning* yang diterapkan pada studi prediksi kendaraan listrik lainnya [15].

2.1 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini adalah Electric Vehicle Market and Pricing Dataset yang diperoleh dari platform Kaggle dan dapat diakses melalui tautan <https://www.kaggle.com/datasets/patelris/electric-vehicle-market-and-pricing-dataset-2026>. Dataset terdiri atas 2.000 observasi dengan lima variabel numerik yang merepresentasikan karakteristik teknis kendaraan listrik dan harga kendaraan. Variabel yang digunakan dalam penelitian meliputi *price_usd* sebagai variabel respon yang merepresentasikan harga kendaraan listrik, serta *year*, *battery_capacity_kwh*, *range_miles*, dan *charging_speed_kw* sebagai variabel prediktor yang masing-masing merepresentasikan tahun produksi kendaraan, kapasitas baterai, jarak tempuh, dan kecepatan pengisian daya.

Dalam penelitian ini, variabel *year*, *battery_capacity_kwh*, *range_miles*, dan *charging_speed_kw* digunakan sebagai masukan pada proses K-Means Clustering, sedangkan *price_usd* digunakan sebagai variabel respon pada model regresi linier berganda. Selanjutnya, hasil *clustering* dimanfaatkan sebagai variabel kategorik tambahan dalam model regresi untuk mengevaluasi kontribusinya terhadap performa prediksi harga kendaraan listrik. Dengan demikian, dataset ini digunakan untuk menganalisis hubungan antara karakteristik teknis kendaraan listrik dan harga kendaraan sekaligus mengevaluasi pengaruh penambahan hasil *clustering* terhadap model prediksi.

2.2 Pra-Pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk memastikan data berada dalam kondisi yang siap digunakan dalam proses analisis dan pemodelan. Tahapan awal diawali dengan proses *data understanding* untuk mengetahui struktur data dan statistik deskriptif setiap variabel. Selanjutnya dilakukan pemeriksaan *missing value*. Data yang memiliki nilai kosong dihapus agar tidak memengaruhi proses analisis.

Karena variabel yang digunakan memiliki skala pengukuran yang berbeda, dilakukan standarisasi data menggunakan metode Z-score. Wongoutong [17] menunjukkan bahwa proses standarisasi dapat meningkatkan akurasi *clustering* pada dataset yang memiliki perbedaan skala antarvariabel. Proses standarisasi bertujuan agar setiap variabel memiliki rata-rata mendekati nol dan simpangan baku sebesar satu sehingga tidak terjadi dominasi variabel tertentu dalam proses K-Means Clustering.

2.3 Segmentasi Clustering

Tahap segmentasi data dilakukan menggunakan algoritma K-Means Clustering untuk mengelompokkan kendaraan listrik berdasarkan karakteristik spesifikasinya [7, 13]. Metode clustering telah banyak digunakan dalam analisis kendaraan listrik untuk mengidentifikasi pola karakteristik kendaraan maupun segmentasi pengguna berdasarkan atribut tertentu sehingga dapat mendukung pengambilan keputusan berbasis data [2].

Variabel yang digunakan dalam proses clustering meliputi tahun kendaraan (*year*), kapasitas baterai (*battery capacity*), jarak tempuh (*range miles*), dan kecepatan pengisian daya (*charging speed*). Variabel harga kendaraan (*price*) tidak diikutsertakan dalam proses clustering karena berperan sebagai variabel respon pada tahap pemodelan regresi linier berganda. Pengecualian variabel harga bertujuan untuk mencegah terjadinya kebocoran informasi (*target leakage*), yaitu kondisi ketika informasi dari variabel target digunakan dalam pembentukan variabel prediktor sehingga dapat menghasilkan performa model yang tampak lebih baik daripada kemampuan sebenarnya.

Penentuan jumlah kluster optimal dilakukan menggunakan Elbow Method, yaitu dengan mengevaluasi perubahan nilai Within Cluster Sum of Squares (WSS) terhadap jumlah kluster [7]. Elbow Method dipilih karena mampu mengidentifikasi jumlah kluster optimal secara sederhana melalui pola perubahan nilai WSS tersebut, sehingga metode ini banyak digunakan pada berbagai penelitian clustering karena kemudahan penerapan dan interpretasinya [7, 13]. Berdasarkan hasil Elbow Method, jumlah kluster optimal ditentukan sebanyak tiga kluster ($k = 3$). Selanjutnya, algoritma K-Means diterapkan untuk mengelompokkan seluruh observasi ke dalam tiga kelompok berdasarkan tingkat kemiripan karakteristik teknis kendaraan.

Hasil clustering kemudian digunakan sebagai variabel kategorik tambahan (*feature engineering*) pada model regresi linier berganda melalui pembentukan variabel *dummy*, serta divisualisasikan untuk mempermudah interpretasi karakteristik masing-masing kluster. Seluruh tahapan analisis, mulai dari pra-pemrosesan data, proses clustering, pemodelan regresi, hingga evaluasi model, dilakukan menggunakan perangkat lunak R, sedangkan visualisasi hasil clustering dibuat menggunakan paket *factoextra*. Untuk memastikan hasil clustering dapat direproduksi, nilai *random seed* ditetapkan sebesar 123 (*set.seed(123)*) sebelum proses K-Means Clustering dijalankan.

2.4 Analisis Komponen Utama (PCA)

Variabel kapasitas baterai dan jarak tempuh menunjukkan korelasi yang sangat tinggi, ditandai dengan nilai Variance Inflation Factor (VIF) yang jauh melampaui ambang batas 10 pada model regresi awal (lihat Bagian 3.3). Kondisi multikolinearitas ini menyebabkan pendugaan koefisien regresi menjadi tidak stabil dan *standard error* yang membesar, sehingga signifikansi masing-masing variabel sulit diinterpretasikan secara individual.

Untuk mengatasi permasalahan tersebut, dilakukan Analisis Komponen Utama (*Principal Component Analysis/PCA*) terhadap kedua variabel yang berkorelasi tinggi tersebut. PCA mentransformasikan kapasitas baterai dan jarak tempuh menjadi satu variabel komposit baru, yaitu

komponen utama pertama (PC1), yang mempertahankan sebagian besar informasi dari kedua variabel asli tanpa membawa permasalahan multikolinearitas. Sebelum dilakukan PCA, kedua variabel terlebih dahulu distandarisasi agar memiliki skala yang setara. Perlu dicatat bahwa PCA hanya diterapkan pada tahap pembentukan variabel prediktor untuk model regresi; variabel asli (kapasitas baterai dan jarak tempuh) tetap digunakan pada proses K-Means Clustering di Bagian 2.3, sehingga hasil segmentasi kluster tidak terpengaruh oleh transformasi ini.

Hasil PCA menunjukkan bahwa PC1 mampu menjelaskan 98,89% dari total variansi kedua variabel asli, sehingga komponen ini dinilai cukup memadai untuk mewakili informasi gabungan kapasitas baterai dan jarak tempuh. Kedua variabel asli memiliki kontribusi (loading) yang hampir setara terhadap PC1 (masing-masing sebesar -0,7071), yang mengonfirmasi bahwa keduanya memang merepresentasikan dimensi informasi yang serupa, yaitu kapabilitas jarak tempuh kendaraan listrik secara keseluruhan. Skor PC1 selanjutnya digunakan menggantikan variabel kapasitas baterai dan jarak tempuh pada model regresi linier berganda (Bagian 2.5), sementara variabel tahun produksi dan kecepatan pengisian daya tetap dipertahankan karena tidak menunjukkan indikasi multikolinearitas yang berarti.

2.5 Regresi Linier Berganda

Setelah proses segmentasi selesai, tahap selanjutnya adalah membangun model prediksi harga kendaraan listrik menggunakan regresi linier berganda. Model ini digunakan untuk menganalisis hubungan antara harga kendaraan listrik sebagai variabel respon dengan karakteristik teknis kendaraan sebagai variabel prediktor, dengan alasan pemilihan model telah dijelaskan pada awal Bagian 2.

Model regresi linier berganda yang digunakan pada penelitian ini adalah sebagai berikut.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_1 + \beta_3 X_2 + \varepsilon$	(1)
---	-----

Keterangan:

Y = harga kendaraan listrik

X_1 = tahun produksi kendaraan

Z_1 = skor komponen utama (PC1) hasil analisis komponen utama dari kapasitas baterai dan jarak tempuh

X_2 = kecepatan pengisian daya

ε = error

Untuk mengevaluasi pengaruh hasil segmentasi terhadap performa prediksi, dibangun dua model regresi, yaitu model tanpa variabel kluster sebagai model dasar (*baseline*) dan model dengan penambahan variabel kluster hasil K-Means Clustering. Pada model kedua, variabel kluster yang terdiri atas tiga kategori (Kluster 1, Kluster 2, dan Kluster 3) direpresentasikan menggunakan dua variabel dummy, yaitu C_2 untuk Kluster 2 dan C_3 untuk Kluster 3. Pendekatan ini mengikuti prinsip bahwa variabel kategorik dengan k kategori direpresentasikan menggunakan $(k - 1)$ variabel dummy untuk menghindari *dummy variable trap*. Kluster 1 digunakan sebagai kategori referensi, sehingga koefisien C_2 dan C_3 diinterpretasikan sebagai perbedaan rata-rata harga kendaraan pada Kluster 2 dan Kluster 3 terhadap Kluster 1 dengan variabel prediktor lainnya dianggap konstan.

2.6 Uji Asumsi Klasik

Sebelum dilakukan interpretasi terhadap model regresi linier berganda, dilakukan pengujian asumsi klasik untuk mengevaluasi apakah model memenuhi asumsi dasar regresi linier sehingga hasil estimasi koefisien dapat diinterpretasikan secara lebih tepat. Pengujian yang dilakukan meliputi uji normalitas residual, uji multikolinearitas, dan uji heteroskedastisitas. Uji normalitas residual dilakukan menggunakan Shapiro-Wilk Test untuk mengetahui apakah residual model berdistribusi normal. Residual dinyatakan tidak berdistribusi normal apabila diperoleh nilai $p\text{-value} < 0,05$. Uji multikolinearitas dilakukan menggunakan Variance Inflation Factor (VIF) untuk mendeteksi adanya multikolinearitas antarvariabel independen. Model dikatakan mengalami

indikasi multikolinearitas apabila nilai VIF melebihi 10. Selanjutnya, uji heteroskedastisitas dilakukan menggunakan Breusch-Pagan Test untuk mengetahui apakah varians residual bersifat konstan. Model dinyatakan mengalami heteroskedastisitas apabila diperoleh nilai $p\text{-value} < 0,05$. Hasil pengujian asumsi klasik digunakan sebagai dasar dalam mengevaluasi kelayakan model regresi serta menjadi pertimbangan dalam menginterpretasikan koefisien regresi yang dihasilkan.

Apabila hasil uji Breusch-Pagan menunjukkan adanya heteroskedastisitas ($p\text{-value} < 0,05$), maka standard error model regresi biasa (Ordinary Least Squares) menjadi tidak valid untuk digunakan dalam pengujian signifikansi koefisien. Pada kondisi tersebut, dilakukan estimasi ulang standard error menggunakan pendekatan heteroscedasticity-consistent standard error (robust standard error) tipe HC3. Nilai koefisien regresi (estimasi titik) tidak berubah dengan pendekatan ini, namun standard error, nilai t, dan p-value dihitung ulang agar valid meskipun asumsi homoskedastisitas tidak terpenuhi. Signifikansi masing-masing variabel prediktor yang dilaporkan pada Bagian 3.1 dan 3.2 mengacu pada hasil robust standard error ini.

2.7 Evaluasi Model

Evaluasi performa model dilakukan menggunakan 10-fold cross-validation untuk memperoleh estimasi kemampuan prediksi yang lebih stabil dan mengurangi ketergantungan hasil terhadap pembagian data latih dan data uji. Pendekatan evaluasi ini umum diterapkan dalam berbagai penelitian prediksi berbasis *machine learning* pada kendaraan listrik [15]. Jumlah 10 lipatan (*fold*) dipilih karena secara empiris memberikan keseimbangan yang baik antara bias dan varians dalam mengestimasi galat model, sekaligus tetap efisien secara komputasi sehingga menjadi skema evaluasi yang banyak digunakan dalam berbagai penelitian *machine learning*.

Pada setiap iterasi, dataset dibagi menjadi 10 bagian (*fold*). Sebanyak sembilan bagian digunakan sebagai data pelatihan (*training set*), sedangkan satu bagian digunakan sebagai data pengujian (*testing set*). Proses ini diulang sebanyak 10 kali hingga setiap *fold* pernah menjadi data pengujian satu kali. Nilai performa model kemudian diperoleh dari rata-rata hasil seluruh iterasi.

Performa model dievaluasi menggunakan tiga metrik, yaitu Root Mean Square Error (RMSE), Mean Absolute Error (MAE), dan koefisien determinasi (R^2). Ketiga metrik tersebut digunakan untuk membandingkan kemampuan prediksi model regresi linier berganda sebelum dan sesudah penambahan variabel kluster hasil K-Means Clustering.

a. Root Mean Square Error (RMSE)

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

b. Mean Absolute Error (MAE)

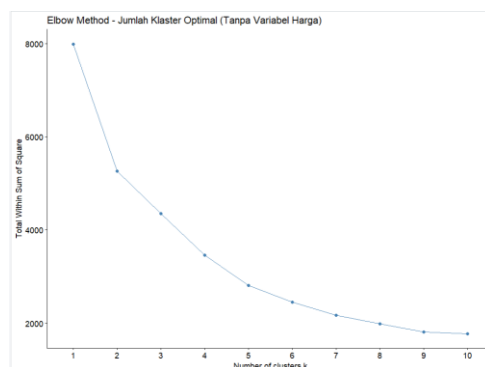
	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $	(3)
--	--	-----

c. Koefisien Determinasi (R^2)

	$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$	(4)
--	---	-----

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil analisis pengelompokan kendaraan listrik menggunakan metode K-Means Clustering dan hasil prediksi harga kendaraan listrik menggunakan regresi linier berganda. Selain itu, dilakukan evaluasi performa model berdasarkan nilai RMSE, MAE, dan R^2 yang diperoleh melalui 10-fold cross-validation. Tahap awal analisis dilakukan dengan menentukan jumlah kluster optimal menggunakan Elbow Method berdasarkan nilai Within Cluster Sum of Squares (WSS), seperti yang diperlihatkan pada Gambar 1.



Gambar 1. Grafik WSS terhadap Jumlah Kluster

Berdasarkan Gambar 1, terlihat bahwa nilai WSS mengalami penurunan yang cukup signifikan hingga $k = 3$. Setelah titik tersebut, penurunan nilai WSS cenderung melambat sehingga membentuk pola siku (*elbow*). Hal ini menunjukkan bahwa penambahan jumlah kluster setelah $k = 3$ tidak memberikan peningkatan yang signifikan terhadap kualitas pengelompokan data. Oleh karena itu, jumlah kluster optimal berdasarkan *Elbow Method* ditentukan sebanyak 3 kluster.

Selain pertimbangan berdasarkan grafik Elbow Method, pemilihan $k = 3$ juga didukung oleh interpretasi yang lebih informatif terhadap karakteristik teknis kendaraan listrik, karena mampu membedakan kelompok kendaraan berdasarkan kombinasi kapasitas baterai, jarak tempuh, dan kecepatan pengisian daya secara lebih rinci. Pemilihan tiga kluster dilakukan karena menghasilkan segmentasi yang lebih representatif terhadap perbedaan spesifikasi dan harga kendaraan listrik dibandingkan penggunaan dua kluster. Dengan demikian, jumlah kluster sebanyak tiga kelompok dinilai lebih sesuai untuk mendukung proses interpretasi data dan analisis prediksi harga kendaraan listrik. Tabel 1 merangkum hasil analisis dari berbagai jumlah kluster.

Tabel 1. Hasil *Clustering K-Means*

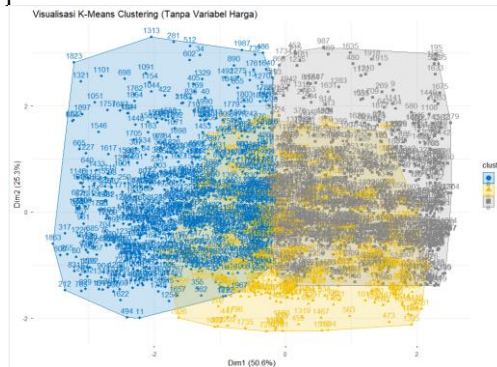
Klaster	n	Harga	year	Battery Capacity	Range (miles)	Charging Speed
1	368	94585	2024	65.9	230	276
2	731	83016	2024	95.5	336	161
3	901	69096	2024	61.3	214	101

Hasil *clustering* pada Tabel 1 menunjukkan adanya perbedaan karakteristik teknis yang jelas antar kelompok kendaraan listrik, meskipun variabel harga tidak dilibatkan dalam proses pembentukan kluster. Klaster 1 dicirikan oleh kecepatan pengisian daya tertinggi (276 kW) dengan kapasitas baterai dan jarak tempuh menengah, yang mengindikasikan kendaraan dengan teknologi pengisian cepat (*fast-charging*). Klaster 2 memiliki kapasitas baterai (95,5 kWh) dan jarak tempuh (336 mil) tertinggi sehingga merepresentasikan kendaraan berjangkauan jauh (*long-range*), sedangkan Klaster 3 memiliki nilai terendah pada hampir seluruh variabel teknis sehingga mencerminkan kendaraan dengan spesifikasi *entry-level*. Secara deskriptif, rata-rata harga tiap kluster adalah \$94.585 pada Klaster 1, \$83.016 pada Klaster 2, dan \$69.096 pada Klaster 3, yang menunjukkan bahwa perbedaan spesifikasi teknis tetap berasosiasi dengan tingkat harga meskipun harga tidak digunakan dalam pembentukan kluster. Ketiga kluster juga memiliki jumlah anggota yang relatif seimbang, yaitu 368 kendaraan (18,4%), 731 kendaraan (36,5%), dan 901 kendaraan (45,1%).

Pola pengelompokan tersebut menunjukkan bahwa Klaster 2 unggul pada kapasitas baterai dan jarak tempuh, mencerminkan hubungan fisik bahwa kapasitas baterai yang lebih besar mendukung jarak tempuh yang lebih jauh, sejalan dengan tingginya nilai *Variance Inflation Factor* pada kedua variabel tersebut (Tabel 4). Sebaliknya, Klaster 1 menonjol pada kecepatan pengisian daya, sedangkan Klaster 3 memiliki spesifikasi teknis terendah. Karena variabel pembentuk klaster saling berkorelasi tinggi, informasi yang dibawa oleh variabel klaster sebagian besar telah tercakup dalam variabel teknis asli sehingga penambahannya hanya memberikan kontribusi yang kecil terhadap akurasi prediksi harga.

Menariknya, meskipun Klaster 2 memiliki kapasitas baterai dan jarak tempuh yang lebih besar, rata-rata harga Klaster 1 (\$94.585) justru lebih tinggi daripada Klaster 2 (\$83.016). Pola ini kemungkinan disebabkan oleh kecepatan pengisian daya yang menjadi proksi tidak langsung tingkat trim atau kelas kendaraan, karena teknologi pengisian cepat umumnya diterapkan pada varian premium. Namun, interpretasi ini masih bersifat tentatif karena penelitian ini tidak memiliki data merek maupun trim kendaraan.

Temuan ini menunjukkan bahwa segmentasi kendaraan listrik berpotensi dimanfaatkan sebagai dasar penyusunan strategi pemasaran, penentuan harga, dan pengembangan infrastruktur pengisian daya sesuai karakteristik masing-masing segmen. Pemisahan dan persebaran observasi pada ketiga segmen kendaraan tersebut dapat diamati secara lebih representatif melalui visualisasi dua dimensi yang disajikan pada Gambar 2.



Gambar 2. Visualisasi Klaster Harga Kendaraan Listrik

3.1 Model Regresi Linier Berganda Sebelum Penambahan Klaster

Model regresi sebelum penambahan variabel klaster mengacu pada persamaan (1) pada Bagian 2. Model ini digunakan sebagai model dasar (*baseline*) untuk mengetahui performa prediksi sebelum penambahan hasil *clustering*. Hasil estimasi model regresi dasar (sebelum penambahan klaster) disajikan pada Tabel 2.

Tabel 2. Hasil Regresi Linier Berganda Sebelum Penambahan Klaster

Variabel	Koefisien	p-value	Keterangan
Year	1.526,7	0,00276	Signifikan
PC1 (Komponen Utama)	-2.969,3	$8,115 \times 10^{-8}$	Signifikan
Charging Speed	123,53	<0,001	Signifikan

Catatan: Nilai p-value dihitung menggunakan heteroscedasticity-consistent standard error (HC3).

3.2 Model Regresi Linier Berganda Setelah Penambahan Klaster

Setelah proses *K-Means Clustering* dilakukan, variabel klaster ditambahkan ke dalam model regresi sebagai variabel kategorik untuk meningkatkan kemampuan model dalam mengenali segmentasi kendaraan listrik.

Model regresi setelah penambahan klaster dituliskan sebagai berikut:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 Z_1 + \beta_3 X_2 + \beta_4 C_2 + \beta_5 C_3 + \varepsilon \quad (5)$$

Keterangan:

C_2 = dummy klaster 2

C_3 = dummy klaster 3

Variabel klaster digunakan untuk merepresentasikan kelompok kendaraan listrik hasil proses *clustering* berdasarkan kemiripan karakteristik kendaraan.

Tabel 3. Hasil Regresi Linier Berganda Setelah Penambahan Klaster

Variabel	Koefisien	p-value	Keterangan
<i>Year</i>	1.487,3	0,00339	Signifikan
<i>PC1 (Komponen Utama)</i>	-3.994,2	$1,261 \times 10^{-5}$	Signifikan
<i>Charging Speed</i>	97,50	$1,631 \times 10^{-9}$	Signifikan
<i>Klaster 2</i>	-8.435,2	0,0293	Signifikan
<i>Klaster 3</i>	-7.001,7	0,0581	Tidak signifikan

Catatan: Nilai p-value dihitung menggunakan heteroscedasticity-consistent standard error (HC3).

Berdasarkan Tabel 3, diperoleh koefisien regresi dan tingkat signifikansi masing-masing variabel prediktor terhadap harga kendaraan listrik menggunakan robust standard error (HC3), mengingat model regresi ini masih menunjukkan indikasi heteroskedastisitas (Bagian 3.3). Hasil regresi menunjukkan bahwa tahun kendaraan dan skor komponen utama (PC1, gabungan kapasitas baterai dan jarak tempuh) berpengaruh signifikan terhadap harga kendaraan listrik. Koefisien PC1 bernilai negatif ($\beta = -3.994,2$; $p = 1,261 \times 10^{-5}$) karena arah komponen utama hasil PCA berlawanan dengan arah kapasitas baterai dan jarak tempuh asli (lihat nilai loading pada Bagian 2.4); secara substansi, hasil ini tetap menunjukkan bahwa kendaraan dengan kapasitas baterai dan jarak tempuh yang lebih besar berasosiasi dengan harga yang lebih tinggi. Variabel kecepatan pengisian daya juga berpengaruh positif dan signifikan terhadap harga ($\beta = 97,50$; $p = 1,631 \times 10^{-9}$), dengan arah yang konsisten dengan model tanpa klaster ($\beta = 123,53$). Konsistensi arah koefisien ini penting dicatat: karena variabel harga tidak dilibatkan dalam pembentukan klaster pada penelitian ini, penambahan variabel klaster tidak menimbulkan perubahan arah hubungan antara kecepatan pengisian daya dan harga, berbeda dengan kondisi yang dapat terjadi apabila variabel target ikut digunakan dalam proses clustering.

Kedua variabel dummy klaster (Klaster 2 dan Klaster 3, dengan Klaster 1 sebagai kategori referensi) bernilai negatif, namun hanya Klaster 2 yang signifikan secara statistik setelah dikoreksi dengan robust standard error ($p = 0,0293$), sedangkan Klaster 3 tidak lagi signifikan pada taraf $\alpha = 0,05$ ($p = 0,0581$). Temuan ini berbeda dari hasil awal menggunakan standard error OLS biasa, yang sebelumnya menunjukkan kedua variabel klaster signifikan. Setelah standard error dikoreksi terhadap heteroskedastisitas, bukti bahwa Klaster 3 berbeda secara meyakinkan dari Klaster 1 menjadi lemah, sehingga kontribusi variabel klaster terhadap model regresi perlu dimaknai secara lebih hati-hati. Hal ini memperkuat indikasi bahwa pengelompokan berdasarkan spesifikasi kendaraan (tanpa melibatkan harga) hanya menangkap sebagian kecil pola yang relevan dengan variasi harga, konsisten dengan hasil evaluasi RMSE, MAE, dan R^2 pada subbab 3.4 yang menunjukkan penambahan variabel klaster tidak memberikan perbaikan performa prediksi yang berarti.

Perlu dicermati bahwa permasalahan multikolinearitas tinggi antara kapasitas baterai dan jarak tempuh yang teridentifikasi pada model awal (VIF masing-masing sebesar 23,13 dan 23,68) telah diatasi melalui transformasi PCA pada Bagian 2.4. Setelah kedua variabel digabungkan menjadi PC1, nilai VIF turun signifikan menjadi 2,82 (Tabel 4), sehingga estimasi koefisien menjadi lebih stabil dan tidak lagi bias akibat redundansi informasi antarvariabel. Meskipun demikian, PCA hanya mengatasi permasalahan multikolinearitas dan tidak menyelesaikan permasalahan heteroskedastisitas maupun non-normalitas residual yang juga terdeteksi pada model (Bagian 3.3);

oleh karena itu, penggunaan robust standard error tetap diperlukan agar inferensi terhadap signifikansi koefisien regresi tetap valid.

3.3 Uji Asumsi Klasik

Pengujian asumsi klasik dilakukan untuk memastikan model regresi memenuhi asumsi dasar regresi linier. Pengujian yang dilakukan meliputi uji normalitas residual menggunakan Shapiro-Wilk Test, uji multikolinearitas menggunakan Variance Inflation Factor (VIF), dan uji heteroskedastisitas menggunakan Breusch-Pagan Test. Tabel 4 menyajikan hasil uji normalitas residual menggunakan Shapiro-Wilk Test.

Tabel 4. Hasil Uji Asumsi Klasik

Pengujian	Metode	Hasil	Kriteria	Kesimpulan
Normalitas Residual	Shapiro-Wilk Test	$p\text{-value} < 2,2 \times 10^{-16}$	$p\text{-value} > 0,05$	Residual tidak berdistribusi normal
Multikolinearitas	Variance Inflation Factor (VIF)	$Year = 1,00$ $PC1$ (Komponen Utama) = 2,82 $Charging\ Speed = 2,62$ $Klaster\ 2 = 4,93$; $Klaster\ 3 = 4,88$	$VIF < 10$ (seluruh variabel)	Tidak terdapat multikolinearitas berarti (VIF seluruh variabel < 10 setelah PCA)
Heteroskedastisitas	Breusch-Pagan Test	$p\text{-value}$ (tanpa klaster) = $8,26 \times 10^{-15}$; $p\text{-value}$ (dengan klaster) = $8,649 \times 10^{-16}$	$p\text{-value} > 0,05$	Terdapat heteroskedastisitas

Berdasarkan Tabel 4, diperoleh nilai statistik $W = 0,974$ dengan $p\text{-value} < 2,2 \times 10^{-16}$, sehingga residual model belum berdistribusi normal secara sempurna. Kondisi ini umum dijumpai pada ukuran sampel yang besar (2.000 observasi), karena uji normalitas menjadi sangat sensitif terhadap penyimpangan kecil dari distribusi normal.

Hasil uji multikolinearitas menunjukkan bahwa seluruh variabel prediktor memiliki nilai VIF di bawah ambang batas 10. Pada model tanpa klaster, nilai VIF tertinggi sebesar 2,82 terdapat pada PC1 (komponen utama hasil PCA dari kapasitas baterai dan jarak tempuh), sedangkan pada model dengan klaster, nilai VIF tertinggi sebesar 4,93 terdapat pada variabel dummy Klaster 2. Hasil ini menunjukkan bahwa transformasi PCA berhasil menghilangkan multikolinearitas tinggi antara kapasitas baterai dan jarak tempuh yang sebelumnya memiliki nilai VIF masing-masing sebesar 23,13 dan 23,68, sehingga estimasi koefisien regresi menjadi lebih stabil.

Selanjutnya, hasil uji heteroskedastisitas menggunakan Breusch-Pagan Test menunjukkan $p\text{-value}$ sebesar $8,26 \times 10^{-15}$ pada model tanpa klaster dan $8,649 \times 10^{-16}$ pada model dengan klaster. Kedua nilai tersebut lebih kecil dari 0,05 sehingga mengindikasikan bahwa model masih mengalami heteroskedastisitas. Hasil ini menegaskan bahwa PCA hanya mengatasi permasalahan multikolinearitas dan tidak memengaruhi heteroskedastisitas karena kedua permasalahan tersebut memiliki penyebab yang berbeda.

Oleh karena itu, seluruh pengujian signifikansi koefisien pada Tabel 2 dan Tabel 3 menggunakan heteroscedasticity-consistent standard error (robust standard error) tipe HC3 agar inferensi terhadap signifikansi koefisien tetap valid meskipun asumsi homoskedastisitas tidak terpenuhi. Sementara itu, pelanggaran asumsi normalitas residual tidak dikoreksi lebih lanjut karena estimasi regresi linier berganda tetap bersifat robust pada ukuran sampel besar, sejalan dengan Central Limit Theorem.

3.4 Evaluasi Model

Hasil evaluasi model menunjukkan nilai RMSE, MAE, dan R^2 pada Tabel 5 merupakan rata-rata hasil 10-fold cross-validation. Standar deviasi RMSE antar-fold adalah 2.088,01 (tanpa klaster)

dan 2.119,08 (dengan klaster); standar deviasi R^2 antar-fold adalah 0,0510 (tanpa klaster) dan 0,0530 (dengan klaster).

Tabel 5. Perbandingan Hasil Evaluasi Model

Metrik	Tanpa klaster	Dengan klaster
RMSE	33.353,66	33.357,02
MAE	26.200,81	26.140,91
R^2	0,1106	0,1109

Berdasarkan Tabel 5, penambahan variabel klaster tidak memberikan perbaikan performa yang meyakinkan. Nilai RMSE bahkan sedikit meningkat dari 33.353,66 menjadi 33.357,02 (selisih 3,36, atau sekitar 0,01%), sedangkan nilai MAE turun tipis dari 26.200,81 menjadi 26.140,91 (selisih 59,90, atau sekitar 0,23%). Selisih tersebut jauh lebih kecil dibandingkan standar deviasi RMSE dan MAE antar-fold pada *cross-validation*, sehingga tidak dapat dianggap sebagai perbaikan performa yang bermakna secara praktis, mengingat selisihnya jauh lebih kecil dibandingkan variasi antar-fold; bahkan untuk metrik RMSE, arah perubahannya justru berlawanan dengan yang diharapkan. Hal ini diduga karena variabel hasil *clustering* hanya merepresentasikan kemiripan karakteristik teknis kendaraan, sedangkan harga kendaraan listrik juga dipengaruhi oleh faktor lain seperti merek, fitur, reputasi produsen, serta kondisi pasar yang tidak tersedia dalam *dataset*.

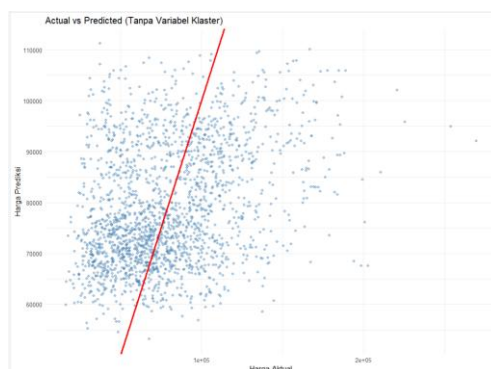
Selain itu, nilai koefisien determinasi (R^2) hanya meningkat dari 0,1106 menjadi 0,1109, yaitu sebesar 0,0003 yang secara praktis tidak berarti. Temuan ini penting untuk digarisbawahi: pada tahap eksplorasi awal penelitian ini, ketika variabel harga turut dilibatkan dalam pembentukan klaster, model dengan variabel klaster menghasilkan peningkatan R^2 yang jauh lebih besar (dari 0,1106 menjadi 0,2791 pada evaluasi *cross-validation*). Perbedaan yang sangat mencolok tersebut konsisten dengan kemungkinan terjadinya kebocoran informasi target (harga) ke dalam variabel klaster pada skenario pertama, bukan mencerminkan kemampuan *clustering* dalam mengungkap pola tersembunyi pada data. Dengan kata lain, ketika klaster dibentuk secara metodologis benar, yaitu hanya dari variabel yang tersedia sebelum harga diketahui, kontribusinya terhadap akurasi prediksi harga kendaraan listrik pada *dataset* ini terbukti sangat terbatas. Temuan ini menunjukkan bahwa *feature engineering* berbasis *clustering* tidak selalu meningkatkan performa model regresi, tetapi bergantung pada karakteristik data serta informasi baru yang mampu diberikan oleh variabel hasil *clustering*.

Meskipun variabel klaster menunjukkan signifikansi statistik pada model regresi (Bagian 3.2), peningkatan kemampuan prediksi model tetap sangat kecil berdasarkan hasil evaluasi RMSE, MAE, dan R^2 pada *cross-validation*. Hal ini mengindikasikan bahwa informasi yang dibawa oleh variabel klaster sebagian besar telah tercermin dalam variabel teknis yang digunakan pada model regresi, sehingga signifikansi statistik tidak selalu diikuti oleh peningkatan kemampuan prediksi yang bermakna secara praktis. Rendahnya nilai R^2 juga menunjukkan bahwa variabel teknis yang digunakan dalam penelitian ini hanya mampu menjelaskan sebagian kecil variasi harga kendaraan listrik. Kondisi ini kemungkinan disebabkan oleh tidak tersedianya variabel penting lain, seperti merek dan reputasi produsen, tipe kendaraan, fitur teknologi, kondisi kendaraan, serta faktor pasar dan preferensi konsumen.

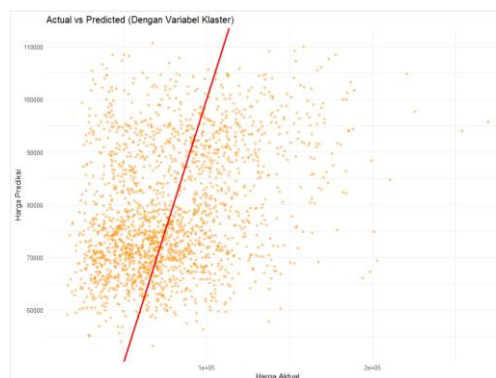
Secara keseluruhan, hasil evaluasi menunjukkan bahwa penambahan variabel klaster hasil K-Means Clustering belum mampu meningkatkan performa model regresi linier berganda secara substantif. Dubey et al. [3] melaporkan bahwa model Random Forest memberikan akurasi prediksi harga EV yang jauh lebih tinggi dibandingkan regresi linier, sedangkan Simsek et al. [12] memperoleh nilai R^2 di atas 90% pada peramalan permintaan EV menggunakan model hybrid deep learning berbasis LSTM-CNN. Perbedaan tersebut kemungkinan disebabkan oleh dua faktor. Pertama, regresi linier berganda merupakan model linier yang terbatas dalam menangkap hubungan non-linier antara spesifikasi teknis dan harga kendaraan, sedangkan Random Forest dan metode *deep learning* lebih mampu memodelkan hubungan tersebut. Kedua, variabel yang tersedia

dalam dataset penelitian ini terbatas pada aspek teknis kendaraan, tanpa mencakup variabel merek, fitur teknologi, maupun faktor pasar yang terbukti berpengaruh signifikan pada penelitian Dubey et al. [3]. Berbeda dengan temuan Zhang dan Kang [18] yang melaporkan bahwa segmentasi berbasis karakteristik pengguna maupun kendaraan mampu mengungkap pola yang tidak tertangkap oleh variabel numerik saja, penelitian ini menunjukkan bahwa strategi feature engineering berbasis clustering tidak selalu memberikan kontribusi praktis yang berarti terhadap performa model regresi, terutama ketika kluster dibentuk secara metodologis benar tanpa melibatkan variabel target. Temuan ini menjadi catatan penting bahwa peningkatan performa model pada pendekatan *clustering-regresi* perlu dievaluasi secara cermat untuk memastikan tidak terjadi kebocoran informasi target.

Perbandingan kemampuan prediksi model sebelum dan sesudah penambahan variabel kluster juga dapat diamati secara visual melalui grafik *Actual vs Predicted* pada Gambar 3 dan Gambar 4. Semakin rapat sebaran titik terhadap garis diagonal merah, semakin baik kemampuan prediksi model.



Gambar 3. Grafik Actual vs Predicted Tanpa Variabel Kluster



Gambar 4. Grafik Actual vs Predicted Dengan Variabel Kluster

Gambar 3 menunjukkan sebaran titik yang acak dan jauh dari garis diagonal, mengindikasikan kemampuan prediksi model tanpa kluster yang masih terbatas. Gambar 4 menunjukkan pola sebaran titik yang serupa dan tidak jauh berbeda dari Gambar 3, konsisten dengan nilai R^2 yang hampir tidak berubah setelah penambahan variabel kluster (dari 0,1106 menjadi 0,1109). Temuan ini menegaskan bahwa penambahan informasi kluster yang dibentuk hanya berdasarkan variabel teknis tanpa melibatkan harga tidak memberikan perbaikan yang berarti terhadap kemampuan prediksi model.

Berdasarkan seluruh hasil analisis, metode K-Means Clustering berhasil mengelompokkan kendaraan listrik ke dalam tiga segmen berdasarkan karakteristik teknisnya. Namun, penambahan variabel kluster ke dalam model regresi linier berganda belum mampu meningkatkan performa prediksi secara berarti dibandingkan model tanpa kluster. Hal ini menunjukkan bahwa informasi

yang dibawa oleh variabel klaster sebagian besar telah tercermin dalam variabel teknis yang digunakan pada model regresi.

4. KESIMPULAN

Penelitian ini berhasil mencapai tujuan yang telah ditetapkan, yaitu menganalisis pengelompokan kendaraan listrik menggunakan metode K-Means Clustering, membangun model prediksi harga kendaraan listrik menggunakan regresi linier berganda, serta mengevaluasi kontribusi hasil *clustering* sebagai variabel tambahan terhadap performa model prediksi. Hasil penelitian menunjukkan bahwa proses *clustering* yang dibentuk hanya dari variabel teknis (tahun produksi, kapasitas baterai, jarak tempuh, dan kecepatan pengisian daya) tanpa melibatkan variabel harga berhasil mengelompokkan kendaraan listrik ke dalam tiga segmen dengan karakteristik teknis yang berbeda. Namun, penambahan variabel klaster pada model regresi hanya memberikan peningkatan performa yang sangat kecil dan secara praktis tidak berarti, ditunjukkan oleh perubahan RMSE dan MAE yang minimal serta kenaikan nilai R^2 dari 0,1106 menjadi 0,1109. Hasil tersebut menunjukkan bahwa informasi yang dibawa oleh variabel klaster sebagian besar telah tercermin dalam variabel teknis yang digunakan pada model regresi.

Penelitian ini juga memberikan kontribusi metodologis dengan menunjukkan bahwa peningkatan performa model yang tampak besar dapat terjadi apabila variabel target dilibatkan dalam proses pembentukan klaster, sehingga berpotensi menimbulkan kebocoran informasi target (*target leakage*). Oleh karena itu, pembentukan fitur (*feature engineering*) berbasis *clustering* perlu dilakukan tanpa melibatkan variabel target agar evaluasi performa model tetap valid dan dapat diinterpretasikan secara tepat.

Penelitian ini masih memiliki keterbatasan karena hanya menggunakan variabel teknis kendaraan dan belum mempertimbangkan faktor lain, seperti merek kendaraan, fitur teknologi, kondisi pasar, dan preferensi konsumen. Penelitian selanjutnya diharapkan dapat menambahkan variabel-variabel tersebut serta mengeksplorasi metode *machine learning* yang mampu menangkap hubungan nonlinier, seperti Random Forest dan Gradient Boosting, dengan tetap menghindari pelibatan variabel target dalam proses pembentukan fitur prediktor. Pengembangan tersebut diharapkan dapat menghasilkan model prediksi harga kendaraan listrik yang lebih akurat sekaligus memperluas penerapan integrasi *clustering* dan regresi secara metodologis benar.

DAFTAR PUSTAKA

- [1] S. Afandizadeh, D. Sharifi, N. Kalantari, and H. Mirzahassein, "Using machine learning methods to predict electric vehicles penetration in the automotive market," *Scientific Reports*, vol. 13, p. 8345, 2023, <https://doi.org/10.1038/s41598-023-35366-3>.
- [2] S. J. Choi and J. Jiao, "Uncovering electric vehicle ownership disparities using K-means clustering analysis: A case study of Austin, Texas," *Journal of Computational Social Science*, vol. 7, pp. 2403–2456, 2024, <https://doi.org/10.1007/s42001-024-00310-6>.
- [3] S. Dubey, A. Renavikar, and S. Kanungo, "Machine learning model for prediction of electric vehicle prices," *Journal of Information Systems Engineering and Management*, vol. 10, no. 50s, pp. 1006–1017, 2025, <https://doi.org/10.52783/jisem.v10i50s.10456>.
- [4] Y. Gan, M. Wang, Z. Lu, and J. Kelly, "Taking into account greenhouse gas emissions of electric vehicles for transportation decarbonization," *Energy Policy*, vol. 155, p. 112353, 2021, <https://doi.org/10.1016/j.enpol.2021.112353>.
- [5] M. S. Hossain, L. Kumar, M. M. Islam, and J. Selvaraj, "A comprehensive review on the integration of electric vehicles for sustainable development," *Journal of Advanced Transportation*, vol. 2022, p. 3868388, 2022, <https://doi.org/10.1155/2022/3868388>.
- [6] I. Hussain, K. B. Ching, C. Uttraphan, K. G. Tay, and A. Noor, "Evaluating machine learning algorithms for energy consumption prediction in electric vehicles: A comparative study," *Scientific Reports*, vol. 15, no. 1, pp. 1–20, 2025, <https://doi.org/10.1038/s41598-025-94946-7>.
- [7] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, 2023, <https://doi.org/10.1016/j.ins.2022.11.139>.

-
- [8] A. Khan, N. Iqbal, Z. Kaleem, Z. Qarnain, and M. M. Bait-Suwailam, "A multi-model approach for predicting electric vehicle specifications and energy consumption using machine learning," *The Journal of Supercomputing*, vol. 81, p. 314, 2025, <https://doi.org/10.1007/s11227-024-06820-4>.
- [9] R. D. Mishra, S. K. Dash, S. Chudjuarjeen, S. Mohanty, S. P. Gautam, R. I. Mohanty, and M. T. Tanvir, "Hybrid LSTM-GBRT based machine learning technique implementation for electric vehicle sales prediction analysis," *International Journal of Sustainable Engineering*, vol. 17, no. 1, pp. 843–858, 2024, <https://doi.org/10.1080/19397038.2024.2411284>.
- [10] S. Pokharel, P. Sah, and D. Ganta, "Improved prediction of total energy consumption and feature analysis in electric vehicles using machine learning and Shapley additive explanations method," *World Electric Vehicle Journal*, vol. 12, no. 3, p. 94, 2021, <https://doi.org/10.3390/wevj12030094>.
- [11] J. A. Sanguesa, V. Torres-Sanz, P. Garrido, F. J. Martinez, and J. M. Marquez-Barja, "A review on electric vehicles: Technologies and challenges," *Smart Cities*, vol. 4, no. 1, pp. 372–404, 2021, <https://doi.org/10.3390/smartcities4010022>.
- [12] A. I. Simsek, E. Koç, B. Desticioglu Tasdemir, A. Aksöz, M. Turkoglu, and A. Sengur, "Deep learning forecasting model for market demand of electric vehicles," *Applied Sciences*, vol. 14, no. 23, p. 10974, 2024, <https://doi.org/10.3390/app142310974>.
- [13] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, <https://doi.org/10.1109/ACCESS.2020.2988796>.
- [14] H. Sonar, H. M. Belal, C. Foropon, R. Manatkar, and V. Sonwaney, "Examining the causal factors of the electric vehicle adoption: A pathway to tackle climate change in resource-constrained environment," *Annals of Operations Research*, vol. 355, no. 1, pp. 59–85, 2023, <https://doi.org/10.1007/s10479-023-05519-8>.
- [15] I. Ullah, K. Liu, T. Yamamoto, R. E. Al Mamlook, and A. Jamal, "A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability," *Energy & Environment*, vol. 33, no. 8, pp. 1583–1612, 2022, <https://doi.org/10.1177/0958305X211044998>.
- [16] J. R. R. Uy, A. K. S. Ong, D. M. B. De Guzman, I. T. Dela Cruz, and J. C. Dela Cruz, "Consumer segmentation and market analysis for sustainable marketing strategy of electric vehicles in the Philippines," *World Electric Vehicle Journal*, vol. 15, no. 7, p. 301, 2024, <https://doi.org/10.3390/wevj15070301>.
- [17] C. Wongoutong, "The impact of neglecting feature scaling in k-means clustering," *PLOS ONE*, vol. 19, no. 12, p. e0310839, 2024, <https://doi.org/10.1371/journal.pone.0310839>.
- [18] B. Zhang and J. Kang, "Profiling user segments and spatial clusters of EV uptake through multi-method modeling," *Sustainable Environment*, vol. 11, no. 1, p. 2544396, 2025, <https://doi.org/10.1080/27658511.2025.2544396>.
-