**Evaluasi Kinerja Random Forest, Support Vector Machine dan Naive Bayes dalam Prediksi Harga Mobil Ford**

**Andika Nuralamsyah 11,2, Denis Firmansyah 23, Nizzami Ramdhan Arraudy 33**

1 Department of Informatics Engineering, UIN Sunan Gunung Djati Bandung, Indonesia

2 Department of Informatics Engineering, UIN Sunan Gunung Djati Bandung, Indonesia

3 Department of Informatics Engineering, UIN Sunan Gunung Djati Bandung, Indonesia

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Article history:***  Received month dd, yyyy  Revised month dd, yyyy  Accepted month dd, yyyy |  | Prediksi harga mobil bekas merupakan tantangan kompleks yang memerlukan pendekatan komputasional yang akurat. Penelitian ini membandingkan kinerja tiga algoritma machine learning—Random Forest, Support Vector Machine (SVM), dan Naive Bayes—dalam memprediksi harga mobil Ford bekas menggunakan dataset dari Kaggle, dengan tahapan meliputi pra-pemrosesan data (pembersihan data, one-hot encoding, normalisasi), pembagian dataset (80:20), serta evaluasi model menggunakan metrik MAE, RMSE, dan MAPE. Hasil menunjukkan bahwa Random Forest memberikan kinerja terbaik dengan R² sebesar 92,92%, diikuti SVM (84,23%) dan Naive Bayes (70,19%), dengan analisis feature importance menunjukkan bahwa atribut numerik seperti tahun dan jarak tempuh memiliki pengaruh dominan dibanding fitur kategorik seperti model dan jenis bahan bakar. Keunggulan Random Forest terletak pada kemampuannya menangkap hubungan non-linear, sementara SVM kurang andal terhadap outlier dan Naive Bayes tidak cocok untuk regresi karena asumsi distribusi Gaussian yang tidak sesuai. Penelitian ini memberikan kontribusi praktis berupa rekomendasi algoritma berdasarkan kebutuhan akurasi dan efisiensi komputasi, serta membuka peluang eksplorasi metode ensemble hybrid atau deep learning untuk meningkatkan akurasi pada prediksi harga ekstrem. |
| ***Keywords:***  Prediksi Harga  Metrik Regresi  Random Forest  Support Vector Machine  ***Naïve Bayes*** |
| *This is an open access article under the* [*CC BY-SA*](https://creativecommons.org/licenses/by-sa/4.0/) *license.*  A picture containing text, clipart  Description automatically generated |
| ***Corresponding Author:***  Asep Ujang  Jurusan Teknik Informatika, UIN Sunan Gunung Djati Bandung  Email: asepujang@gmail.com | | |

1. **PENDAHULUAN**

Perkembangan teknologi *machine learning* di Indonesia terus menunjukkan tren positif, terutama dalam sektor otomotif. Banyak platform prediksi harga mobil bekas yang semakin marak digunakan oleh dealer dan konsumen untuk membantu menentukan harga jual dan beli yang lebih objektif dan akurat.

Studi oleh Dewi, Haikal, Sulistyowati, Fitriani, dan Kuswandono (2024) menunjukkan bahwa algoritma *Random Forest,* dengan mempertimbangkan faktor seperti model, tahun produksi, kapasitas mesin, jarak tempuh, transmisi, dan bahan bakar, mampu menjelaskan hingga 90% variabilitas harga mobil bekas dalam aplikasi berbasis web. Selain itu, Winata (2024) membandingkan algoritma *Gradient Boosting Machine* dan *Random Forest* dalam konteks yang sama, di mana *Random Forest* menunjukkan akurasi mencapai 96 % dalam klasifikasi harga mobil bekas.

Namun demikian, sebagian besar penelitian sebelumnya hanya berfokus pada satu atau dua algoritma, terutama algoritma *ensemble* seperti *Random Forest* atau *Gradient Boosting Machine* (GBM), tanpa membandingkannya secara komprehensif dengan algoritma klasik lain seperti *Support Vector Machine* (SVM) dan *Naive Bayes*. Padahal, studi lain menunjukkan bahwa SVM dan *Naive Bayes* juga memiliki akurasi tinggi dalam domain klasifikasi dan prediksi kendaraan. Nugroho dan Wibowo (2023) menggunakan *Naive Bayes* dengan teknik seleksi fitur *forward selection*, menghasilkan akurasi hingga 97,8 % untuk data klasifikasi mobil bekas. Selain itu, studi oleh Paul, Wiguna, dan Santoso (2023) menerapkan algoritma SVM dan *Naive Bayes* untuk mengklasifikasi jenis mobil terlaris berdasarkan data produksi. Hasilnya, akurasi *Naive Bayes* berada di kisaran 97,4 %, sedangkan SVM mencapai 96,5%.

Selain prediksi harga, teknologi ini juga banyak digunakan dalam analisis atribut kendaraan. Contohnya, Luthfi Rifky dkk. (2022) berhasil mengklasifikasikan merek mobil paling laris laku di pasar dengan akurasi mencapai 92,2% menggunakan *Naive Bayes* dari data GAIKINDO. Temuan-temuan ini memperkuat pentingnya mengevaluasi ketiga metode secara komparatif untuk memahami kekuatan mereka dalam konteks dataset Ford.

Sebagai solusi atas kesenjangan tersebut, penelitian ini menggunakan pendekatan eksperimen komparatif untuk mengevaluasi kinerja tiga algoritma machine learning dalam memprediksi harga mobil bekas menggunakan dataset *Ford Car Price Prediction* dari Kaggle. Pendekatan ini melibatkan tahapan pra-pemrosesan data, pelatihan model, pengujian performa, serta evaluasi menggunakan metrik regresi seperti *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), dan *Mean Absolute Percentage Error* (MAPE). Dengan menggunakan metrik yang relevan, studi ini bertujuan untuk memperoleh gambaran menyeluruh mengenai keunggulan dan kelemahan masing-masing algoritma dalam konteks prediksi harga mobil berdasarkan atribut-atribut seperti tahun produksi, transmisi, kapasitas mesin, dan jarak tempuh.

Berbeda dengan penelitian sebelumnya yang cenderung hanya menguji satu atau dua algoritma secara terpisah, penelitian ini menghadirkan pendekatan komparatif langsung terhadap tiga algoritma berbeda yang mewakili pendekatan *ensemble*, *margin-based*, dan *probabilistik*. Selain itu, penggunaan dataset *Ford Car Price Prediction* dari Kaggle memungkinkan analisis dilakukan secara fokus pada satu merek mobil, sehingga hasil evaluasi menjadi lebih spesifik, kontekstual, dan relevan untuk studi kasus otomotif yang tersegmentasi. Nilai kebaruan lain dari studi ini adalah pada penggunaan tiga metrik evaluasi regresi secara simultan (MAE, RMSE, dan MAPE) yang memberikan sudut pandang lebih menyeluruh dalam menilai performa model prediktif.

Hasil dari studi ini diharapkan dapat menjadi kontribusi dalam domain sistem pendukung keputusan harga otomotif berbasis data di Indonesia, serta memberikan gambaran pilihan algoritma mana yang sebaiknya digunakan sesuai kebutuhan bisnis atau riset.

1. **METODE**

Penelitian ini menggunakan desain eksperimen dengan pendekatan komparatif. Tujuan utamanya adalah untuk mengevaluasi dan membandingkan kinerja tiga algoritma *machine learning* dalam memprediksi harga mobil Ford bekas menggunakan dataset *Ford Car Price Prediction*.

* 1. ***Prosedur Penelitian***

Langkah-langkah penelitian dilakukan sebagai berikut:

2.1.1. Akuisisi Data

Dataset diunduh dari situs Kaggle yang berisi data mobil merek Ford. Atribut yang tersedia meliputi: *year, mileage, transmission, fuelType, tax, mpg, engineSize*, dan *price* sebagai variabel target.

2.1.2. Pra-pemrosesan Data

Tahapan pra-pemrosesan yang dilakukan mencakup:

1.Menghapus duplikat dan nilai kosong (*missing values*).

2.Mengubah variabel kategorikal (*transmission, fuelType*) menggunakan teknik *One-Hot Encoding*.

3.Menormalisasi fitur numerik (*mileage, tax, mpg, engineSize*) menggunakan *Min-Max Scaling.*

4.Membagi data menjadi data latih dan data uji dengan rasio 80:20.

2.1.3. Pelatihan dan Evaluasi Model

Ketiga algoritma dilatih menggunakan data latih dan diuji menggunakan data uji. Evaluasi dilakukan menggunakan tiga metrik regresi berikut:

1.Mean Absolute Error (MAE)

2.Root Mean Square Error (RMSE)

3.Mean Absolute Percentage Error (MAPE)

Ketiga metrik ini dipilih untuk mengevaluasi sejauh mana prediksi model mendekati nilai aktual.

***2.2. Pseudocode***

Berikut adalah pseudocode umum alur kerja pelatihan dan evaluasi model:

Input: Dataset Ford D

Output: MAE, RMSE, MAPE

1. Bersihkan D dari data kosong dan duplikat

2. Encode variabel kategorikal

3. Normalisasi fitur numerik

4. Bagi D menjadi data latih dan uji (80:20)

5. Untuk setiap algoritma dalam {*Random Forest, SVM, Naive Bayes*}:

    a. Latih model dengan data latih

    b. Prediksi harga pada data uji

    c. Hitung nilai MAE, RMSE, dan MAPE

6. Bandingkan hasil ketiga model

***2.3. Implementasi***

Seluruh eksperimen dilakukan menggunakan bahasa pemrograman *Python*, dengan pustaka utama *Scikit-learn*. Untuk pengolahan data digunakan *Pandas* dan *NumPy*, serta visualisasi dilakukan menggunakan *Matplotlib* dan *Seaborn*

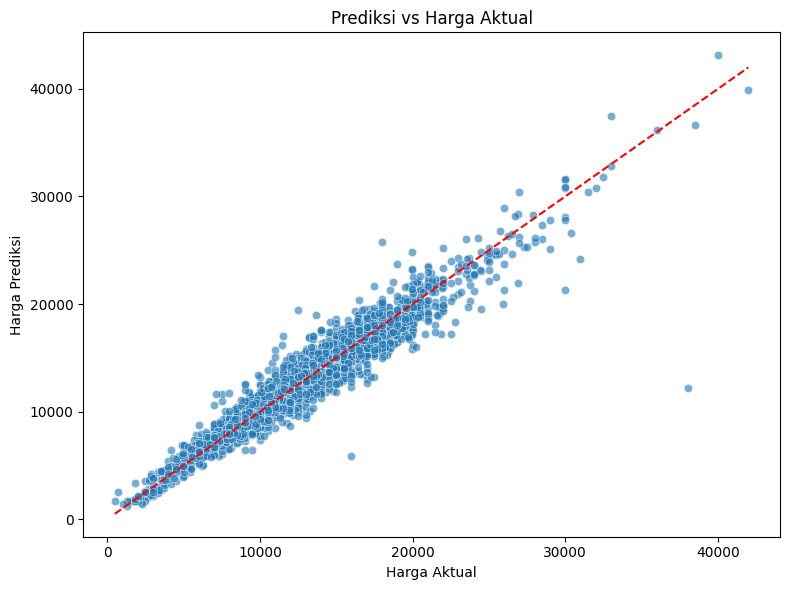
1. **HASIL PENELITIAN**

Penelitian ini menyajikan hasil evaluasi kinerja dari tiga algoritma *machine learning* berbeda, yaitu *Random Forest*, *Support Machine Vector,* serta *Naive Bayes* untuk memprediksi harga mobil Ford. Ketiga model ini dilatih dengan menggunakan 80% dari *dataset,* serta divalidasi dengan 20% dari sisanya. Hasil pengujian menunjukkan tingkat akurasi yang berbeda untuk ketiga algoritma.

***3.1.  Random Forest***

Model *Random Forest* menunjukkan kinerja yang menjanjikan dalam memprediksi harga. Model ini mencatat nilai R-*squared* sebesar 92,92. Nilai R-*squared* ini mengindikasikan bahwa sebagian besar prediksi harga yang dihasilkan oleh model *Random Forest* sesuai dengan harga aktual di dataset pengujian.

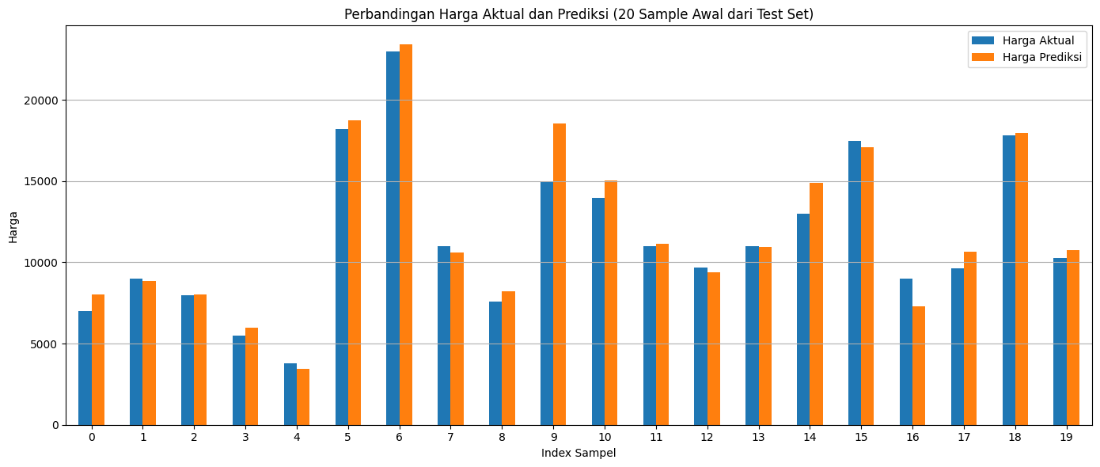
Untuk memberikan gambaran yang lebih jelas mengenai performa model, kami melakukan perbandingan antara harga prediksi yang dihasilkan oleh *Random Forest* dan harga aktual dari dataset pengujian. Hasilnya menunjukkan kecenderungan prediksi yang sangat mendekati harga aktual, meskipun terdapat sedikit deviasi pada beberapa titik data ekstrem.



*Gambar 3.1. Plot perbandingan nilai prediksi dan aktual Random Forest*

Secara visual, plot perbandingan antara harga prediksi dan harga aktual menunjukkan pola yang sangat mirip, dengan titik-titik data yang tersebar di sekitar garis identitas. Deviasi antara harga prediksi dan harga aktual relatif kecil, mengindikasikan kemampuan model untuk melakukan generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya.

Untuk visualisasi lebih lanjut mengenai perbandingan harga aktual dan prediksi yang dihasilkan oleh model *Random Forest*, *Gambar 3.2* menampilkan data untuk 20 sampel awal dari *test set.*



*Gambar 3.2 Grafik perbandingan harga aktual dan prediksi 20 index pertama Random Forest*

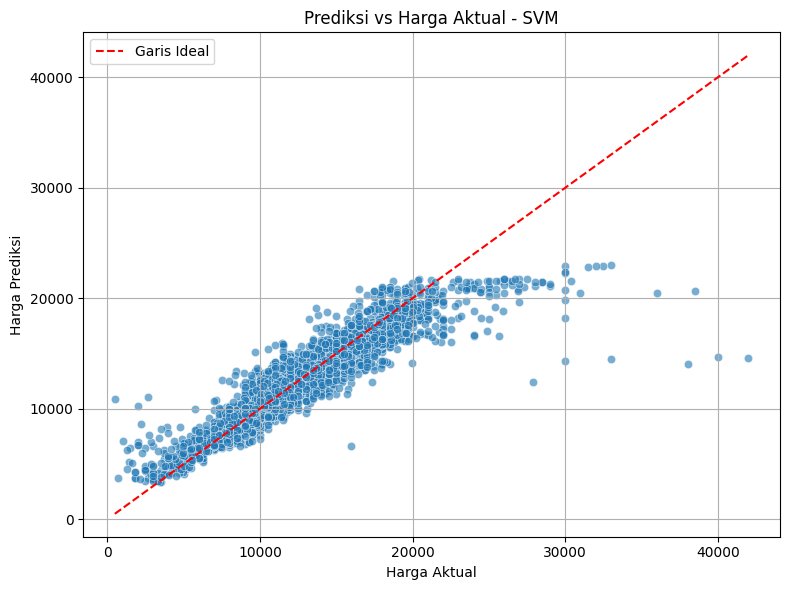
Dari *Gambar 3.2*, terlihat bahwa terdapat perbedaan yang relatif kecil antara harga prediksi dan harga aktual untuk sebagian besar indeks. Meskipun demikian, pada beberapa kasus seperti indeks 9, dapat diamati adanya deviasi yang cukup signifikan.

Tingkat akurasi yang tinggi ini dapat diatribusikan pada sifat *Random Forest* yang merupakan ensemble learning, di mana model ini menggabungkan prediksi dari banyak pohon keputusan untuk mengurangi overfitting dan meningkatkan stabilitas model. Hal ini memungkinkan *Random Forest* untuk mengatasi kompleksitas data harga yang mungkin memiliki hubungan non-linear antar variabel.

***3.2.  Support Vector Machine***

Model *Support Vector Machine* menunjukan kinerja yang cukup baik dalam memprediksi harga, ditunjukan dengan hasil evaluasi mencapai R-squaredi 0.8423. Nilai ini menunjukkan bahwa model SVM mampu mengklasifikasikan 84,23% data testing dengan benar ke dalam rentang harga yang sesuai.

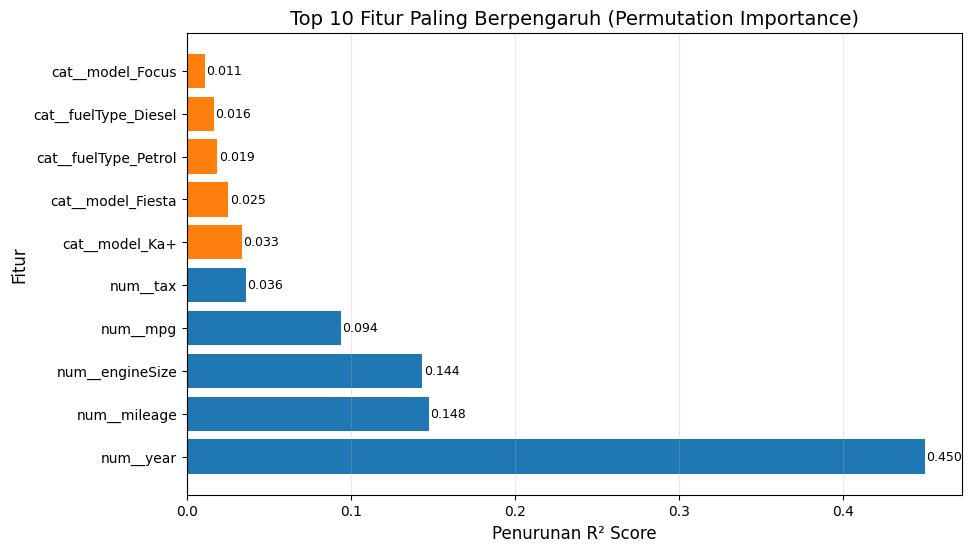
Untuk mengevaluasi kinerja model *Support Vector Machine* (SVM**)** dalam memprediksi harga mobil Ford, kami membandingkan harga prediksi dengan harga aktual dari data uji. Hasilnya menunjukkan bahwa SVM mampu menghasilkan prediksi yang relatif dekat dengan nilai sebenarnya, terutama untuk mobil dengan harga menengah. Namun, terdapat beberapa deviasi pada titik data ekstrem, seperti mobil dengan harga sangat tinggi atau rendah.



*Gambar 3.3. Plot perbandingan nilai prediksi dan aktual Support Vector Machine*

Berdasarkan visualisasi, dapat disimpulkan bahwa model ini menunjukkan performa yang relatif baik dalam memprediksi harga, dengan sebagian besar titik data mengelompok di sekitar garis ideal. Namun, terlihat adanya penyimpangan yang lebih signifikan pada beberapa kasus tertentu, terutama untuk mobil dengan karakteristik ekstrem, yang mungkin disebabkan oleh sensitivitas SVM terhadap outlier atau ketidakmampuan model sepenuhnya dalam menangkap hubungan non-linear yang kompleks pada data.

Penyebaran titik yang semakin melebar di area harga tertentu mengindikasikan bahwa model mengalami kesulitan dalam memprediksi nilai-nilai di rentang tersebut secara konsisten, menyarankan perlunya optimasi lebih lanjut terhadap parameter model atau teknik preprocessing data untuk meningkatkan akurasi prediksi



*Gambar 3.4 Grafik perbandingan harga aktual dan prediksi 10 index pertama SVM*

Grafik ini menunjukkan 10 fitur paling berpengaruh dalam prediksi harga mobil Ford bekas menggunakan model SVM, diukur berdasarkan Permutation Importance. Fitur numerik seperti tahun produksi (num\_year) dan jarak tempuh (num\_mileage) memiliki pengaruh terbesar, yang menegaskan bahwa usia dan penggunaan mobil menjadi faktor penentu utama harga. Sementara itu, fitur kategorik seperti model mobil (cat\_model\_Focus, cat\_model\_Fiesta) dan jenis bahan bakar (cat\_fuelType\_Diesel/Petrol) juga signifikan, meskipun dampaknya lebih kecil. Hasil ini mengonfirmasi bahwa karakteristik teknis (numerik) lebih dominan dalam memengaruhi prediksi harga dibandingkan fitur kategorik, dengan tahun produksi muncul sebagai faktor paling kritis.

***3.3.  Naive Bayes***

Model algoritma Naive Bayes digunakan untuk memprediksi harga kendaraan berdasarkan fitur-fitur seperti model, jenis transmisi, dan tipe bahan bakar. Dataset yang digunakan telah melalui tahap pra-pemrosesan, termasuk teknik one-hot encoding untuk variabel kategorikal, serta pemisahan data menjadi data latih dan data uji dengan rasio 80:20.

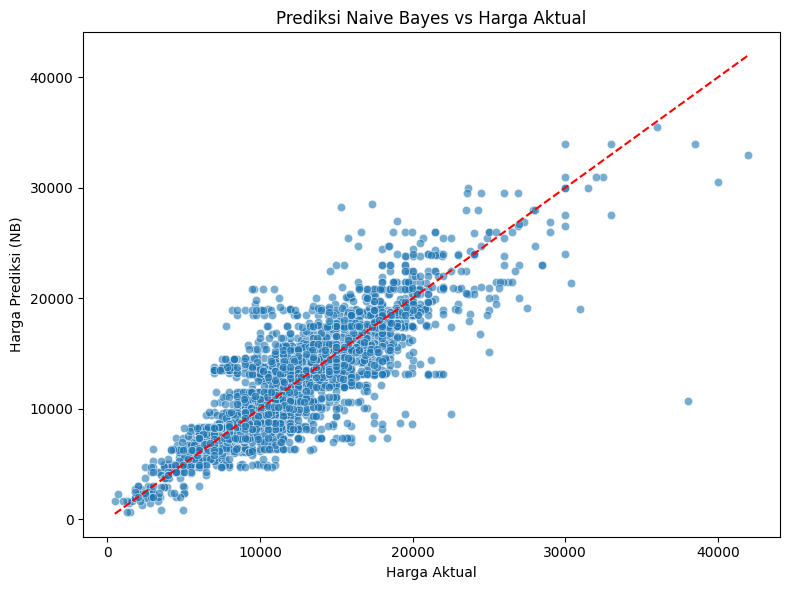
Setelah model *Gaussian Naive Bayes* dilatih, evaluasi dilakukan menggunakan metrik *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan koefisien determinasi (*R-squared*). Hasil evaluasi ditunjukkan pada Tabel berikut:

|  |  |
| --- | --- |
| **Metrik Evaluasi** | **Nilai** |
| MAE | 179.11 |
| MSE | 6,680,573.76 |
| R2 | 0.7019 |

*Tabel 3.5. Hasil Evaluasi Naïve Bayes*

Dari hasil tersebut, terlihat bahwa model Naive Bayes menghasilkan MAE sebesar 1797.11, yang mengindikasikan bahwa secara rata-rata, prediksi harga kendaraan meleset sekitar 1.797 satuan mata uang dari nilai sebenarnya. Nilai MSE yang besar menunjukkan adanya sejumlah prediksi dengan galat yang cukup besar. Namun demikian, nilai *R-squared*sebesar 0.7019 mengindikasikan bahwa model mampu menjelaskan sekitar 70.19% variansi dari data harga aktual, yang cukup baik mengingat bahwa Naive Bayes bukanlah algoritma yang secara khusus dirancang untuk regresi numerik.

Visualisasi pada Gambar 3.6 menunjukkan sebaran prediksi harga terhadap harga aktual. Garis merah putus-putus menandai garis ideal dimana prediksi tepat sama dengan nilai aktual. Titik-titik data yang menyebar di sekitar garis tersebut menandakan tingkat galat dari masing-masing prediksi. Meskipun terlihat bahwa sebagian besar titik berada dekat dengan garis ideal, terdapat pula beberapa prediksi ekstrem yang jauh dari garis, yang menyebabkan meningkatnya nilai MSE.



*Gambar 3.6. Plot perbandingan nilai prediksi dan aktual Naive Bayes*

Secara keseluruhan, hasil ini menunjukkan bahwa meskipun Naive Bayes dapat digunakan untuk memprediksi harga, algoritma ini kurang optimal untuk regresi numerik, terutama karena asumsi distribusi data yang digunakan tidak sesuai dengan sifat data harga yang cenderung tidak mengikuti distribusi Gaussian murni. Eksperimen ini dilakukan semata-mata sebagai pembanding atau *baseline model* terhadap algoritma regresi lain yang lebih sesuai, seperti *Random Forest*, XGBoost, atau Regresi Linier.

1. **CONCLUSION**

Berdasarkan analisis komparatif tiga algoritma machine learning untuk prediksi harga mobil Ford bekas, dapat disimpulkan bahwa Random Forest menunjukan performa terbaik dengan nilai R² sebesar 92,92%, mengungguli SVM dan Naive Bayes. Keunggulan Random Forest ini disebabkan kemampuannya menangkap hubungan non-linear dalam data melalui pendekatan ensemble learning. Meskipun SVM menunjukkan hasil yang cukup baik dengan R² 84,23%, model ini mengalami kesulitan dalam memprediksi harga ekstrem karena sensitivitas terhadap outlier. Di sisi lain, Naive Bayes menjadi yang paling tidak cocok untuk kasus ini dengan R² hanya 70,19% dan MAE 1797,11, karena asumsi distribusi Gaussian yang tidak sesuai dengan karakteristik data harga mobil.

Analisis feature importance mengungkapkan bahwa fitur numerik seperti tahun produksi dan jarak tempuh menjadi faktor penentu utama harga, sementara fitur kategorik seperti model spesifik dan jenis bahan bakar memiliki pengaruh lebih kecil. Temuan ini konsisten di ketiga model dan sesuai dengan logika pasar mobil bekas dimana kondisi teknis kendaraan lebih berpengaruh daripada faktor kategorikal.

Untuk implementasi praktis, Random Forest direkomendasikan sebagai pilihan utama karena akurasinya yang tinggi, meski membutuhkan sumber daya komputasi lebih besar. SVM dapat dipertimbangkan sebagai alternatif dengan optimasi parameter lebih lanjut, sementara Naive Bayes kurang cocok digunakan kecuali sebagai baseline. Penelitian lanjutan dapat mengeksplorasi teknik feature engineering tambahan dan pendekatan hybrid untuk meningkatkan akurasi prediksi, khususnya pada kisaran harga ekstrem yang masih menjadi tantangan bagi semua model.

**DAFTAR PUSTAKA**

[1] Winata, A. J. (2024, Desember). Prediksi harga mobil bekas menggunakan algoritma Gradient Boosting Machine dan Random Forest. *Jurnal Inovasi Pendidikan Kreatif*, 5(4).

[2] Purnomo, D., Firgiawan, W., & Nur, N. (2025). Komparasi Algoritma Random Forest, Naïve Bayes, dan SVM pada Sentimen Kebijakan PPN 12%. *Jurnal Tekno Kompak*, *19*(2), 155 - 167. <https://doi.org/10.33365/teknokompak.v19i2.122>

[3] Ningsih, W., Alfianda, B., Rahmaddeni, R., & Wulandari, D. (2024). Perbandingan Algoritma SVM dan Naïve Bayes dalam Analisis Sentimen Twitter pada Penggunaan Mobil Listrik di Indonesia : Comparison of Naive Bayes and SVM Algorithms in Twitter Sentiment Analysis on Electric Car Use in Indonesia. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(2), 556-562. <https://doi.org/10.57152/malcom.v4i2.1253>

[4] Belinda Eka Sarah Dewi, Haikal, S., Sulistyowati, H., Fitriani, R., & Pranowo Kuswandono, D. (2024). PENERAPAN MACHINE LEARNING MENGGUNAKAN ALGORITMA RANDOM FOREST UNTUK PREDIKSI HARGA MOBIL BEKAS. *Jurnal Teknologi Informasi Dan Digital*, *2*(1), 20- 31. Retrieved from <https://banisalehjurnal.ubs.ac.id/index.php/tridi/article/view/38>.

[5] Nugroho, H. A., & Wibowo, S. (2023). Penerapan algoritma Naive Bayes berbasis forward selection untuk memprediksi penjualan mobil bekas. *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*.

[6] Paul, H., Wiguna, A. S., & Santoso, H. (2023). Penerapan algoritma Support Vector Machine dan Naive Bayes untuk klasifikasi jenis mobil terlaris berdasarkan produksi di Indonesia. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 7(1).

[7] Rifky, L., Nugraha, Z., Saputra, B., Pratama, D., Raswir, E., & Pratama, Y. (2022). Implementasi data mining untuk penjualan mobil menggunakan metode Naive Bayes. *Jurnal Informatika dan Rekayasa Komputer (JAKAKOM)*, 2(2).