
PREDIKSI HARGA MOBIL BEKAS MENGGUNAKAN ALGORITMA GRADIENT BOOSTING MACHINE DAN RANDOM FOREST

Ardiansyah Jaya Winata¹

¹Universitas Tarumanagara, Indonesia
ardiansyah.535210014@stu.untar.ac.id

ABSTRAK

Penelitian ini bertujuan untuk memprediksi harga mobil bekas menggunakan dua algoritma *machine learning*, yaitu *Gradient Boosting Machine (GBM)* dan *Random Forest*. Data yang digunakan berasal dari platform *Cars24*, yang mencakup informasi mengenai spesifikasi mobil, seperti tahun pembuatan, jarak tempuh, jenis bahan bakar, dan kapasitas mesin. Kedua algoritma dievaluasi berdasarkan metrik *Precision*, *Recall*, *F1-Score*, dan *Accuracy* pada tiga kategori harga: rendah, sedang, dan tinggi. Hasil penelitian menunjukkan bahwa *GBM* memiliki akurasi keseluruhan sebesar 0,97, sedikit lebih unggul dibandingkan *Random Forest* yang memiliki akurasi 0,96. Kedua model menunjukkan performa yang baik pada kategori harga rendah dan tinggi, tetapi kesulitan dalam mengklasifikasikan kategori harga sedang. Hal ini mengindikasikan bahwa meskipun *GBM* dan *Random Forest* efektif untuk prediksi harga ekstrem, keduanya memiliki keterbatasan dalam mendeteksi harga di kategori menengah.

Kata Kunci: Mobil Bekas, Prediksi Harga, *Machine Learning*, *Gradient Boosting Machine*, *Random Forest*.

ABSTRACT

This study aims to predict used car prices using two machine learning algorithms: Gradient Boosting Machine (GBM) and Random Forest. The data used in this study were obtained from the Cars24 platform, which includes information on car specifications such as manufacturing year, mileage, fuel type, and engine capacity. Both algorithms were evaluated based on Precision, Recall, F1-Score, and Accuracy metrics across three price categories: low, medium, and high. The results indicate that GBM achieved an overall accuracy of 0.97, slightly outperforming Random Forest with an accuracy of 0.96. Both models demonstrated strong performance in predicting low and high price categories but struggled with classification in the medium price category. This suggests that, while GBM and Random Forest are effective for predicting extreme price categories, they face limitations in detecting prices within the medium range.

Keywords: *Used Car, Price Prediction, Machine Learning, Gradient Boosting Machine, Random Forest.*

A. PENDAHULUAN

Mobil bekas adalah kendaraan yang sebelumnya dimiliki dan digunakan oleh satu atau lebih pemilik sebelum dijual kembali kepada pembeli baru (Zhu, 2023). Penjualan mobil bekas telah menjadi salah satu sektor penting dalam industri otomotif, terutama bagi konsumen yang mencari kendaraan dengan harga lebih terjangkau dibandingkan mobil baru. Dengan pertumbuhan populasi dan peningkatan kebutuhan transportasi pribadi, mobil bekas menjadi pilihan menarik bagi konsumen yang ingin memiliki kendaraan tanpa harus mengeluarkan biaya yang signifikan (Bharambe et al., 2022)(Kriswantara & Sadikin, 2022).

Industri mobil bekas tidak hanya menawarkan manfaat bagi konsumen, tetapi juga menciptakan peluang bisnis besar. Di banyak negara, pasar mobil bekas berkembang pesat dengan meningkatnya permintaan. Pelaku usaha di sektor ini menyediakan berbagai macam kendaraan bekas dari berbagai merek, tahun produksi, dan kondisi, sehingga konsumen dapat memilih mobil yang sesuai dengan anggaran dan kebutuhan pengguna (Yadav et al., 2021). Kemajuan teknologi juga mendorong perkembangan industri ini, dengan munculnya platform digital dan situs *e-commerce* yang menyediakan layanan jual-beli mobil bekas secara daring. Layanan ini memudahkan konsumen menelusuri pilihan, membandingkan harga, serta mendapatkan informasi spesifikasi dan riwayat kendaraan (Varshitha et al., 2022)(Hankar et al., 2022)(Lin & Hashim, 2023).

Salah satu platform *e-commerce* yang aktif dalam penjualan mobil bekas adalah [Cars24](#). Platform ini menyediakan layanan komprehensif untuk jual-beli mobil bekas, termasuk evaluasi kondisi kendaraan, layanan keuangan, dan garansi bagi pembeli. Data dari Cars24 mencakup berbagai informasi penting, termasuk tahun pembuatan, jarak tempuh, jenis bahan bakar, tipe transmisi, kapasitas mesin, jumlah kursi, dan harga jual. Dataset Cars24 yang tersedia di [Kaggle](#), diperoleh melalui proses *web scraping* dan mencakup ribuan mobil bekas yang tercantum di platform tersebut. Penelitian ini menggunakan analisis prediktif dengan metode *machine learning*, yaitu *Gradient Boosting Machine* (GBM) dan *Random Forest*, untuk memprediksi harga mobil bekas berdasarkan berbagai fitur dalam dataset.

B. METODE PENELITIAN

1. Data

Dalam penelitian ini, data yang digunakan diperoleh dari platform [Kaggle](#). Dataset ini mencakup 7.801 entri dengan 13 kolom, mencakup informasi penting tentang mobil bekas seperti nama model, lokasi, tahun pembuatan, jarak tempuh, jenis bahan bakar, transmisi, status kepemilikan, efisiensi bahan bakar, kapasitas mesin, daya, jumlah kursi, harga jual mobil bekas.

Beberapa tahap pra-pemrosesan data dilakukan untuk mempersiapkan dataset agar siap digunakan dalam model prediksi. Langkah-langkah ini meliputi penghapusan data kosong, transformasi kolom kategori seperti "Jenis Bahan Bakar" dan "Transmisi" menjadi data numerik, serta konversi tipe data pada kolom "Tahun". Langkah-langkah ini bertujuan untuk memastikan bahwa data siap diolah oleh algoritma machine learning, sehingga dapat menghasilkan prediksi harga mobil bekas yang lebih akurat.

2. *Pra-pemrosesan*

Dataset awal dibagi menjadi dua bagian, yaitu data latih (80%) dan data uji (20%). Pembagian ini bertujuan untuk menguji performa setiap metode dalam mengklasifikasikan penerbangan yang mengalami keterlambatan dan yang tidak. Hasil eksperimen dievaluasi dan dibandingkan berdasarkan nilai akurasi yang dihasilkan oleh setiap metode.

3. *Random Forest*

Random Forest merupakan salah satu algoritma machine learning untuk klasifikasi data dalam jumlah yang besar (Wanli Sitorus et al., 2021). Metode ini merupakan sebuah ensemble (kumpulan) metode pembelajaran menggunakan pohon keputusan sebagai *base classifier* yang dibangun dan dikombinasikan. *Decision tree* atau pohon pengambil keputusan adalah sebuah diagram alir yang berbentuk seperti pohon, yang memiliki sebuah *root node* yang digunakan untuk mengumpulkan data (Baldanullah et al., 2023)(Schonlau & Zou, 2020). *Random Forest* biasanya digunakan untuk analisa klasifikasi dan regresi (Suliztia, 2020).

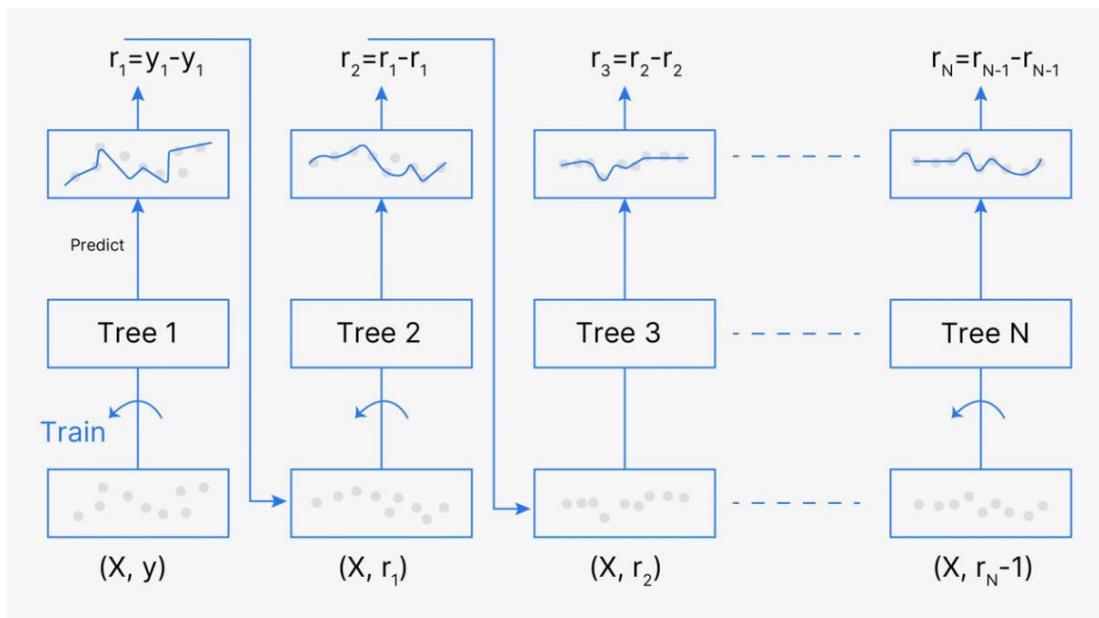
$$\text{Gini} = 1 - \sum_{i=1}^c (P_i)^2 \quad (1)$$

Saat melakukan *Random Forests* berdasarkan data klasifikasi, indeks Gini digunakan sebagai faktor penentu dalam membuat keputusan tentang percabangan node pada pohon keputusan. Indeks Gini memungkinkan perhitungan ketidakmurnian atau ketidakjelasan di setiap cabang node, dan ini membantu dalam pemilihan cabang yang memiliki kemungkinan

lebih tinggi untuk menghasilkan prediksi yang akurat. Dalam rumus ini, " p_i " mengacu pada frekuensi relatif kelas yang diamati dalam data, dan " c " adalah jumlah kelas yang ada dalam masalah klasifikasi.

4. *Gradient Boosting Machine (GBM)*

Gradient Boosting Machine (GBM) adalah teknik pembelajaran mesin yang digunakan untuk tugas regresi dan klasifikasi (Sai et al., 2023). GBM membangun model prediktif yang kuat dengan menggabungkan sejumlah model sederhana, yang dikenal sebagai "weak learners," secara bertahap untuk meningkatkan akurasi prediksi (Omotehinwa et al., 2023). Biasanya, model sederhana ini berupa pohon keputusan dengan kedalaman rendah. Untuk Gambaran prosesnya dapat dilihat melalui **Gambar 1**.



Gambar 1. Proses Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) bekerja dengan membangun model secara bertahap untuk memprediksi target dengan akurasi tinggi. Proses dimulai dengan model awal sederhana yang seringkali hanya memprediksi rata-rata target. Setelah itu, GBM menghitung *residual* atau selisih antara nilai aktual dan prediksi model awal. Model baru dilatih untuk memprediksi residual ini, sehingga secara bertahap memperbaiki kesalahan prediksi dari model sebelumnya. Setiap model berikutnya ditambahkan ke model keseluruhan dengan kontribusi yang dikontrol oleh parameter *learning rate*, yang mengatur seberapa besar perbaikan yang dilakukan oleh setiap model tambahan. Iterasi ini terus berlanjut hingga

jumlah model yang diinginkan atau hingga kesalahan prediksi mencapai tingkat minimal yang diharapkan (Guo et al., 2023)(Liu et al., 2023)(Mienye & Sun, 2022). Secara matematis, pada setiap iterasi m , model GBM memperbarui prediksi dengan rumus:

$$F_M(x) = F_{M-1}(x) + \eta \times h_m(x) \tag{2}$$

dimana $F_M(x)$ adalah model pada iterasi ke- m , $F_{M-1}(x)$ adalah model dari iterasi sebelumnya, $h_m(x)$ adalah model baru yang dilatih untuk memprediksi residual, dan η adalah *learning rate* yang mengatur kontribusi model baru dalam mengurangi kesalahan prediksi. Kombinasi iteratif ini menghasilkan prediksi akhir yang akurat dengan memanfaatkan koreksi residual pada setiap langkah.

5. Metode Evaluasi

Metode evaluasi pada penelitian ini menggunakan *confusion matrix* untuk menilai akurasi prediksi harga mobil bekas yang dihasilkan oleh model *Gradient Boosting Machine* (GBM) dan *Random Forest*. *Confusion matrix* memberikan gambaran mengenai distribusi prediksi model dalam berbagai kategori harga yang telah didefinisikan. Untuk membuat kategori ini, harga mobil diubah menjadi kelas diskrit, yaitu rendah, sedang, dan tinggi, dengan menggunakan binning. Data harga aktual dan hasil prediksi model dikonversi ke dalam kategori tersebut sebelum dihitung *confusion matrix*-nya. *Confusion matrix* menyajikan informasi jumlah prediksi yang benar untuk setiap kategori, serta kesalahan di mana harga sebenarnya berbeda dari hasil prediksi. Dari *confusion matrix*, dihitung juga nilai *precision*, *recall*, dan *F1-score* untuk setiap model, yang memberikan evaluasi yang lebih komprehensif mengenai performa prediksi pada setiap kategori harga. Matriks ini serta nilai metrik evaluasi tambahan membantu mengidentifikasi kekuatan dan kelemahan model dalam memprediksi harga mobil bekas secara akurat di berbagai rentang harga.

C. HASIL DAN PEMBAHASAN

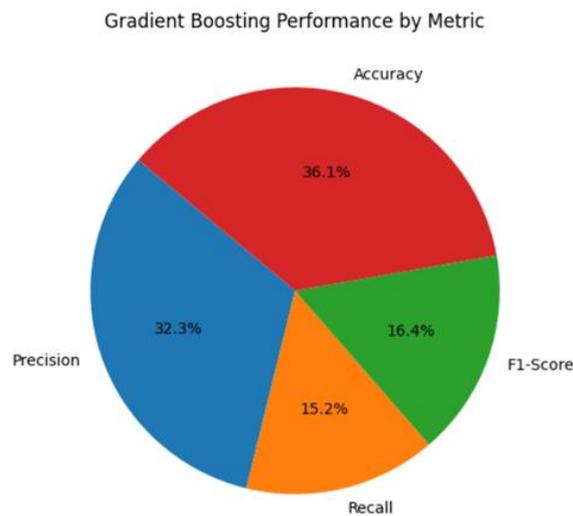
Hasil dari prediksi menggunakan dua metode algoritma ini mendapatkan laporan prediksi yang memberikan informasi lebih tentang performa model.

Tabel 1 Laporan Prediksi dari GBM dan *Random Forest*

	Gradient Boosting			Random Forest		
	R	S	T	R	S	T

Precision	0.91	0.64	1.00	0.98	0.34	1.00
Recall	0.99	0.23	0	0.98	0.31	0
F1-score	0.98	0.34	0	0.98	0.32	0
Accuracy	0.97	0.97	0.96	0.96	0.96	0.96

Dalam penelitian ini, dua algoritma pembelajaran mesin, yaitu *Gradient Boosting Machine (GBM)* dan *Random Forest*, telah digunakan untuk memprediksi harga mobil bekas dan menilai performa prediksi berdasarkan *Precision*, *Recall*, *F1-Score*, dan *Accuracy* pada kategori harga rendah, sedang, dan tinggi. Hasil dari masing-masing algoritma dianalisis untuk memahami kekuatan dan kelemahan model dalam mengklasifikasikan harga berdasarkan tiga kategori harga. **Tabel 1** menyajikan laporan prediksi dari kedua model. Gradient Boosting Machine (GBM) memberikan hasil yang cukup memuaskan dengan nilai *Precision* tinggi untuk kategori rendah dan tinggi, yaitu masing-masing sebesar 0.91 dan 1.00. rinciannya dapat dilihat melalui **Gambar 2**.

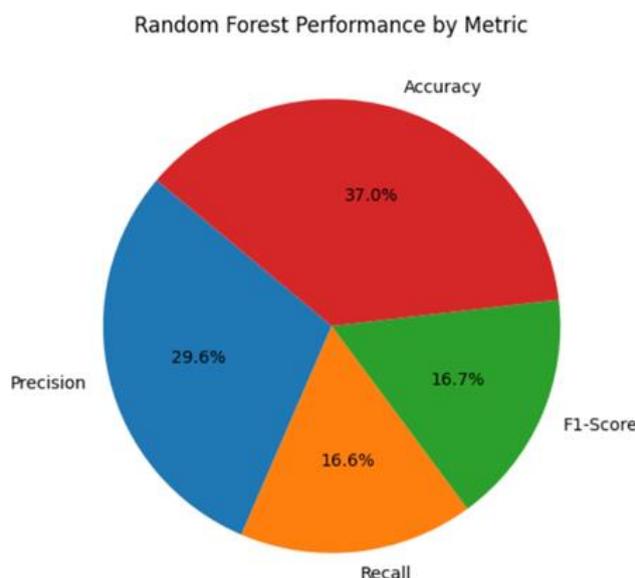


Gambar 2. Hasil rincian dari GBM

Ini menunjukkan bahwa *GBM* dapat secara akurat mengidentifikasi mobil dalam kategori rendah dan tinggi. Namun, pada kategori harga sedang, nilai *Precision* turun menjadi 0.64, yang menandakan bahwa model ini cenderung melakukan kesalahan prediksi pada kategori tersebut. Dari sisi *Recall*, *GBM* memiliki hasil yang sangat tinggi pada kategori rendah dengan nilai 0.99, namun rendah pada kategori sedang dan tinggi (masing-masing

0.23 dan 0.00). Hal ini menunjukkan bahwa meskipun *GBM* mampu mengenali banyak mobil dalam kategori rendah, ia kurang berhasil dalam mendeteksi mobil pada kategori sedang dan tinggi. Nilai *F1-Score* mengikuti pola yang sama, dengan nilai tertinggi pada kategori rendah (0.98), namun rendah pada kategori sedang (0.34) dan 0 pada kategori tinggi. Akurasi keseluruhan *GBM* adalah 0.97, menunjukkan bahwa secara umum model ini dapat memberikan prediksi yang akurat untuk sebagian besar data, terutama pada kategori rendah dan tinggi.

Di sisi lain, *Random Forest*, memiliki performa yang berbeda dibandingkan dengan *GBM*. *Precision* untuk kategori rendah dan tinggi masing-masing adalah 0.98 dan 1.00, untuk rinciannya dapat dilihat melalui **Gambar 3**.

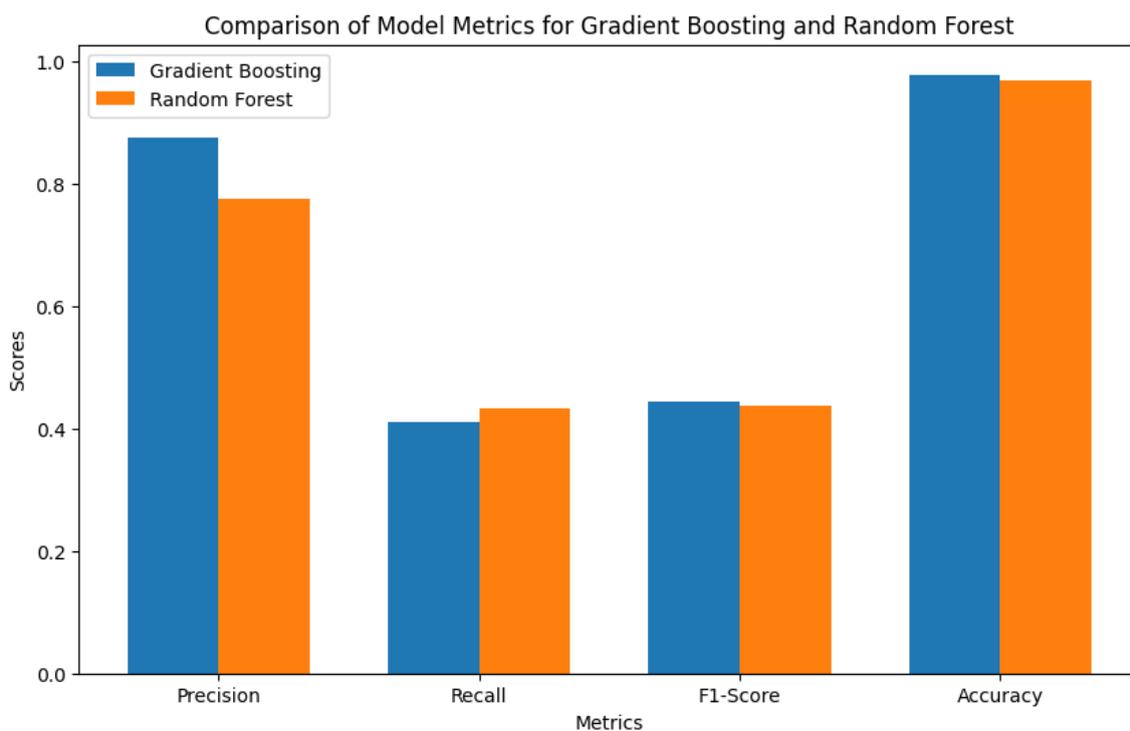


Gambar 3. Hasil rincian dari *Random Forest*

Hal ini menunjukkan performa yang serupa dengan *GBM* pada kategori ini. Namun, *Precision* untuk kategori sedang lebih rendah, hanya 0.34, yang menandakan ketidakmampuan model untuk membedakan kategori harga sedang dengan baik. *Recall* dari *Random Forest* adalah 0.98 untuk kategori rendah, yang cukup tinggi, namun turun menjadi 0.31 dan 0.00 untuk kategori sedang dan tinggi. Hal ini serupa dengan *GBM*, di mana model ini juga menunjukkan kelemahan dalam mengidentifikasi kategori sedang dan tinggi. *F1-Score* mengikuti pola ini, dengan nilai tertinggi pada kategori rendah (0.98) dan sangat rendah pada kategori sedang dan tinggi. Akurasi keseluruhan *Random Forest* adalah 0.96,

sedikit lebih rendah dibandingkan *GBM*, yang menunjukkan performa keseluruhan yang baik, meskipun terdapat kelemahan pada kategori sedang dan tinggi.

Secara keseluruhan, baik *GBM* maupun *Random Forest* menunjukkan performa yang baik dalam memprediksi kategori harga rendah dan tinggi, namun keduanya kesulitan dalam mengidentifikasi kategori harga sedang, dengan nilai *Precision* dan *Recall* yang rendah pada kategori tersebut.



Gambar 4. Gambaran hasil kedua algoritma menggunakan diagram batang

Gambar 4 menunjukkan perbandingan nilai *Precision*, *Recall*, *F1-Score*, dan *Accuracy* dari kedua model dalam bentuk diagram batang, memudahkan visualisasi untuk menganalisis kekuatan dan kelemahan dari masing-masing algoritma, *Gradient Boosting Machine (GBM)* dan *Random Forest*. Secara keseluruhan, kelebihan utama *GBM* adalah kemampuannya untuk mencapai nilai akurasi keseluruhan yang lebih tinggi, yakni 0.97, dengan performa yang baik dalam mengklasifikasikan kategori harga rendah dan tinggi. Namun, *GBM* masih menunjukkan kelemahan dalam mendeteksi kategori harga sedang, sebagaimana tercermin dari nilai *Recall* dan *Precision* yang rendah pada kategori ini. Di sisi lain, *Random Forest* menunjukkan stabilitas yang mirip dengan *GBM*, meskipun dengan

akurasi keseluruhan sedikit lebih rendah pada 0.96. Meskipun *Random Forest* juga mengalami kesulitan dalam membedakan kategori harga sedang, algoritma ini tetap andal dalam mengidentifikasi kategori harga rendah dan tinggi dengan nilai *Precision* dan *Recall* yang tinggi pada kategori tersebut. Dari hasil ini, dapat disimpulkan bahwa baik *GBM* maupun *Random Forest* cocok untuk aplikasi yang membutuhkan pengklasifikasian harga dengan akurasi tinggi pada kategori harga rendah dan tinggi. Namun, jika diperlukan klasifikasi yang lebih akurat pada kategori harga sedang, perbaikan lebih lanjut pada kedua model ini atau penyesuaian parameter tambahan mungkin diperlukan.

D. KESIMPULAN

Penelitian ini menunjukkan bahwa *Gradient Boosting Machine (GBM)* dan *Random Forest* merupakan algoritma yang efektif dalam memprediksi harga mobil bekas, khususnya pada kategori harga rendah dan tinggi. Kedua model menunjukkan performa yang baik, dengan *GBM* mencatatkan akurasi keseluruhan sebesar 0,97 dan *Random Forest* sedikit di bawahnya dengan akurasi 0,96. Meskipun demikian, kedua model mengalami kesulitan dalam mengklasifikasikan harga pada kategori sedang, yang terlihat dari nilai *Precision* dan *Recall* yang rendah pada kategori tersebut. Hal ini mengindikasikan bahwa model mungkin kurang sensitif terhadap variasi harga di rentang sedang, sehingga menghasilkan prediksi yang kurang akurat pada kategori ini. Kendati *GBM* menunjukkan performa yang sedikit lebih unggul dibandingkan *Random Forest*, kedua model ini memiliki kelebihan dalam mengidentifikasi kategori harga yang lebih jelas, seperti rendah dan tinggi, yang sangat berguna dalam menyediakan rekomendasi harga bagi konsumen dan pelaku industri.

Sebagai langkah pengembangan, beberapa strategi dapat dilakukan untuk meningkatkan performa prediksi pada kategori harga sedang. Pertama, penyesuaian lebih lanjut terhadap *hyperparameter* masing-masing algoritma, seperti *learning rate*, *max depth*, dan *n_estimators* pada *GBM* serta *max_features* dan *n_estimators* pada *Random Forest*, dapat membantu meningkatkan akurasi secara keseluruhan. Kedua, penggunaan metode *ensemble* yang lebih kompleks, seperti *stacking* atau *blending* dengan algoritma tambahan seperti *XGBoost*, berpotensi memperbaiki performa pada kategori yang lebih sulit diprediksi. Ketiga, menambah data yang lebih variatif—misalnya dari wilayah berbeda atau tipe mobil yang lebih beragam—dapat membantu model mengenali pola yang lebih representatif dari

pasar mobil bekas. Terakhir, menambahkan fitur tambahan, seperti usia pemilik atau riwayat perawatan, dapat memberikan konteks lebih dalam analisis harga. Langkah-langkah ini diharapkan dapat menghasilkan model prediksi yang lebih akurat dan adaptif, khususnya dalam mendukung pengambilan keputusan konsumen dan bisnis di pasar mobil bekas.

DAFTAR PUSTAKA

- Baldanullah, N. I. S., Mulyarizki, N., Permatasari, I., Naufal, I. P., & Pratama, D. C. (2023). Parallel Processing Pada Pemodelan Machine Learning Menggunakan Random Forest. *Journal of Informatics and Advanced Computing (JIAC)*, 4(1).
- Bharambe, Prof. P., Bagul, B., Dandekar, S., & Ingle, P. (2022). Used Car Price Prediction using Different Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, 10(4). <https://doi.org/10.22214/ijraset.2022.41300>
- Guo, J., Yun, S., Meng, Y., He, N., Ye, D., Zhao, Z., Jia, L., & Yang, L. (2023). Prediction of heating and cooling loads based on light gradient boosting machine algorithms. *Building and Environment*, 236. <https://doi.org/10.1016/j.buildenv.2023.110252>
- Hankar, M., Birjali, M., & Beni-Hssane, A. (2022). Used Car Price Prediction using Machine Learning: A Case Study. *11th International Symposium on Signal, Image, Video and Communications, ISIVC 2022 - Conference Proceedings*. <https://doi.org/10.1109/ISIVC54825.2022.9800719>
- Kriswantara, B., & Sadikin, R. (2022). Used Car Price Prediction with Random Forest Regressor Model. *Journal of Information Systems, Informatics and Computing Issue Period*, 6(1).
- Lin, X., & Hashim, S. (2023). Research on an Intelligent Evaluation System for Used Car Prices Based on Improved Machine Learning Algorithms. *Journal of Network Intelligence*, 8(2).
- Liu, J., Li, C., Ouyang, P., Liu, J., & Wu, C. (2023). Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. *Journal of Forecasting*, 42(5). <https://doi.org/10.1002/for.2931>
- Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms,

- Applications, and Prospects. In *IEEE Access* (Vol. 10).
<https://doi.org/10.1109/ACCESS.2022.3207287>
- Omotehinwa, T. O., Oyewola, D. O., & Dada, E. G. (2023). A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. *Healthcare Analytics*, 4. <https://doi.org/10.1016/j.health.2023.100218>
- Sai, M. J., Chettri, P., Panigrahi, R., Garg, A., Bhoi, A. K., & Barsocchi, P. (2023). An Ensemble of Light Gradient Boosting Machine and Adaptive Boosting for Prediction of Type-2 Diabetes. *International Journal of Computational Intelligence Systems*, 16(1). <https://doi.org/10.1007/s44196-023-00184-y>
- Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, 20(1). <https://doi.org/10.1177/1536867X20909688>
- Suliztia, M. L. (2020). Penerapan Analisis Random Forest Pada Prototype Sistem Prediksi Harga Kamera Bekas Menggunakan Flask. *Fakultas Matematika Dan Ilmu Pengetahuan Alam*.
- Varshitha, J., Jahnavi, K., & Lakshmi, C. (2022). Prediction Of Used Car Prices Using Artificial Neural Networks And Machine Learning. *2022 International Conference on Computer Communication and Informatics, ICCCI 2022*. <https://doi.org/10.1109/ICCCI54379.2022.9740817>
- Wanli Sitorus, Y., Sukarno, P., Mandala, S., Informatika, F., & Telkom, U. (2021). Analisis Deteksi Malware Android menggunakan metode Support Vector Machine & Random Forest. *E-Proceeding of Engineering*, 8(6).
- Yadav, A., Kumar, E., & Yadav, P. K. (2021). Object detection and used car price predicting analysis system (UCPAS) using machine learning technique. *Linguistics and Culture Review*, 5(S2). <https://doi.org/10.21744/lingcure.v5ns2.1660>
- Zhu, Y. (2023). Prediction of the price of used cars based on machine learning algorithms. *Applied and Computational Engineering*, 6(1). <https://doi.org/10.54254/2755-2721/6/20230917>