ROUTERS: Jurnal Sistem dan Teknologi Informasi
Vol. 3, No. 2, Juli 2025, pp. 94-103
ISSN: 2962-1224, DOI: 10.25181/rt.v3i2.4249

Prediksi Harga Mobil Bekas Menggunakan Klasterisasi K-Means dan Perbandingan Empat Algoritma Regresi Supervised Learning

Gilang Rizki Ramadhan [1], Amanda Bulan Nayla [2], Syahreza Riatma[3]
[1, 2, 3] Politeknik Negeri Lampung, Teknologi Rekayasa Internet, Bandar Lampung

INFORMASI ARTIKEL

Diterima 4 Juli 2025
Direvisi 10 Juli 2025
Diterbitkan 11 Juli 2025

Kata kunci:

Machine Learning;
Segmentasi Harga;
Regresi;
K-Nearest Neighbors;
Random Forest Regressor;
Polynomial Regression;
K-Means Clustering

ABSTRAK

Prediksi harga mobil bekas menjadi tantangan penting dalam industri otomotif seiring meningkatnya kebutuhan akan sistem valuasi yang akurat dan objektif. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan kinerja empat algoritma regresi yaitu Multiple Linear Regression (MLR), Polynomial Regression (PR) derajat 2, K-Nearest Neighbors Regression (KNNR), dan Random Forest Regressor (RFR) dalam memprediksi harga mobil bekas berbasis segmentasi harga menggunakan K-Means Clustering. Dataset yang digunakan merupakan *Used Cars Dataset* dari India yang berisi 14.993 entri dan 11 fitur kendaraan.Segmentasi dilakukan menggunakan algoritma K-Means dengan pendekatan Elbow Method dan Silhouette Score, yang menunjukkan hasil optimal pada k = 2 dengan nilai Silhouette Score sebesar 0,7170. Klaster pertama mencakup mobil dengan rentang harga antara 15.000 hingga 1.183.000 dengan rata-rata sebesar 514.411, sedangkan klaster kedua mencakup harga antara 1.185.000 hingga 2.271.000 dengan rata-rata sebesar 1.855.157. Hasil segmentasi ini kemudian digunakan sebagai fitur tambahan dalam proses pelatihan model regresi.Evaluasi performa menggunakan MAE, RMSE, R^2, dan MAPE menunjukkan bahwa Random Forest Regressor memberikan hasil terbaik, dengan nilai MAE = 92.439,34, RMSE = 146.838,26, R^2 = 0,9477, dan MAPE = 18,25%, yang mengindikasikan tingkat akurasi prediksi yang tinggi. Integrasi teknik segmentasi harga dan model regresi non-linear terbukti meningkatkan performa prediksi secara signifikan. Hasil penelitian ini dapat menjadi landasan dalam pengembangan sistem rekomendasi harga kendaraan bekas berbasis data yang lebih adaptif dan presisi.

Used Car Price Prediction Using K-Means Clustering and Comparison of Four Supervised Learning Regression Algorithms

ARTICLE INFO

Received July 4, 2025
Revised July 10, 2025
Published July 11, 2025

Keyword:

Machine Learning;
Price Segmentation;
Regression;
K-Nearest Neighbors;
Random Forest Regressors;
Polynomial Regression;
K-Means Clustering

ABSTRACT

The Used car price prediction is a crucial challenge in the automotive industry as the need for an accurate and objective valuation system increases. This study aims to implement and compare the performance of four regression algorithms, namely Multiple Linear Regression (MLR), Polynomial Regression (PR) degree 2, K-Nearest Neighbors Regression (KNNR), and Random Forest Regressor (RFR) in predicting used car prices based on price segmentation using K-Means Clustering. The dataset used is the Used Cars Dataset from India containing 14,993 entries and 11 vehicle features. Segmentation is performed using the K-Means algorithm with the Elbow Method and Silhouette Score approaches, which show optimal results at k = 2 with a Silhouette Score value of 0.7170. The first cluster includes cars with a price range between 15,000 and 1,183,000 with an average of 514,411, while the second cluster includes prices between 1,185,000 and 2,271,000 with an average of 1,855,157. The results of this segmentation are then used as additional features in the regression model training process. Performance evaluation using MAE, RMSE, R^2, and MAPE shows that Random Forest Regressor provides the best results, with MAE = 92,439.34, RMSE = 146,838.26, R^2 = 0.9477, and MAPE = 18.25%, which indicates a high level of prediction. The integration of price segmentation techniques and non-linear regression models is proven to significantly improve prediction performance. The results of this study can be a basis for developing a more adaptive and precise data-based used vehicle price recommendation system.

Corresponding Author:

Gilang Rizki Ramadhan, Teknologi Rekayasa Internet Politeknik Negeri Lampung
Email: gilangrizkiramadhan1906@gmail.com

1. PENDAHULUAN

Perkembangan teknologi analitik prediktif dalam bidang data science telah mendorong pemanfaatan algoritma regresi secara luas di berbagai sektor, termasuk industri otomotif. Salah satu aplikasi signifikan dari algoritma ini adalah dalam pemodelan harga kendaraan bekas, yang merupakan aspek penting dalam sistem informasi penjualan dan pengambilan keputusan konsumen. Harga kendaraan bekas dipengaruhi oleh beragam atribut, seperti usia kendaraan, jarak tempuh, merek, jenis transmisi, kondisi fisik, dan preferensi pasar.

Algoritma regresi menyediakan pendekatan kuantitatif untuk mengestimasi nilai berdasarkan relasi antara variabel-variabel input dan output. Regresi linear, sebagai salah satu metode dasar, telah menunjukkan efektivitas dalam memodelkan hubungan linier multivariat. Studi sebelumnya menunjukkan bahwa regresi linear dapat digunakan untuk memprediksi biaya asuransi berdasarkan fitur demografis seperti usia dan status merokok, dengan tingkat akurasi yang signifikan [1].

Dalam konteks prediksi harga kendaraan, penelitian menunjukkan bahwa penggunaan teknik *hyperparameter tuning* pada regresi linear mampu meningkatkan akurasi prediksi hingga mencapai 80% [2]. Di sisi lain, algoritma non-regresif seperti C4.5 juga telah digunakan dalam prediksi harga

kendaraan bekas, menghasilkan tingkat akurasi yang sangat tinggi, yaitu mencapai 99% pada skenario tertentu [3]. Meskipun algoritma yang digunakan berbeda, temuan-temuan tersebut menunjukkan bahwa pemilihan model dan parameter sangat berpengaruh terhadap performa sistem prediksi.

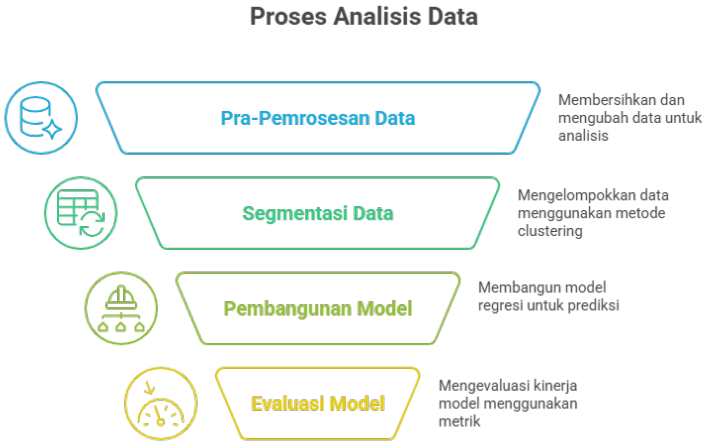
Struktur pasar kendaraan bekas yang terdiri dari beberapa segmen harga turut memengaruhi performa model prediksi. Penelitian terdahulu mengindikasikan bahwa variasi antarsegmen dapat memengaruhi akurasi model, sehingga diperlukan pendekatan yang mempertimbangkan karakteristik data per segmen [4]. Studi komparatif terhadap beberapa algoritma prediktif juga telah dilakukan dalam berbagai konteks, yang menunjukkan bahwa efektivitas model sangat bergantung pada kesesuaian algoritma terhadap struktur data yang digunakan [5].

Selain akurasi, kinerja algoritma regresi juga dipengaruhi oleh efisiensi komputasi, pemilihan fitur yang relevan, serta strategi praproses data. Oleh karena itu, evaluasi terhadap berbagai algoritma regresi perlu dilakukan secara sistematis untuk mengidentifikasi model yang optimal berdasarkan kriteria tertentu, terutama dalam konteks prediksi harga kendaraan bekas dengan mempertimbangkan segmentasi harga.

Penelitian ini bertujuan untuk melakukan analisis komparatif terhadap beberapa algoritma regresi, termasuk regresi linear, regresi berganda, dan algoritma machine learning lainnya, dalam memprediksi harga kendaraan bekas berdasarkan segmentasi harga. Evaluasi dilakukan terhadap aspek akurasi prediksi, waktu komputasi, serta sensitivitas terhadap pemilihan fitur. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan sistem pendukung keputusan berbasis data dalam sektor penjualan kendaraan bekas.

2. METODE

Penelitian ini mengadopsi pendekatan kombinatif antara *unsupervised learning* dan *supervised learning* untuk membangun model prediksi harga mobil bekas berbasis segmentasi. Dimulai dari proses pengumpulan dan praproses data, dilanjutkan dengan segmentasi data menggunakan algoritma K-Means Clustering. Setelah itu, dilakukan pembangunan model regresi menggunakan beberapa algoritma *supervised learning*, kemudian dievaluasi kinerjanya menggunakan metrik tertentu, dan diakhiri dengan analisis hasil serta interpretasi terhadap temuan yang diperoleh. Tahapan metodologi disusun secara sistematis sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Tahapan Proses Analisis Data

2.1. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berjudul **"Used Cars Dataset"** yang diperoleh dari platform Kaggle melalui tautan https://www.kaggle.com/datasets/mohitkumar282/used-car-dataset. Dataset ini merepresentasikan kondisi pasar mobil bekas di India dan terdiri dari 14.993 entri dengan 11 fitur utama yang mencakup aspek teknis maupun historis kendaraan, seperti merek, model, tahun pembuatan, usia kendaraan, jarak tempuh (*km driven*), jenis transmisi, status

kepemilikan, jenis bahan bakar, tanggal unggah ke platform, informasi tambahan mengenai kendaraan, serta harga jual yang diminta. Data dikumpulkan hingga bulan November 2024, mencerminkan tren terkini dalam penjualan mobil bekas di wilayah tersebut.

Secara umum, dataset ini memiliki karakteristik sebagai berikut: total entri sebanyak 14.993 data dengan 11 atribut yang terdiri dari kombinasi tipe data numerik dan kategorikal, serta ukuran file sekitar ±1,94 MB. Informasi ini menjadi fondasi penting dalam proses eksplorasi, segmentasi, dan pemodelan prediktif terhadap harga mobil bekas yang dilakukan dalam penelitian ini.

2.2. Pra-Pemrosesan Data

Pra-pemrosesan data dalam penelitian ini dilakukan melalui dua tahap utama, yaitu *data understanding* dan *data preprocessing*. Tahap *data understanding* diawali dengan memuat data dan memastikan tidak terdapat kesalahan atau error dalam proses pembacaan. Selanjutnya, dilakukan normalisasi nama kolom agar konsisten (menggunakan huruf kecil dan tanpa spasi berlebih), serta pembersihan nilai pada kolom kmDriven dan AskPrice dengan menghapus simbol atau satuan yang tidak diperlukan. Kolom PostedDate dikonversi ke dalam format waktu yang valid untuk kemudian dihitung selisih harinya sebagai fitur baru DaysSincePosted. Analisis statistik deskriptif juga dilakukan terhadap kolom numerik untuk memperoleh informasi seperti nilai rata-rata, median, minimum, maksimum, dan standar deviasi [6]. Selain itu, statistik untuk fitur kategorikal juga dianalisis, termasuk jumlah kategori unik dan nilai modus [7]. Konten pada kolom teks seperti AdditionalInfo turut dianalisis untuk memahami informasi tambahan yang mungkin relevan, serta dilakukan identifikasi dan dokumentasi terhadap nilai-nilai yang hilang.

Tahap selanjutnya adalah *data preprocessing* yang berfokus pada transformasi teknis terhadap data agar siap digunakan dalam proses analisis dan pemodelan. Nilai pada kolom kmDriven dan AskPrice dikonversi menjadi data numerik agar dapat diolah secara kuantitatif, serta dilakukan transformasi waktu pada PostedDate untuk menghasilkan fitur DaysSincePosted. Penanganan *missing values* dilakukan menggunakan metode imputasi rata-rata (*mean imputation*), sementara *outlier* pada fitur Age, kmDriven, dan AskPrice ditangani dengan teknik *capping* [8]. Untuk fitur kategorikal seperti Brand, Model, Transmission, Owner, dan FuelType, diterapkan metode *target encoding* guna mengkonversinya ke bentuk numerik berbasis target. Seluruh fitur numerik kemudian dinormalisasi menggunakan metode *standardization* berbasis Z-score [9]. Dataset akhir disimpan dalam format yang siap pakai untuk proses analisis dan pemodelan lebih lanjut.

2.3. Segmentasi Clustering

Tahapan segmentasi data dilakukan dengan memanfaatkan data hasil pra-pemrosesan dan seleksi fitur yang telah disiapkan sebelumnya. Seluruh fitur kategorikal dikonversi ke dalam bentuk numerik melalui teknik *one-hot encoding* untuk memastikan kompatibilitas dengan algoritma klasterisasi [10]. Selanjutnya, jumlah klaster yang optimal ditentukan menggunakan dua pendekatan utama, yaitu *Elbow Method* yang mengandalkan analisis nilai *inertia*, serta *Silhouette Score* yang mengevaluasi kualitas klaster berdasarkan tingkat kohesi dan separasi antar anggota klaster. Setelah jumlah klaster ditetapkan, algoritma K-Means digunakan untuk mengelompokkan data mobil bekas ke dalam segmen-segmen yang memiliki karakteristik serupa.

Untuk memahami struktur hasil klasterisasi secara visual, dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA) dan divisualisasikan dalam ruang dua dimensi. Setiap titik pada grafik visualisasi merepresentasikan satu entri mobil bekas, dengan warna yang menunjukkan klaster tempatnya tergolong. Setelah klasterisasi selesai, dilakukan analisis harga pada masing-masing klaster, mencakup nilai minimum, maksimum, rata-rata, serta jumlah entri dalam setiap segmen. Hasil akhir segmentasi kemudian disimpan dalam format file terstruktur untuk keperluan analisis lanjutan dan pembangunan model prediktif berbasis segmentasi tersebut.

2.4. Regresi Harga

Setelah proses segmentasi, tahap selanjutnya adalah membangun model prediktif harga mobil bekas menggunakan pendekatan *supervised learning* [11]. Empat algoritma regresi diterapkan dalam penelitian ini, yaitu Multiple Linear Regression (MLR), Polynomial Regression (PR) derajat 2, K-Nearest Neighbors Regression (KNNR), dan Random Forest Regressor (RFR). MLR digunakan untuk memodelkan hubungan linier antara variabel prediktor dan harga, sedangkan PR derajat 2

diaplikasikan untuk menangkap hubungan non-linier antara fitur dan variabel target. KNNR dipilih sebagai pendekatan berbasis instance yang mengestimasi harga berdasarkan kemiripan lokal antar data, sementara RFR digunakan sebagai metode ensemble berbasis pohon keputusan untuk menangani kompleksitas data dan interaksi antar fitur secara lebih fleksibel. Seluruh model dilatih menggunakan dataset hasil pra-pemrosesan yang telah ditambahkan atribut klaster sebagai variabel prediktor tambahan, sehingga informasi segmentasi harga turut dimanfaatkan dalam proses prediksi. Pembagian data dilakukan menggunakan skema *train-test split* dengan rasio 80:20 untuk memastikan adanya evaluasi performa yang objektif terhadap data uji yang belum pernah dilihat oleh model.

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil implementasi algoritma segmentasi dan regresi dalam memprediksi harga mobil bekas. Hasil ditampilkan dalam bentuk grafik dan visualisasi guna mempermudah pemahaman pembaca terhadap pola dan struktur data yang terbentuk. Selanjutnya, dilakukan pembahasan mendalam terhadap hasil tersebut serta perbandingannya dengan temuan dari penelitian sebelumnya untuk mendukung validitas dan kebaruan pendekatan yang diusulkan.

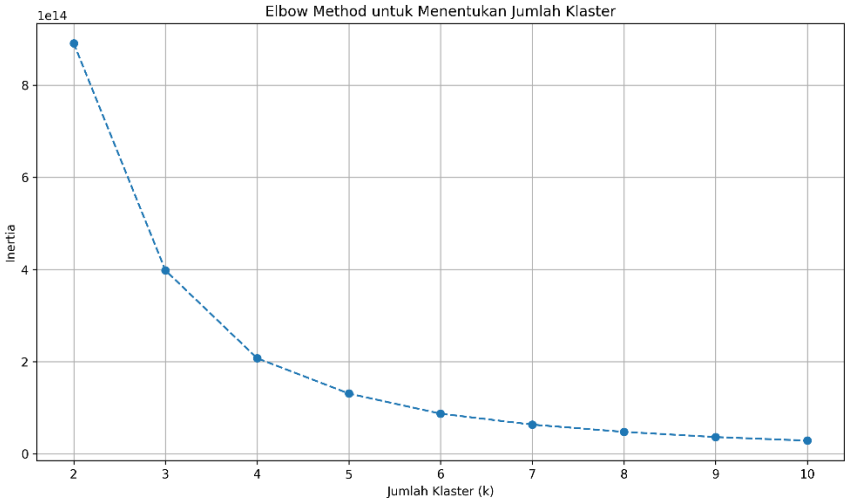
3.1. Penentuan Jumlah Klaster Optimal

Analisis segmentasi difokuskan pada atribut numerik, khususnya harga dan variabel kuantitatif lainnya, guna mengidentifikasi pola distribusi yang bermakna antar klaster. Untuk menentukan jumlah klaster yang optimal, digunakan dua metrik evaluasi utama, yaitu *inertia* dan *silhouette score*. Nilai *inertia* mengukur tingkat kekompakan data dalam masing-masing klaster, di mana nilai yang lebih rendah menunjukkan hasil segmentasi yang lebih baik. Sementara itu, *silhouette score* mengevaluasi sejauh mana objek berada pada klaster yang tepat, dengan nilai mendekati 1 menunjukkan pemisahan klaster yang optimal. Tabel 1 menyajikan nilai *inertia* dan *silhouette score* untuk berbagai jumlah klaster yang diuji dalam penelitian ini.

Tabel 1. Nilai Inertia dan Silhouette Score untuk Berbagai Jumlah Klaster K

K	Inertia	Silhouette Score
2	891,313,937,893,027.5	**0.7170**
3	398,232,054,949,807.0	0.6289
4	207,528,528,975,619.7	0.5957
5	130,795,105,182,154.7	0.5809
6	87,149,069,257,756.6	0.5793
7	63,385,954,807,172.0	0.5758
8	47,641,071,583,542.0	0.5833
9	36,377,777,641,814.7	0.5778
10	28,615,813,233,249.0	0.5803

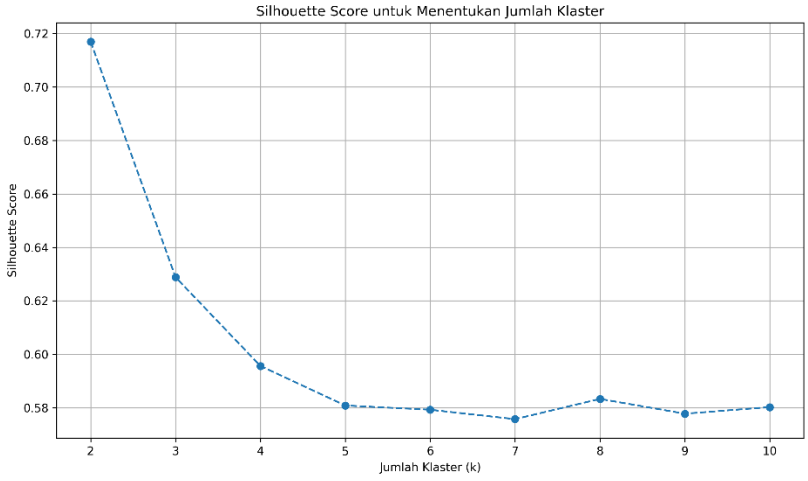
Penurunan inertia yang paling signifikan terjadi antara K=2 dan K=3, namun setelah itu penurunannya cenderung melandai. Hal ini mengindikasikan bahwa penambahan jumlah klaster tidak lagi memberikan manfaat signifikan terhadap kekompakan klaster. Di sisi lain, nilai Silhouette Score tertinggi dicapai pada K=2 (0.7170), yang berarti dua klaster memberikan pemisahan data yang paling optimal.Untuk mendukung analisis numerik pada Tabel 1, visualisasi Elbow Method pada Gambar 1 memperlihatkan hubungan antara jumlah klaster (k) dan nilai inertia.



Gambar 2. Grafik terhadap Jumlah Klaster Berdasarkan Elbow Method

Visualisasi Elbow Method pada Gambar 2 menunjukkan hubungan antara jumlah klaster (k) dan nilai inertia (Within-Cluster Sum of Squares/ WCSS). Grafik tersebut memperlihatkan penurunan inertia yang masih cukup signifikan hingga k = 4, kemudian melandai, yang mengindikasikan bahwa k = 4 dapat dianggap sebagai titik *elbow*, yakni titik optimal di mana penambahan klaster tidak lagi memberikan peningkatan berarti terhadap kekompakan data.

Namun demikian, evaluasi lebih lanjut menggunakan Silhouette Score sebagaimana ditampilkan pada Gambar 3 menunjukkan bahwa nilai tertinggi dicapai pada k = 2 dengan skor 0,7170. Hal ini menunjukkan bahwa dua klaster menghasilkan segmentasi paling optimal, baik dari segi kohesi internal maupun pemisahan antar klaster. Oleh karena itu, meskipun Elbow Method mengindikasikan k = 4 sebagai kandidat potensial, jumlah klaster optimal dalam penelitian ini ditetapkan pada k = 2 karena memberikan struktur klaster yang lebih representatif dan berkualitas tinggi terhadap distribusi data.



Gambar 3. Grafik Nilai Silhouette Score terhadap Jumlah Klaster

Berdasarkan hasil evaluasi numerik dan visualisasi, jumlah klaster optimal dalam penelitian ini ditetapkan sebanyak dua klaster. Keputusan ini menjadi dasar untuk analisis lanjutan terhadap karakteristik masing-masing klaster, guna mengidentifikasi perbedaan distribusi harga serta fitur numerik lainnya dalam dataset. Pemahaman terhadap struktur dan pola unik pada tiap klaster

diharapkan dapat memperkuat proses segmentasi dalam konteks prediksi harga dan pengelompokan mobil bekas secara lebih akurat.

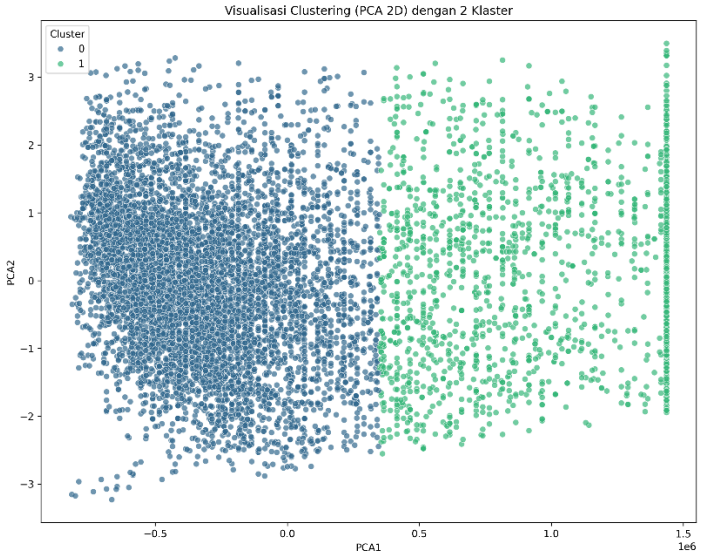
3.2. Analisis Segmentasi Harga Menggunakan K-Means Klastering

Setelah menetapkan jumlah klaster optimal (K=2), algoritma K-Means diterapkan untuk membagi data menjadi dua kelompok berdasarkan kemiripan fitur numerik. Hasil distribusi harga antar klaster ditunjukkan pada Tabel 2.

Tabel 2. Statistik Harga Mobil Bekas Berdasarkan Klaster

Klaster	Harga Minimum	Harga Maksimum	Harga Rata-rata	Jumlah Data
0	15.000	1.183.000	514.411	7.294
1	1.185.000	2.271.000	1.855.157	2.288

Interpretasi terhadap hasil segmentasi menunjukkan bahwa Klaster 0 merepresentasikan segmen kendaraan dengan harga rendah hingga menengah, yang umumnya terdiri dari mobil-mobil dengan spesifikasi standar dan ditujukan untuk pasar massal. Sementara itu, Klaster 1 cenderung merepresentasikan segmen kendaraan dengan harga lebih tinggi, yang mengindikasikan keberadaan mobil-mobil premium atau unit dengan fitur tambahan yang lebih lengkap dan tahun produksi yang lebih baru. Untuk memvisualisasikan hasil klasterisasi secara lebih intuitif, dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA), yang kemudian divisualisasikan dalam ruang dua dimensi (2D). Hasil visualisasi ditampilkan pada Gambar 4.



Gambar 4. Visualisasi Klaster Mobil Bekas dalam Ruang PCA 2D

Visualisasi hasil PCA menampilkan distribusi data dalam dua dimensi utama: PCA1 pada sumbu X dan PCA2 pada sumbu Y. Kedua sumbu ini merepresentasikan komponen utama yang dihasilkan oleh Principal Component Analysis, yaitu kombinasi linear dari fitur-fitur asli yang menangkap variabilitas terbesar dalam data. Melalui reduksi dimensi ini, struktur data dapat divisualisasikan secara lebih sederhana tanpa mengurangi informasi penting yang terkandung di dalamnya [12].

Setiap titik pada grafik merepresentasikan satu unit mobil bekas, dan diberi warna berdasarkan klaster hasil dari algoritma K-Means. Titik berwarna biru menunjukkan Klaster 0, yang mencerminkan mobil dengan harga rendah hingga menengah. Sementara itu, titik berwarna hijau merepresentasikan Klaster 1, yang berisi mobil-mobil dengan harga tinggi atau kategori premium. Secara visual, klaster biru cenderung terkonsentrasi di sisi kiri (nilai PCA1 rendah), sedangkan

klaster hijau mendominasi sisi kanan (nilai PCA1 tinggi), dengan garis batas pemisahan terlihat di sekitar nilai PCA1 antara 0.2 hingga 0.3.

Klaster 0 (biru) tampak lebih padat dan beragam dalam hal distribusi, mencerminkan adanya variasi pada kelompok mobil dengan harga yang lebih terjangkau. Ini bisa mencakup berbagai model dan usia kendaraan, tetapi umumnya memiliki spesifikasi standar. Sebaliknya, Klaster 1 (hijau) menunjukkan penyebaran yang lebih luas terutama di wilayah PCA1 tinggi, menandakan adanya keragaman fitur pada mobil berharga mahal, seperti fitur tambahan, tahun produksi yang lebih baru, atau merek premium.

Pemisahan yang cukup jelas antara kedua klaster menunjukkan bahwa K-Means dengan K = 2 adalah pilihan yang tepat untuk segmentasi harga mobil bekas dalam dataset ini. Hal ini diperkuat oleh nilai Silhouette Score sebesar 0.7170 yang menunjukkan kualitas klasterisasi yang cukup baik. Visualisasi ini memberikan dasar yang kuat untuk melanjutkan ke tahap analisis lanjutan, seperti pemodelan prediksi harga berdasarkan klaster atau penerapan strategi pemasaran yang berbeda untuk masing-masing segmen pasar.

3.3. Analisis Regresi Berdasarkan Segmentasi Harga

Setelah proses segmentasi harga dilakukan menggunakan algoritma K-Means, langkah selanjutnya adalah membangun model regresi untuk memprediksi harga mobil bekas secara lebih akurat di masing-masing klaster yang merepresentasikan segmen harga rendah, menengah, dan tinggi. Tujuan dari pendekatan ini adalah untuk membandingkan kinerja beberapa algoritma **regresi** dalam konteks pasar yang telah disegmentasi, sehingga dapat diketahui model mana yang paling sesuai untuk masing-masing segmen. Dalam penelitian ini, empat algoritma regresi yang digunakan meliputi: Multiple Linear Regression (MLR) untuk menangani hubungan linier antar variabel, Polynomial Regression (PR) derajat 2 untuk mendeteksi pola non-linear, K-Nearest Neighbors Regression (KNNR) yang bekerja berdasarkan kedekatan fitur lokal, serta Random Forest Regressor (RFR) sebagai metode ensemble yang kuat dalam mengatasi kompleksitas dan interaksi antar variabel [11].

Evaluasi kinerja keempat model dilakukan menggunakan empat metrik umum dalam regresi, yaitu: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Koefisien Determinasi (R^2), dan Mean Absolute Percentage Error (MAPE) [13]. Metrik ini digunakan untuk menilai sejauh mana model mampu memprediksi harga mobil bekas secara akurat, khususnya setelah ditambahkan fitur klaster hasil segmentasi harga. Dengan pendekatan berbasis segmentasi, model dapat mempelajari pola harga yang lebih homogen sesuai dengan karakteristik masing-masing klaster, sehingga diharapkan performa prediksi meningkat secara signifikan dibandingkan jika model diterapkan langsung pada keseluruhan data tanpa segmentasi.

3.4. Perbandingan Kinerja Model Regresi Berdasarkan Hasil Evaluasi

Setelah dilakukan pelatihan pada empat model regresi yang berbeda, yaitu Multiple Linear Regression (MLR), Polynomial Regression (PR) derajat 2, K-Nearest Neighbors Regression (KNNR), dan Random Forest Regressor (RFR), evaluasi performa dilakukan menggunakan empat metrik utama: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Koefisien Determinasi (R^2), dan Mean Absolute Percentage Error (MAPE).Hasil evaluasi disajikan dalam Tabel 3, yang menunjukkan kinerja masing-masing model terhadap data uji untuk menilai sejauh mana model mampu memprediksi harga mobil bekas secara akurat, terutama setelah ditambahkan fitur klaster hasil segmentasi harga.

Tabel 3. Statistik Harga Mobil Bekas Berdasarkan Klaster

Model	MAE	RMSE	R^2	MAPE
MLR	164,992.61	227,317.44	0.8746	29.79%
PR (deg=2)	158,760.78	220,369.90	0.8821	28.31%
KNNR (k=5)	132,769.11	197,619.89	0.9052	24.83%
RFR	92,439.34	146,838.26	0.9477	18.25%

Berdasarkan hasil evaluasi pada Tabel 3, Random Forest Regressor (RFR) menunjukkan performa paling unggul dibandingkan model lainnya. Model ini memiliki nilai MAE terendah sebesar 92.439,34, RMSE sebesar 146.838,26, dan R^2 tertinggi sebesar 0,9477, yang menunjukkan bahwa model mampu menjelaskan hampir 95% variasi harga mobil bekas pada data uji. Selain itu, MAPE sebesar 18,25% juga merupakan yang paling rendah, mengindikasikan bahwa kesalahan prediksi relatif terhadap nilai aktual cukup kecil.

Perbedaan performa antara model regresi Multiple Linear Regression (MLR), Polynomial Regression (PR), K-Nearest Neighbors Regression (KNNR), dan Random Forest Regressor (RFR) dapat dijelaskan melalui karakteristik dasar masing-masing algoritma dan bagaimana mereka menangani kompleksitas data. Model Multiple Linear Regression (MLR) dan Polynomial Regression (PR) memiliki keterbatasan dalam menangkap hubungan non-linear yang kompleks antar fitur. Multiple Linear Regression (MLR) secara khusus mengasumsikan hubungan linier antara fitur prediktor dan target, yang menyebabkan penurunan akurasi ketika hubungan data bersifat non-linier. Polynomial Regression (PR) memang mencoba memodelkan hubungan non-linier, namun hanya pada tingkat polinomial tertentu yang dapat menyebabkan *overfitting* jika derajatnya terlalu tinggi, atau *underfitting* jika terlalu rendah.

Sementara itu, K-Nearest Neighbors Regression (KNNR) bekerja dengan prinsip kesamaan lokal. Meskipun lebih fleksibel dibandingkan regresi linier, KNNR sangat sensitif terhadap skala data dan noise, serta kinerjanya cenderung menurun pada data berdimensi tinggi (*curse of dimensionality*), terutama ketika data memiliki distribusi yang tidak seragam antar klaster.Sebaliknya, Random Forest Regressor (RFR) unggul karena merupakan metode *ensemble* berbasis pohon keputusan yang mampu memodelkan relasi non-linier dan interaksi antar fitur tanpa asumsi distribusi tertentu. Kemampuan RFR dalam melakukan *bagging* serta menangani *outlier* dan variabel yang saling berinteraksi menjadikannya lebih tangguh dalam menangkap pola-pola kompleks dalam data, termasuk perbedaan harga antar segmen yang telah dibentuk melalui K-Means.

Selain itu, RFR juga memiliki kemampuan *feature selection* internal yang membantu dalam mengurangi *noise*, sehingga meningkatkan akurasi prediksi secara signifikan. Dengan demikian, keunggulan performa RFR dibandingkan algoritma lain dalam studi ini dapat dikaitkan langsung dengan kemampuannya dalam menangani kompleksitas data, relasi non-linier, dan fitur-fitur interaktif yang relevan dalam prediksi harga mobil bekas.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa integrasi antara *unsupervised learning* dan *supervised learning* dapat secara signifikan meningkatkan akurasi prediksi harga kendaraan bekas. Segmentasi data menggunakan K-Means Clustering memungkinkan model mengenali kelompok harga dengan karakteristik serupa, yang kemudian digunakan sebagai fitur tambahan dalam proses pemodelan. Evaluasi terhadap jumlah klaster menggunakan metrik *inertia* dan *silhouette score* menunjukkan bahwa dua klaster memberikan segmentasi optimal, yang divisualisasikan secara jelas menggunakan PCA.

Eksperimen dengan empat algoritma regresi yaitu Multiple Linear Regression (MLR), Polynomial Regression (PR), K-Nearest Neighbors Regression (KNNR), dan Random Forest Regressor (RFR) dengan menunjukkan bahwa RFR memberikan performa terbaik dengan nilai MAE, RMSE, dan MAPE terendah serta nilai R^2 tertinggi. Keunggulan ini disebabkan oleh kemampuannya dalam menangani relasi non-linear, interaksi antar fitur, serta ketahanan terhadap outlier dan noise. Sebaliknya, model MLR dan PR memiliki keterbatasan dalam menangkap hubungan kompleks, sementara KNNR sensitif terhadap dimensi tinggi dan distribusi data yang tidak seimbang.

Dengan demikian, pendekatan kombinatif ini tidak hanya meningkatkan performa prediksi tetapi juga memberikan nilai tambah dalam pengembangan sistem pendukung keputusan berbasis data di sektor otomotif, khususnya pada penjualan kendaraan bekas. Temuan ini dapat menjadi dasar bagi penelitian selanjutnya untuk mengintegrasikan metode segmentasi yang lebih kompleks (seperti DBSCAN atau *Hierarchical Clustering*), teknik *feature engineering* yang lebih mendalam, serta

pendekatan *deep learning* berbasis data dan citra kendaraan untuk sistem prediksi yang lebih canggih dan adaptif terhadap dinamika pasar.

DAFTAR PUSTAKA

[1] M. Sholeh, S. Suraya, dan D. Andayati, "Machine Linear Untuk Analisis Regresi Linier Biaya Asuransi Kesehatan Dengan Menggunakan Python Jupyter Notebook," *Jurnal Edukasi Dan Penelitian Informatika (Jepin)*, vol. 8, no. 1, hlm. 20, 2022, doi: 10.26418/jp.v8i1.48822.
[2] A. Amalia, M. Radhi, S. H. Sinurat, D. R. H. Sitompul, dan E. Indra, "Prediksi Harga Mobil Menggunakan Algoritma Regressi Dengan Hyper-Parameter Tuning," *Jurnal Sistem Informasi Dan Ilmu Komputer Prima(jusikom Prima)*, vol. 4, no. 2, hlm. 28–32, 2022, doi: 10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2479.
[3] S. Mulyanda, S. Defit, dan S. Sumijan, "Analisis Data Mining Menggunakan Algoritma C4.5 Untuk Prediksi Harga Pasar Mobil Bekas," *Jurnal Komtekinfo*, hlm. 116–121, 2023, doi: 10.35134/komtekinfo.v10i3.427.
[4] H. Syahputra dan M. F. Anwar, "Penerapan Metode Multifactor Evaluation Process (MFEB) Pada Sistem Penunjang Keputusan Untuk Pemilihan Mobil Bekas Terbaik Berbasis Web," *Jurnal Pustaka Robot Sister*, vol. 2, no. 2, hlm. 27–31, 2024, doi: 10.55382/jurnalpustakarobotsister.v2i2.743.
[5] F. H. Hamdanah dan D. Fitrianah, "Analisis Performansi Algoritma Linear Regression Dengan Generalized Linear Model Untuk Prediksi Penjualan Pada Usaha Mikra, Kecil, Dan Menengah," *Jurnal Nasional Pendidikan Teknik Informatika (Janapati)*, vol. 10, no. 1, hlm. 23, 2021, doi: 10.23887/janapati.v10i1.31035.
[6] A. Alcalde-Barros, D. García-Gil, S. García, dan F. Herrera, "DPASF: A Flink Library for Streaming Data Preprocessing," *Big Data Analytics*, vol. 4, no. 1, 2019, doi: 10.1186/s41044-019-0041-8.
[7] I. Chahid, A. K. Elmiad, dan M. E. Badaoui, "Data Preprocessing for Machine Learning Applications in Healthcare: A Review," hlm. 1–6, 2023, doi: 10.1109/sita60746.2023.10373591.
[8] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, dan O. Tabona, "A Survey on Missing Data in Machine Learning," *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00516-9.
[9] S. Diehn *dkk.*, "Discrimination of Grass Pollen of Different Species by FTIR Spectroscopy of Individual Pollen Grains," *Analytical and Bioanalytical Chemistry*, vol. 412, no. 24, hlm. 6459–6474, 2020, doi: 10.1007/s00216-020-02628-2.
[10] A. F. H. Marsandy, M. N. Hayati, dan M. Fauziyah, "Klasterisasi Prevalensi Stunting Menggunakan K-Prototype Pada Data Campuran," *Metik Jurnal*, vol. 8, no. 2, hlm. 48–54, 2024, doi: 10.47002/metik.v8i2.824.
[11] Y. Zhang, "Utilize the Machine Learning Models to Forecast Home Values.:Seattle U.S.," *Tcsisr*, vol. 5, hlm. 330–336, 2024, doi: 10.62051/7gzntt12.
[12] I. Valova, N. Gueorguieva, T. Mai, dan R. Chen, "Fuzzy Clustering in High-Dimensional Spaces: Visualization and Performance Metrics," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 28, no. 2, hlm. 313–333, 2024, doi: 10.3233/kes-221614.
[13] J. Ye, "Analysis on E-Commerce Order Cancellations Using Market Segmentation Approach," hlm. 33–40, 2021, doi: 10.1145/3450588.3450596.