

IMPLEMENTASI *GRADIENT BOOSTING MACHINES* UNTUK PREDIKSI HARGA RUMAH PADA JAKARTA SELATAN

Fauzan Asyraf Rangkuti¹, Khairunnisa², Siti Sundari^{*3}

^{1,2,3}Teknik Informatika, Fakultas Teknik dan Komputer, Universitas Harapan Medan, Medan, Indonesia

Email: [1fauzanasyrif2000@gmail.com](mailto:fauzanasyrif2000@gmail.com), [2khairunnisajv2@gmail.com](mailto:khairunnisajv2@gmail.com), [3sundaristth@gmail.com](mailto:sundaristth@gmail.com)

(Diterima : 22 April 2025, Direvisi : 10 Mei 2025, Disetujui : 20 Mei 2025)

Abstrak

Prediksi harga properti menjadi salah satu tantangan penting dalam bidang real estat, khususnya di wilayah perkotaan seperti Jakarta Selatan yang mengalami perkembangan pesat. Informasi mengenai estimasi harga rumah sangat dibutuhkan oleh pembeli, penjual, maupun investor untuk pengambilan keputusan yang lebih akurat. Penelitian ini bertujuan untuk membangun model prediksi harga rumah dengan menerapkan metode *Gradient Boosting Machine* (GBM), sebuah teknik *machine learning* yang menggabungkan sejumlah *decision tree* secara iteratif untuk meningkatkan akurasi prediksi. Dataset yang digunakan mencakup beberapa variabel penting, antara lain luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan garasi. Berdasarkan hasil pelatihan dan pengujian model, diketahui bahwa luas tanah dan luas bangunan memiliki pengaruh paling dominan terhadap harga rumah. Evaluasi performa model menunjukkan nilai *mean squared error* (MSE) sebesar $1.6e+19$ dan *mean absolute error* (MAE) sekitar 30 miliar rupiah, yang menunjukkan masih adanya tingkat kesalahan prediksi yang signifikan, terutama pada rumah dengan harga sangat tinggi. Selain itu, analisis visual terhadap distribusi galat dan *residual plot* mengungkapkan adanya sejumlah *outlier* yang perlu diperhatikan. Oleh karena itu, diperlukan optimasi lanjutan pada model agar mampu menangkap pola data yang lebih kompleks dan menghasilkan prediksi yang lebih presisi.

Kata kunci : *gradient boosting machines*, harga rumah, luas bangunan, luas tanah.

IMPLEMENTATION OF *GRADIENT BOOSTING MACHINES* FOR HOUSE PRICE PREDICTION IN SOUTH JAKARTA

Abstract

House price prediction is one of the major challenges in the real estate sector, especially in urban areas such as South Jakarta, which is experiencing rapid development. Information regarding estimated house prices is highly needed by buyers, sellers, and investors to support more accurate decision-making. This study aims to build a house price prediction model by applying the *Gradient Boosting Machine* (GBM) method, a machine learning technique that combines multiple *decision trees* iteratively to improve prediction accuracy. The dataset used includes several important variables, such as land area, building area, number of bedrooms, number of bathrooms, and garage. Based on the training and testing results, it was found that land area and building area have the most dominant influence on house prices. The model's performance evaluation showed a *mean squared error* (MSE) of $1.6e+19$ and a *mean absolute error* (MAE) of around 30 billion rupiah, indicating a significant level of prediction error, particularly for high-priced houses. Additionally, visual analysis of the error distribution and *residual plot* revealed the presence of several *outliers* that need to be addressed. Therefore, further optimization of the model is required to capture more complex data patterns and produce more precise predictions.

Keywords: *building area, gradient boosting machines, house price, land area.*

1. PENDAHULUAN

Jakarta Selatan dikenal sebagai salah satu wilayah dengan perkembangan properti yang pesat di Indonesia. Harga rumah di kawasan ini sangat dipengaruhi oleh berbagai faktor seperti lokasi, aksesibilitas, fasilitas umum, hingga tren permintaan di pasar. Tingginya minat masyarakat untuk memiliki hunian di Jakarta Selatan menjadikan wilayah ini sebagai pusat perhatian dalam dunia properti. Dengan berbagai macam proyek perumahan yang terus

bermunculan, harga rumah di daerah ini cenderung meningkat dari tahun ke tahun, membuat prediksi harga rumah menjadi salah satu informasi penting bagi pembeli maupun investor properti [1].

Namun, permasalahan yang muncul adalah harga rumah sering kali sulit diprediksi dengan akurat. Fluktuasi pasar, kebijakan pemerintah, dan faktor eksternal lainnya membuat harga rumah cenderung tidak stabil. Selain itu, kompleksitas dari berbagai variabel yang memengaruhi harga seperti luas tanah, tipe bangunan, jarak dari pusat kota, serta fasilitas publik menjadikan tugas prediksi harga rumah semakin sulit. Alhasil, tanpa alat analisis yang tepat, pelaku pasar properti dan masyarakat umum sering kali kesulitan dalam membuat keputusan berdasarkan proyeksi harga yang akurat.

Sebagai solusi untuk permasalahan ini, *Gradient Boosting Machines* (GBM) dapat diterapkan sebagai salah satu metode *Machine learning* yang efektif untuk prediksi harga rumah. GBM bekerja dengan membangun model prediksi dari serangkaian pohon keputusan yang saling menguatkan satu sama lain. Melalui pendekatan ini, model dapat mempelajari pola kompleks dari data dan menghasilkan prediksi yang lebih akurat. Dengan implementasi GBM, variabel-variabel kompleks yang memengaruhi harga rumah di Jakarta Selatan dapat dianalisis secara lebih mendalam, sehingga dapat membantu pelaku pasar dalam membuat keputusan yang lebih tepat berdasarkan hasil prediksi yang lebih akurat.

Penelitian terdahulu yang berkaitan dilakukan oleh [2] Hasil penelitian dari jurnal menunjukkan bahwa penerapan algoritma *Gradient Boosting* dan *Linear Regression* digunakan untuk memprediksi harga mobil bekas. Penelitian ini menggunakan data sekunder yang diambil dari situs online, meliputi variabel seperti tahun produksi, harga, jenis transmisi, tipe bahan bakar, dan jarak tempuh kendaraan. Melalui pemodelan dengan teknik *Machine learning*, peneliti bertujuan untuk menghasilkan prediksi harga mobil bekas yang lebih akurat guna membantu konsumen dalam menentukan harga yang wajar. Penelitian [3] menerapkan algoritma regresi linear untuk memprediksi harga rumah di Tebet, Jakarta Selatan, menggunakan variabel seperti luas tanah, bangunan, kamar tidur, kamar mandi, dan garasi. Hasil menunjukkan bahwa luas tanah dan bangunan paling berpengaruh, dengan RMSE sekitar 3,3 miliar rupiah. Namun, model masih memiliki error yang besar, tidak membandingkan dengan metode lain, dan tidak dijelaskan jumlah serta variasi data secara rinci, sehingga akurasi dan generalisasi model masih terbatas, [4] penelitian ini membandingkan algoritma *Multiple Linear Regression* dan *Random Forest Regression* untuk memprediksi harga rumah, dengan hasil Random Forest lebih akurat (81,6%). Namun, tidak ada evaluasi lanjutan seperti *cross-validation* dan eksplorasi algoritma lain yang lebih canggih. Penelitian [5] menerapkan metode prediksi penjualan rumah dengan menggunakan metode *Regresi Linier*. Hasil prediksi harga jual rumah dengan menggunakan metode regresi linear dari harga 35,000 dan terprediksi sekitar 31,000 dari harga jual, di tahun sebelumnya. Hasil *Root Mean Squared Error* (RMSE) pada *Performance Regresi Linear* ini menunjukkan : 3326.243 +/- 0.000. Pada [6] menggunakan *Multiple Linear Regression* untuk memprediksi harga rumah yang sesuai spesifikasi dengan mendapatkan hasil dengan menggunakan algoritma regresi dengan nilai akurasi sebesar 66%, sedangkan [7] menggunakan *General Regression Neural Network* untuk prediksi harga rumah, dan hasil yang didapatkan berupa *data actual* dan *data predicted* yang divisualisasikan dengan *line plot*. Dilakukan juga uji akurasi dan uji kinerja pemodelan menggunakan 3 jenis evaluasi dengan hasil dari ketiga jenis evaluasi tersebut adalah 58,72 untuk skor MSE, skor 7,66 untuk RMSE dan 5,99 untuk skor MAE.

Tujuan penelitian ini adalah untuk menerapkan metode *Gradient Boosting Machine* (GBM) dalam memprediksi harga rumah di wilayah Jakarta Selatan dengan mempertimbangkan berbagai variabel seperti luas tanah, luas bangunan, jumlah kamar tidur, dan jumlah kamar mandi. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor yang paling berpengaruh terhadap harga properti serta mengevaluasi kinerja model dalam menghasilkan prediksi harga yang akurat, guna memberikan wawasan yang lebih baik bagi pasar properti di wilayah tersebut.

2. METODE PENELITIAN

2.1. Prediksi

Prediksi adalah suatu proses untuk memperkirakan atau memperhitungkan kejadian yang akan datang berdasarkan data, informasi, atau pola yang ada saat ini. Prediksi dapat diterapkan dalam berbagai bidang, seperti cuaca, ekonomi, olahraga, dan ilmu pengetahuan. Proses prediksi bertujuan untuk memberikan gambaran mengenai kemungkinan masa depan, yang dapat membantu individu atau organisasi dalam mengambil keputusan yang lebih baik dan terinformasi [8]. Dalam dunia modern, prediksi sering kali melibatkan penggunaan model matematika dan algoritma statistik yang didukung oleh teknologi komputasi. Dalam konteks ilmu pengetahuan dan teknologi, prediksi sering didasarkan pada analisis data. Misalnya, dalam ilmu meteorologi, prediksi cuaca dilakukan dengan memanfaatkan data atmosfer yang dikumpulkan dari berbagai sumber, seperti satelit dan stasiun cuaca, untuk memperkirakan kondisi cuaca di masa depan. Demikian pula, dalam dunia bisnis, prediksi penjualan dilakukan dengan menganalisis tren data penjualan historis, pola permintaan, dan kondisi pasar untuk memproyeksikan angka penjualan di periode berikutnya. Dengan kemajuan teknologi, metode prediksi semakin berkembang dan semakin presisi [9]–[11].

2.2. Machine learning

Machine learning adalah cabang dari kecerdasan buatan (*artificial intelligence*) yang fokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit. Dalam *Machine learning*, sistem komputer menggunakan data historis atau pengalaman sebelumnya untuk mengenali pola, menganalisis informasi, dan membuat keputusan atau prediksi secara otomatis. Dengan kata lain, *Machine learning* memungkinkan mesin untuk “belajar” dan meningkatkan kinerja dari waktu ke waktu berdasarkan data yang diberikan. Proses pembelajaran dalam *Machine learning* terjadi melalui model matematis yang dibangun berdasarkan data masukan. Algoritma *Machine learning* akan mengidentifikasi pola atau hubungan dalam data tersebut, dan menggunakan pola-pola ini untuk memprediksi hasil pada data baru yang belum pernah dilihat sebelumnya. Salah satu contoh sederhana adalah model prediksi harga rumah, di mana algoritma dilatih dengan data historis tentang harga rumah berdasarkan faktor-faktor seperti lokasi, ukuran, dan jumlah kamar untuk memprediksi harga rumah di masa depan [12], [13].

2.3. Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) adalah salah satu algoritma *Machine learning* yang berbasis ensemble, yang digunakan untuk tugas-tugas prediksi seperti klasifikasi dan regresi. Algoritma ini bekerja dengan membangun model secara bertahap, di mana setiap model baru berusaha memperbaiki kesalahan prediksi yang dibuat oleh model sebelumnya. Konsep utama di balik GBM adalah menggabungkan beberapa model sederhana (biasanya berupa pohon keputusan) untuk menciptakan model yang kuat, di mana setiap model baru dioptimalkan untuk mengurangi kesalahan yang tersisa dari model sebelumnya. Pada intinya, GBM memanfaatkan metode boosting yang bekerja dengan menggabungkan beberapa model lemah secara berurutan, sehingga masing-masing model berkontribusi pada perbaikan keseluruhan. Setiap model baru yang ditambahkan ke rangkaian ini berfokus untuk mengurangi kesalahan residu (*residual errors*) dari model sebelumnya, yang dilakukan dengan meminimalkan fungsi loss (*loss function*) menggunakan teknik gradient descent [14]. Secara matematis, GBM dapat dijelaskan dalam beberapa langkah sebagai berikut:

- a. Fungsi Prediksi Awal:
- b. Pada awalnya, GBM membuat prediksi awal dengan model dasar (biasanya pohon keputusan sederhana). Fungsi prediksi awal ini biasanya disimbolkan sebagai $F_0(x)$ dan sering kali berupa prediksi rata-rata pada kasus regresi:

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (1)$$

Di mana $\sum_{i=1}^n L(y_i, \gamma)$ adalah fungsi loss (kerugian) untuk observasi ke-iii dengan target y_i, γ

2.4. Python

Python adalah bahasa pemrograman tingkat tinggi yang diciptakan oleh Guido van Rossum dan pertama kali dirilis pada tahun 1991. Python dirancang dengan fokus pada keterbacaan kode dan kemudahan penulisan, menjadikannya salah satu bahasa pemrograman yang paling populer di dunia saat ini. Salah satu alasan utama di balik popularitas Python adalah sintaksnya yang sederhana dan mudah dipahami, sehingga cocok untuk pemula maupun programmer berpengalaman. Python menggunakan indentasi untuk menandai blok kode, bukan tanda kurung kurawal atau kata kunci seperti yang digunakan dalam banyak bahasa lain. Python adalah bahasa pemrograman serbaguna yang dapat digunakan untuk berbagai tujuan, mulai dari pengembangan web, aplikasi desktop, hingga data science dan *Machine learning*. Bahasa ini bersifat *interpreted* dan *dynamically typed*, yang berarti kode Python dijalankan secara langsung oleh interpreter tanpa harus melalui tahap kompilasi seperti bahasa seperti C++ atau Java. Python mendukung berbagai paradigma pemrograman, termasuk pemrograman berorientasi objek (OOP), pemrograman prosedural, dan pemrograman fungsional, yang memberikan fleksibilitas dalam pendekatan pengembangan perangkat lunak [15].

3. HASIL DAN PEMBAHASAN

Penerapan metode *Gradient Boosting Machines* (GBM) dalam penelitian ini bertujuan untuk membangun model prediksi yang mampu memperkirakan harga rumah secara akurat berdasarkan berbagai variabel yang memengaruhi, seperti luas tanah, luas bangunan, jumlah kamar, dan fasilitas lainnya. GBM dipilih karena kemampuannya dalam menangani data yang kompleks dan memberikan hasil prediksi yang akurat dengan menggabungkan beberapa model prediksi yang lebih lemah menjadi model yang lebih kuat. Dengan metode ini, diharapkan model dapat mengenali pola dan hubungan antara variabel-variabel properti yang beragam, sehingga mampu memberikan estimasi harga rumah yang lebih tepat dan relevan.

```
# Langkah 2: Menerapkan Gradient Boosting Machines (GBM)
# Inisialisasi model GBM
gbm = GradientBoostingRegressor()

# Melatih model dengan data latih
gbm.fit(X_train, y_train)

# Langkah 3: Memprediksi harga pada data uji
y_pred = gbm.predict(X_test)
```

Gambar 1. Implementasi Metode GBM

Gambar tersebut menunjukkan proses penerapan metode *Gradient Boosting Machines* (GBM) untuk prediksi harga rumah. Pada langkah pertama, model GBM diinisialisasi menggunakan fungsi *GradientBoostingRegressor()* dari pustaka *scikit-learn*. Inisialisasi ini menciptakan sebuah model yang siap dilatih dengan data latih untuk mempelajari hubungan antara fitur-fitur properti (seperti luas tanah, luas bangunan, jumlah kamar, dll.) dengan harga rumah. Pada langkah kedua, model dilatih menggunakan data latih (*X_train* dan *y_train*). Fungsi *gbm.fit(X_train, y_train)* akan membuat model GBM mempelajari pola dari data latih tersebut. Model ini secara iteratif membangun serangkaian pohon keputusan (decision trees) yang bertujuan untuk meminimalkan kesalahan prediksi, sehingga mampu memprediksi harga rumah dengan lebih akurat. Setelah model dilatih, langkah ketiga adalah melakukan prediksi harga pada data uji menggunakan fungsi *gbm.predict(X_test)*. Fungsi ini akan menghasilkan prediksi harga (*y_pred*) berdasarkan fitur-fitur yang ada di data uji (*X_test*). Prediksi ini kemudian dapat dibandingkan dengan harga aktual untuk mengevaluasi performa model. Secara keseluruhan, langkah-langkah ini menggambarkan proses membangun dan menguji model prediksi harga rumah menggunakan metode *Gradient Boosting Machines*.

```
# Menghitung Mean Squared Error (MSE), Mean Absolute Error (MAE) dan R-Squared untuk evaluasi
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"Mean Absolute Error: {mae}")
print(f"R-Squared: {r2}")

# Menyimpan hasil prediksi dalam DataFrame baru
test_data['Prediksi_Harga'] = y_pred
```

Gambar 2. Perhitungan MSE

Gambar tersebut menunjukkan proses evaluasi kinerja model prediksi dan penyimpanan hasil prediksi ke dalam DataFrame baru. Langkah pertama adalah menghitung beberapa metrik evaluasi utama, yaitu *Mean Squared Error* (MSE), *Mean Absolute Error* (MAE), dan *R-Squared* (R2). MSE mengukur rata-rata kuadrat dari kesalahan prediksi, di mana nilai yang lebih rendah menunjukkan akurasi prediksi yang lebih baik. MAE, di sisi lain, mengukur rata-rata dari kesalahan absolut antara harga aktual dan harga prediksi, yang memberikan gambaran tentang seberapa besar kesalahan rata-rata yang terjadi dalam satuan harga rumah. Sementara itu, *R-Squared* (R2) menunjukkan seberapa baik variabel-variabel fitur dapat menjelaskan variasi dalam harga rumah. Nilai R2 yang mendekati 1 menunjukkan bahwa model memiliki kemampuan yang baik dalam memprediksi harga. Setelah menghitung metrik-metrik tersebut, hasilnya ditampilkan agar pengguna dapat mengevaluasi performa model secara langsung. Langkah selanjutnya adalah menyimpan hasil prediksi harga ke dalam DataFrame *test_data* dengan menambahkan kolom baru bernama 'Prediksi_Harga'. Ini memungkinkan kita untuk melihat harga yang diprediksi oleh model berdampingan dengan harga aktual pada data uji. Penyimpanan hasil prediksi ini penting untuk analisis lebih lanjut serta untuk membuat laporan yang lebih komprehensif. Dengan demikian, proses ini mencakup evaluasi lengkap dari performa model dan penyimpanan hasil prediksi untuk referensi dan analisis lebih lanjut.

```
Mean Squared Error: 1.6002171142209343e+19
Mean Absolute Error: 3032601441.1220183
R-Squared: 0.6572149166557151
```

Gambar 3. Hasil Perhitungan MSE

Gambar tersebut menampilkan hasil evaluasi kinerja model prediksi harga rumah menggunakan beberapa metrik utama. *Mean Squared Error* (MSE) sebesar $1.6002171142209343e+19$ menunjukkan bahwa rata-rata kuadrat kesalahan prediksi sangat besar, yang berarti model cenderung membuat kesalahan yang signifikan dalam

memprediksi harga rumah. Nilai MSE yang besar ini bisa disebabkan oleh adanya outlier atau variabilitas yang tinggi dalam data harga rumah. *Mean Absolute Error* (MAE) sebesar 30.326.014.411,12 menunjukkan bahwa rata-rata kesalahan absolut dalam prediksi harga mencapai sekitar 30 miliar rupiah. Ini menandakan bahwa model secara rata-rata meleset sekitar 30 miliar rupiah dari harga rumah yang sebenarnya, yang merupakan kesalahan prediksi yang cukup besar dan mengindikasikan bahwa model perlu dioptimalkan lebih lanjut. *R-Squared* (R2) yang sebesar 0,657 menunjukkan bahwa sekitar 65,7% variasi dalam harga rumah dapat dijelaskan oleh variabel-variabel fitur yang digunakan dalam model. Meskipun nilai ini menunjukkan model dapat menjelaskan sebagian besar variasi, namun masih ada sekitar 34,3% variasi yang tidak dapat dijelaskan oleh model, yang berarti masih ada ruang untuk perbaikan. Secara keseluruhan, hasil evaluasi ini mengindikasikan bahwa performa model masih bisa ditingkatkan dengan melakukan optimasi lebih lanjut, seperti penghapusan outlier, penambahan fitur yang lebih relevan, atau penyesuaian parameter model.

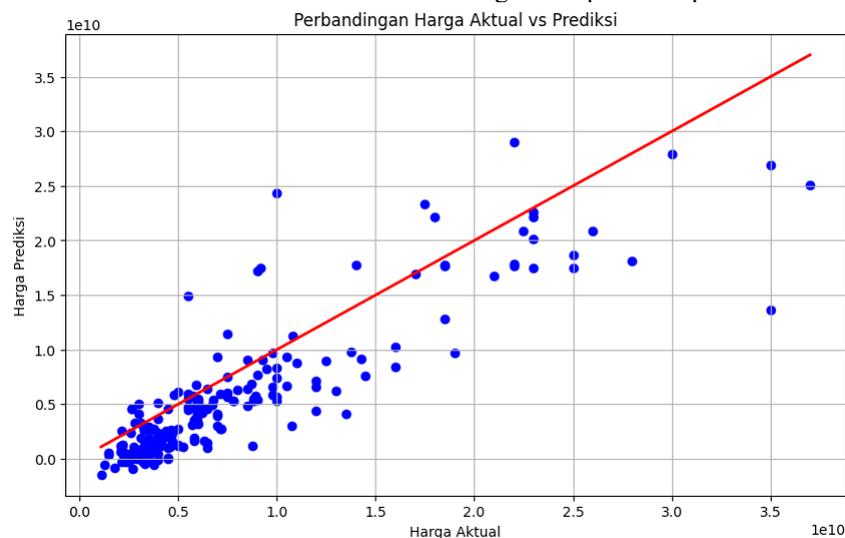
Data Uji beserta Hasil Prediksi:

NO	NAMA RUMAH	LB	LT	KT	KM	\
0	630 Rumah Bagus Siap Huni JUAL CEPAT ! Tebet Utara	300	300	5	4	
1	789 Dijual Rumah, Siap Huni,Strategis, Di Tebet Ja...	148	150	4	3	
2	685 Rumah Cantik Minimalis Di Tebet	300	200	5	4	
3	517 For Sale British Tuscan House At Tebet Mas, Ja...	500	720	5	5	
4	530 Dijual Rumah Tebet Raya Lokasi Premium Area Ko...	300	520	4	3	

GRS	HARGA	Prediksi_Harga
0	2 8900000000	5.786953e+09
1	3 6500000000	1.051816e+09
2	4 6500000000	6.368346e+09
3	2 3700000000	2.504976e+10
4	3 1850000000	1.775967e+10

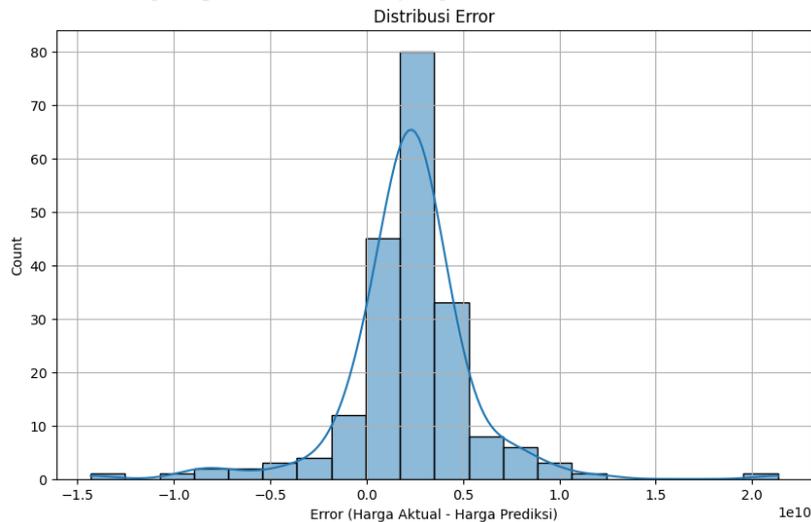
Gambar 4. Hasil Prediksi

Gambar tersebut menampilkan data uji beserta hasil prediksi harga rumah yang dihasilkan oleh model *Gradient Boosting Machines* (GBM). Setiap baris menunjukkan informasi detail mengenai properti, termasuk kolom NO yang berisi nomor urut, NAMA RUMAH yang berisi deskripsi singkat properti, LB (Luas Bangunan) dan LT (Luas Tanah) dalam meter persegi, serta jumlah KT (Kamar Tidur) dan KM (Kamar Mandi). Selain itu, kolom GRS menunjukkan jumlah garasi yang tersedia. Kolom HARGA berisi harga rumah sebenarnya dari data uji, sementara kolom Prediksi_Harga menunjukkan hasil prediksi harga yang dihasilkan oleh model. Misalnya, pada baris pertama, rumah dengan deskripsi "Rumah Bagus Siap Huni JUAL CEPAT! Tebet Utara" memiliki harga asli 8,9 miliar rupiah, sementara prediksi harga yang dihasilkan oleh model adalah sekitar 5,79 miliar rupiah. Pada baris kedua, terdapat rumah dengan harga asli 6,5 miliar rupiah dan prediksi harga model sekitar 1,05 miliar rupiah, yang menunjukkan adanya perbedaan signifikan antara harga aktual dan prediksi. Secara umum, tabel ini menunjukkan bahwa terdapat kesenjangan antara harga sebenarnya dan hasil prediksi, yang mengindikasikan bahwa model masih perlu dioptimalkan lebih lanjut untuk meningkatkan akurasi prediksinya. Perbedaan yang cukup besar antara harga aktual dan harga prediksi dapat disebabkan oleh berbagai faktor, seperti fitur yang belum dioptimalkan, adanya outlier, atau kompleksitas hubungan antar variabel yang belum sepenuhnya ditangkap oleh model. Analisis lebih lanjut diperlukan untuk memahami kesalahan model dan meningkatkan performa prediksi.



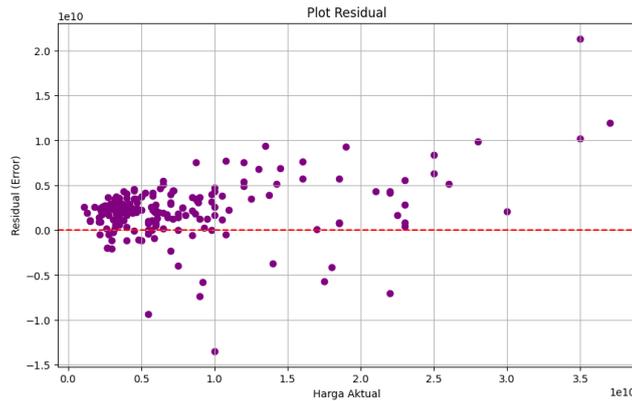
Gambar 5. Perbandingan Harga Aktual dan Prediksi

Gambar tersebut menunjukkan scatter plot yang membandingkan harga rumah aktual dengan harga prediksi yang dihasilkan oleh model. Setiap titik biru di plot ini mewakili satu rumah dalam data uji, dengan sumbu horizontal (x) menunjukkan harga aktual rumah dan sumbu vertikal (y) menunjukkan harga yang diprediksi oleh model. Garis merah diagonal merupakan garis referensi yang menunjukkan prediksi sempurna, di mana harga prediksi sama dengan harga aktual. Jika model prediksi sempurna, semua titik biru akan berada tepat di garis merah ini. Dari plot ini, terlihat bahwa sebagian besar titik tidak berada di garis merah, yang menunjukkan bahwa ada perbedaan antara harga yang diprediksi model dengan harga sebenarnya. Beberapa titik berada jauh dari garis, yang menandakan bahwa model melakukan kesalahan prediksi yang cukup besar untuk beberapa rumah. Hal ini menunjukkan bahwa meskipun model dapat menangkap beberapa pola dasar, masih ada ruang yang signifikan untuk peningkatan akurasi prediksi. Kecenderungan titik-titik yang menyebar di sekitar garis merah menunjukkan bahwa model mampu menangkap tren umum dari data, tetapi mungkin kurang akurat dalam memprediksi harga rumah dengan harga yang sangat tinggi atau sangat rendah. Hal ini dapat disebabkan oleh berbagai faktor, seperti fitur yang tidak cukup kuat untuk menjelaskan variabilitas harga, keberadaan outlier, atau kompleksitas hubungan antar variabel yang belum sepenuhnya ditangkap oleh model. Untuk meningkatkan performa, perlu dilakukan optimasi lebih lanjut pada model dan mungkin penambahan fitur yang lebih relevan.



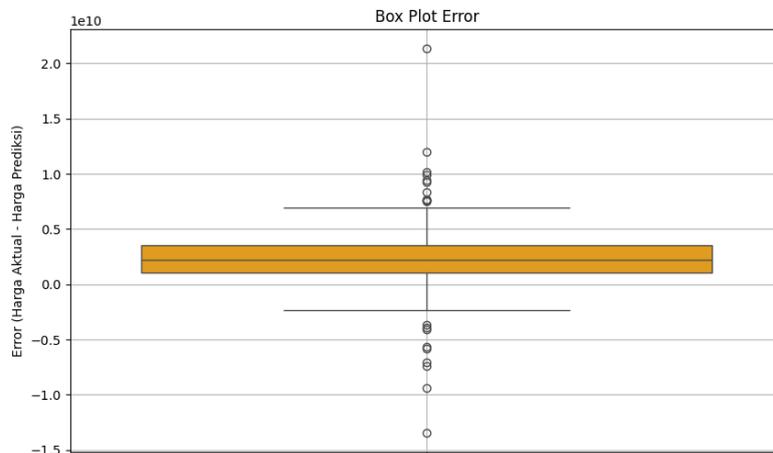
Gambar 6. Distribusi Error

Gambar tersebut menampilkan histogram distribusi error antara harga rumah aktual dan harga prediksi yang dihasilkan oleh model. Sumbu horizontal (x) menunjukkan nilai error, yaitu selisih antara harga aktual dan harga prediksi (Harga Aktual - Harga Prediksi), sedangkan sumbu vertikal (y) menunjukkan jumlah rumah (*count*) yang memiliki nilai error tertentu. Dari histogram ini, terlihat bahwa sebagian besar error berada di sekitar nol, yang menunjukkan bahwa model memiliki tingkat kesalahan yang cukup rendah untuk banyak rumah. Kurva distribusi yang overlay (garis biru) menunjukkan bentuk distribusi error secara keseluruhan, yang mendekati distribusi normal dengan puncak di sekitar nol. Ini menunjukkan bahwa sebagian besar prediksi model mendekati harga aktual rumah. Namun, terdapat juga beberapa error yang signifikan, yang terlihat dari adanya bar di kedua ujung histogram, baik di sisi negatif maupun positif. Error negatif berarti model memprediksi harga lebih tinggi dari harga sebenarnya, sedangkan error positif berarti model memprediksi harga lebih rendah dari harga sebenarnya. Distribusi error yang relatif simetris ini menunjukkan bahwa model tidak bias terhadap overestimasi atau underestimasi harga rumah. Namun, ada beberapa outlier dengan nilai error yang sangat besar, yang mengindikasikan bahwa model mengalami kesulitan dalam memprediksi beberapa properti dengan harga yang sangat tinggi atau sangat rendah. Hal ini bisa disebabkan oleh variabel-variabel tertentu yang tidak sepenuhnya ditangkap oleh model, atau karakteristik unik dari properti-properti tersebut yang tidak terwakili dengan baik dalam data latih. Untuk meningkatkan performa model, analisis lebih lanjut diperlukan untuk memahami penyebab kesalahan besar ini dan melakukan optimasi model lebih lanjut.



Gambar 7. Plot Residual

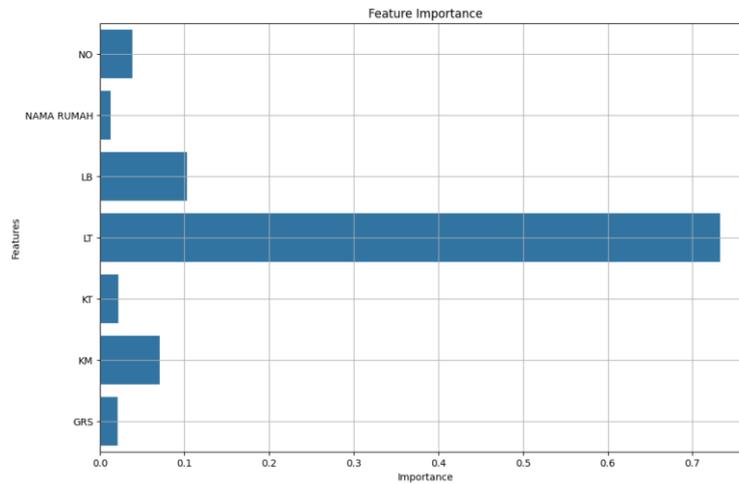
Gambar tersebut menunjukkan plot residual yang menggambarkan perbedaan antara harga aktual dan harga prediksi (*residual/error*) terhadap harga aktual rumah. Sumbu horizontal (x) merepresentasikan harga aktual rumah, sedangkan sumbu vertikal (y) menunjukkan nilai residual atau error, yaitu selisih antara harga aktual dan harga prediksi (Harga Aktual - Harga Prediksi). Garis merah horizontal di titik nol menunjukkan titik di mana prediksi model tepat sama dengan harga aktual, yang berarti tidak ada error. Dari plot ini, kita dapat melihat bahwa sebagian besar titik residual terkonsentrasi di sekitar garis nol, yang menunjukkan bahwa untuk banyak rumah, model mampu memprediksi harga dengan cukup akurat. Namun, terdapat beberapa titik yang jauh dari garis merah, yang menunjukkan adanya error prediksi yang cukup besar untuk rumah-rumah tersebut. Ini berarti model membuat kesalahan yang lebih besar untuk beberapa properti, baik itu overestimasi (titik di atas garis merah) maupun underestimasi (titik di bawah garis merah). Distribusi titik-titik di sepanjang sumbu horizontal menunjukkan bahwa error tidak sepenuhnya acak. Misalnya, ada kecenderungan bahwa rumah dengan harga yang lebih tinggi (lebih dari 2 miliar rupiah) memiliki error prediksi yang lebih besar, yang dapat menunjukkan bahwa model mengalami kesulitan dalam memprediksi harga properti yang mahal dengan tepat. Ini mungkin disebabkan oleh data yang kurang representatif untuk properti dengan harga tinggi, atau karena variabel-variabel yang digunakan dalam model tidak cukup kuat untuk menjelaskan harga properti yang sangat mahal.



Gambar 8. Box Plot Error

Gambar tersebut menampilkan *Box Plot* yang menggambarkan distribusi error prediksi dari model, yaitu selisih antara harga rumah aktual dan harga prediksi (Harga Aktual - Harga Prediksi). Sumbu vertikal menunjukkan nilai error, sedangkan *Box Plot* memberikan gambaran tentang persebaran dan distribusi error tersebut. Pada *Box Plot* ini, kotak (*box*) berwarna oranye merepresentasikan rentang interkuartil (*interquartile range*, IQR), yang menunjukkan distribusi error antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Garis di dalam kotak menunjukkan nilai median error. Median yang dekat dengan nol menunjukkan bahwa setengah dari error prediksi berada di bawah dan setengahnya lagi berada di atas nol, yang mengindikasikan tidak adanya bias signifikan dalam prediksi model secara keseluruhan. Garis-garis vertikal yang memanjang dari kotak disebut *Box Plot*, yang menunjukkan rentang distribusi error yang berada di luar rentang interkuartil. Titik-titik di luar *Box Plot* adalah *outliers*, yaitu error yang jauh lebih besar atau lebih kecil dari distribusi utama. Dari gambar ini, terlihat beberapa outlier di kedua sisi, baik di bawah maupun di atas, yang menunjukkan adanya beberapa prediksi model yang sangat meleset dari harga aktual. Outlier di bagian atas menunjukkan kasus-kasus di mana model secara signifikan meremehkan harga rumah (harga prediksi jauh lebih rendah dari harga aktual), sedangkan outlier di bagian bawah menunjukkan kasus di mana model

terlalu tinggi dalam memprediksi harga (harga prediksi jauh lebih tinggi dari harga aktual). Kehadiran outlier ini menandakan bahwa ada beberapa properti yang harga prediksinya sangat jauh dari harga aktual, yang mungkin disebabkan oleh faktor-faktor khusus dari properti tersebut atau kekurangan pada model yang digunakan.



Gambar 9. Feature Importance

Gambar tersebut menunjukkan feature importance plot yang menggambarkan seberapa besar pengaruh masing-masing fitur terhadap prediksi harga rumah yang dilakukan oleh model *Gradient Boosting Machines* (GBM). Sumbu horizontal (x) menunjukkan tingkat kepentingan atau kontribusi relatif dari setiap fitur, sedangkan sumbu vertikal (y) menunjukkan nama-nama fitur yang digunakan dalam model. Dari plot ini, terlihat bahwa fitur LT (Luas Tanah) memiliki pengaruh paling besar terhadap prediksi harga rumah, dengan nilai importance yang jauh lebih tinggi dibandingkan fitur lainnya. Ini menunjukkan bahwa luas tanah merupakan faktor paling signifikan dalam menentukan harga rumah di dataset ini. Fitur LB (Luas Bangunan) juga memberikan kontribusi yang cukup besar, meskipun tidak sebesar luas tanah. Fitur-fitur lain seperti KT (Jumlah Kamar Tidur), KM (Jumlah Kamar Mandi), dan GRS (Garasi) memiliki pengaruh yang relatif lebih kecil, yang mungkin menunjukkan bahwa mereka tidak terlalu menentukan harga dibandingkan dengan luas tanah dan bangunan. Fitur NO dan NAMA RUMAH memiliki importance yang sangat rendah, yang menunjukkan bahwa informasi ini tidak memberikan kontribusi signifikan dalam memprediksi harga rumah. Plot ini memberikan wawasan yang berguna untuk memahami bagaimana model menggunakan fitur-fitur yang tersedia untuk memprediksi harga rumah.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa penerapan metode *Gradient Boosting Machine* (GBM) dalam penelitian ini mampu memberikan gambaran prediktif terhadap harga rumah di wilayah Jakarta Selatan, dengan mempertimbangkan variabel-variabel seperti luas tanah, luas bangunan, jumlah kamar tidur, dan jumlah kamar mandi. Meskipun hasil prediksi tidak sepenuhnya akurat, model ini tetap memberikan representasi yang cukup baik terhadap tren harga properti di wilayah tersebut. Dari hasil analisis *feature importance*, diketahui bahwa luas tanah (LT) dan luas bangunan (LB) merupakan dua fitur yang paling dominan dalam menentukan harga rumah, sementara fitur lainnya seperti jumlah kamar tidur (KT), jumlah kamar mandi (KM), dan garasi (GRS) memiliki pengaruh yang relatif lebih kecil. Evaluasi terhadap performa model menunjukkan bahwa nilai *Mean Squared Error* (MSE) mencapai sekitar $1.6e+19$ dan *Mean Absolute Error* (MAE) sebesar 30 miliar rupiah, yang mengindikasikan masih tingginya tingkat kesalahan prediksi, terutama pada rumah-rumah dengan harga yang sangat tinggi. Selain itu, analisis distribusi error menunjukkan bahwa sebagian besar prediksi berada di sekitar harga aktual, namun terdapat beberapa *outlier* signifikan yang menunjukkan bahwa model kesulitan dalam menangani prediksi harga properti yang berada di luar rentang umum. Temuan ini juga diperkuat oleh plot residual dan scatter plot yang memperlihatkan adanya kesenjangan mencolok antara harga aktual dan hasil prediksi, terutama pada rumah dengan harga sangat tinggi. Oleh karena itu, model perlu dilakukan proses optimasi lebih lanjut agar mampu menangkap pola yang lebih kompleks dalam data dan menghasilkan prediksi yang lebih akurat serta andal.

DAFTAR PUSTAKA

- [1] G. T. Rethabfahisa and E. Ariyanto, "ANALISIS PREFERENSI GENERASI MILENIAL DALAM MEMILIH HUNIAN DI KOTA JAMBI," *J. Ilm. Manajemen, Ekon. Akunt.*, vol. 9, no. 1, pp. 1596–1621, 2025.

- [2] Putra Nasyuli Leriensyah, Lubis Imran, and Marwan Elhanafi Andi, “Penerapan Model *Machine learning* Algoritma Gradient Boosting dan Linear Regression Melakukan Prediksi Harga Kendaraan Bekas Application Of *Machine learning* Models and Gradient Boosting Algorithms Doing Linear Regression Vehicle Price Prediction Used,” *J. Ilmu Komput. dan Sist. Informasi(JIRSI)*, vol. 2, no. 2, pp. 299–310, 2023.
- [3] A. Saiful, “Prediksi Harga Rumah Menggunakan Web Scrapping dan *Machine learning* Dengan Algoritma Linear Regression,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 1, pp. 41–50, 2021, doi: 10.35957/jatisi.v8i1.701.
- [4] C. Haryanto, N. Rahaningsih, and F. Muhammad Basysyar, “Komparasi Algoritma *Machine learning* Dalam Memprediksi Harga Rumah,” *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 1, pp. 533–539, 2023, doi: 10.36040/jati.v7i1.6343.
- [5] G. Khalda Rifdan, N. Rahaningsih, A. Bahtiar, I. Ali, and N. Dienwati Nuris, “Ramalan Penjualan Rumah Menggunakan Algoritma Linear Regresi Di Tebet Jakarta Selatan,” *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 2, pp. 1847–1851, 2024, doi: 10.36040/jati.v8i2.9022.
- [6] M. L. Mu’tashim, T. Muhayat, S. A. Damayanti, H. N. Zaki, and R. Wirawan, “Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan *Multiple Linear Regression*,” *Inform. J. Ilmu Komput.*, vol. 17, no. 3, p. 238, 2021, doi: 10.52958/iftk.v17i3.3635.
- [7] E. F. Rahayuningtyas, F. N. Rahayu, and Y. Azhar, “Prediksi Harga Rumah Menggunakan General Regression Neural Network,” *J. Inform.*, vol. 8, no. 1, pp. 59–66, 2021, doi: 10.31294/ji.v8i1.9036.
- [8] R. M. Sari, *Klasifikasi Data Mining*. Serasi Media Teknologi, 2024.
- [9] D. K. H. Putra, “Analisis Perbandingan Model Prediksi Harga Saham melalui Pendekatan Long-Short Term Memory (LSTM) Dan Autoregressive Integrated Moving Average (ARIMA)(Studi Kasus: Alfamart dan Alfamidi).” Sekolah Tinggi Teknologi Terpadu Nurul Fikri, 2024.
- [10] B. Imran, Zaeniah, Sriasih, S. Erniwati, and Salman, “Data Mining Using a Support Vector Machine , Decision Tree , Logistic Regression and Random Forest for,” *J. Infokum*, vol. 10, no. 2, pp. 792–802, 2022.
- [11] Heliyanti Susana, “Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet,” *J. Ris. Sist. Inf. dan Teknol. Inf.*, vol. 4, no. 1, pp. 1–8, 2022, doi: 10.52005/jursistekni.v4i1.96.
- [12] R. Rosidin, R. Novianti, K. P. Ningsih, D. Haryadi, G. Chrisnawati, and N. Anripa, “Peran Kecerdasan Buatan Dalam Pengembangan Sistem Otomatisasi Proses Bisnis,” *J. Rev. Pendidik. Dan Pengajaran*, vol. 7, no. 3, pp. 9320–9329, 2024.
- [13] Y. Findawati, U. Indahyanti, Y. Rahmawati, and R. Puspitasari, “Sentiment Analysis of Potential Presidential Candidates 2024: A Twitter-Based Study,” *Acad. Open*, vol. 8, no. 1, pp. 6–17, 2023, doi: 10.21070/acopen.8.2023.7138.
- [14] E. Febrywinata, “Pengenalan Dan Klasifikasi Jenis Buah Menggunakan Metode CNN Secara Sederhana Dengan Menggunakan Google Colab,” *Merkurius J. Ris. Sist. Inf. dan Tek. Inform.*, vol. 2, no. 4, pp. 185–193, 2024.
- [15] Y. Dios Azhari, “PEMBUATAN APLIKASI UNTUK MODUL PEMBELAJARAN SISTEM KOMUNIKASI DI POLITEKNIK CALTEX RIAU MENGGUNAKAN SIMULINK.” Politeknik Caltex Riau, 2024.