

OPTIMASI XGBOOST DENGAN GRID SEARCH CROSS VALIDATION UNTUK PENINGKATAN AKURASI PREDIKSI HARGA RUMAH

Muh. Junaidi ¹⁾, Rina Firliana ²⁾, Arie Nugroho ³⁾

^{1,2,3)} Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri
Email: muhjunaidi183@gmail.com¹⁾

Dikirim: 13 Mei 2026; Disetujui: 16 Juni 2026; Dipublikasikan: 31 Juli 2026

ABSTRAK

Harga rumah dipengaruhi oleh berbagai faktor seperti lokasi, luas bangunan, dan fasilitas sehingga prediksi harga secara akurat menjadi tantangan dalam sektor properti. Penelitian ini bertujuan mengoptimalkan prediksi harga rumah menggunakan algoritma XGBoost Regressor dengan Grid Search Cross Validation (GridSearchCV). Dataset yang digunakan adalah Jakarta House Price Dataset dengan tahapan preprocessing berupa penanganan missing value, transformasi logaritmik pada variabel target, dan One-Hot Encoding pada fitur kategorikal. Model dibangun dalam pipeline terintegrasi dan dioptimasi untuk memperoleh kombinasi hyperparameter terbaik. Hasil penelitian menunjukkan bahwa model yang diusulkan memperoleh nilai R^2 sebesar 0,8889, MAE sebesar 0,2418, MSE sebesar 0,1300, dan RMSE sebesar 0,3606, serta memiliki kinerja yang lebih baik dibandingkan Random Forest. Kontribusi penelitian ini terletak pada integrasi preprocessing dan optimasi hyperparameter XGBoost dalam satu pipeline untuk meningkatkan akurasi prediksi harga rumah pada dataset properti Indonesia. Dengan demikian, model yang dihasilkan dapat digunakan sebagai pendukung pengambilan keputusan di bidang properti.

Kata Kunci : *XGBoost, GridSearchCV, Prediksi Harga Rumah, Machine Learning, Hyperparameter Tuning.*

ABSTRACT

House prices are influenced by various factors, such as location, building size, and property facilities, making accurate price prediction a challenging task in the real estate sector. This study aims to optimize house price prediction using the XGBoost Regressor algorithm with Grid Search Cross Validation (GridSearchCV). The research employs the Jakarta House Price Dataset and applies several preprocessing techniques, including missing value handling, logarithmic transformation of the target variable, and One-Hot Encoding for categorical features. The model is developed within an integrated pipeline and optimized to obtain the best hyperparameter combination. The results show that the proposed model achieves an R^2 score of 0.8889, MAE of 0.2418, MSE of 0.1300, and RMSE of 0.3606, outperforming the Random Forest model. The main contribution of this study lies in the integration of data preprocessing and XGBoost hyperparameter optimization within a single pipeline to improve house price prediction performance on an Indonesian property dataset. Therefore, the developed model can serve as a reliable decision-support tool in the property sector.

Keywords : *XGBoost, GridSearchCV, House Price Prediction, Machine Learning, Hyperparameter Tuning.*

1. PENDAHULUAN

Harga rumah merupakan indikator penting dalam sektor properti yang memengaruhi berbagai keputusan seperti investasi, pembelian, dan penjualan. Prediksi harga rumah yang akurat sangat dibutuhkan agar pengambilan keputusan dapat dilakukan secara rasional dan berbasis data [1]. Seiring perkembangan teknologi, penerapan *machine learning* menjadi solusi yang efektif untuk memprediksi harga rumah dengan memanfaatkan data historis dan berbagai atribut properti [2]. Berbagai algoritma seperti *Linear Regression*, *K-Nearest Neighbor (KNN)*, *Artificial Neural Network (ANN)*, *Random Forest*, hingga *Extreme Gradient Boosting (XGBoost)* telah digunakan untuk memodelkan hubungan kompleks antara fitur properti dan harga rumah [3]. Pemanfaatan kecerdasan buatan dalam sektor properti terus berkembang dan terbukti mampu meningkatkan efisiensi serta akurasi dalam proses penentuan harga rumah secara otomatis [4]. Selain itu, sektor properti memiliki karakteristik data yang dinamis dan kompleks sehingga membutuhkan pendekatan analisis berbasis data yang mampu menangani hubungan non-linear antar variabel secara efektif.

Penelitian terdahulu menunjukkan bahwa model berbasis ensemble memiliki performa yang baik dalam prediksi harga rumah. [5] menunjukkan bahwa *Random Forest* yang dioptimasi menggunakan *Optuna* mampu meningkatkan performa model dibandingkan algoritma lainnya. Namun demikian, masih ditemukan adanya selisih prediksi yang cukup besar pada beberapa data tertentu sehingga model belum sepenuhnya optimal dalam menangkap pola data yang kompleks [3]. Kondisi tersebut menunjukkan bahwa meskipun akurasi model secara umum sudah tinggi, masih terdapat kasus *underprediction* maupun *overprediction* yang cukup signifikan. Permasalahan ini dapat disebabkan oleh distribusi data yang tidak merata, keberadaan fitur yang kurang optimal, serta konfigurasi *hyperparameter* yang belum berada pada kombinasi terbaik [6]. Selain itu, data yang tidak seimbang juga dapat memengaruhi performa model dan menyebabkan hasil prediksi menjadi kurang stabil [7].

Selain faktor algoritma, kualitas *preprocessing* data memiliki peran penting

dalam meningkatkan performa model *machine learning*. Data harga rumah umumnya terdiri atas kombinasi *fitur numerik* dan *kategorikal* yang memerlukan proses pengolahan yang tepat. Kesalahan dalam penanganan *missing value*, transformasi data, maupun *encoding* fitur kategorikal dapat menyebabkan model tidak mampu mempelajari pola data secara optimal [8]. Oleh karena itu, diperlukan pendekatan yang sistematis dalam *preprocessing* data serta optimasi *hyperparameter* untuk meningkatkan performa model prediksi [3]. Beberapa penelitian juga menunjukkan bahwa transformasi logaritmik pada variabel target mampu meningkatkan stabilitas model regresi dengan mengurangi pengaruh distribusi data yang condong (*skewed distribution*) [9].

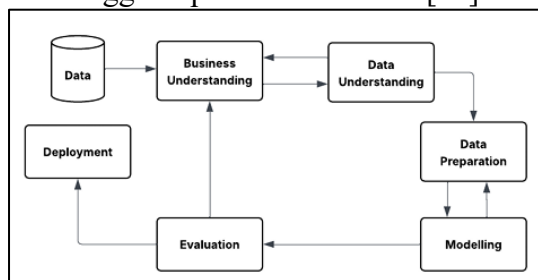
XGBoost merupakan salah satu algoritma *machine learning* berbasis gradient boosting yang memiliki performa tinggi dalam permasalahan *regresi* maupun *klasifikasi*. Algoritma ini mampu menangani hubungan *non-linear* antar variabel serta dilengkapi mekanisme *regularisasi* untuk mengurangi risiko *overfitting* [10]. Beberapa penelitian menunjukkan bahwa *XGBoost* mampu menghasilkan performa yang lebih baik dibandingkan metode regresi tradisional maupun algoritma berbasis *tree* lainnya [11]. Selain itu, penerapan *hyperparameter tuning* menggunakan *GridSearchCV* terbukti mampu meningkatkan performa model melalui pencarian kombinasi parameter terbaik secara sistematis dan tervalidasi [12].

Meskipun penelitian mengenai prediksi harga rumah menggunakan *algoritma ensemble* telah banyak dilakukan, masih terdapat kesenjangan penelitian pada integrasi *preprocessing* data dan optimasi model dalam satu pipeline terpadu yang diterapkan pada dataset properti Indonesia. Sebagian besar penelitian terdahulu berfokus pada peningkatan akurasi model, namun belum secara khusus mengevaluasi pengaruh kombinasi transformasi logaritmik target, *One-Hot Encoding* berbasis *ColumnTransformer*, serta optimasi *hyperparameter* *XGBoost* menggunakan *GridSearchCV* pada *Jakarta House Price Dataset*. Selain itu, implementasi metode tersebut pada konteks pasar properti Indonesia masih relatif terbatas sehingga efektivitasnya perlu dievaluasi lebih lanjut.

Berdasarkan permasalahan tersebut, penelitian ini menggunakan algoritma *XGBoost Regressor* yang dioptimasi menggunakan *GridSearchCV* untuk meningkatkan akurasi prediksi harga rumah. Selain itu, dilakukan transformasi logaritmik pada variabel target serta penerapan *One-Hot Encoding* menggunakan *ColumnTransformer* untuk mengolah fitur kategorikal. Dataset yang digunakan adalah *Jakarta House Price Dataset* sebagai representasi kondisi pasar properti di Indonesia. Kebaruan penelitian ini terletak pada integrasi *preprocessing* data, transformasi logaritmik target, dan optimasi *hyperparameter XGBoost* dalam satu pipeline yang diterapkan pada dataset properti Indonesia. Penelitian ini diharapkan mampu meningkatkan akurasi prediksi serta mengurangi selisih antara nilai aktual dan hasil prediksi sehingga menghasilkan model yang lebih optimal dan dapat diandalkan sebagai pendukung pengambilan keputusan di bidang properti.

2. METODE

Penelitian ini menggunakan pendekatan Cross Industry Standard Process for Data Mining (CRISP-DM) yang terdiri atas enam tahapan, yaitu business understanding, data understanding, data preparation, modeling, evaluation, dan deployment. Pendekatan ini dipilih karena mampu memberikan kerangka kerja yang sistematis dalam pengembangan model machine learning mulai dari identifikasi masalah hingga implementasi model [12].



Gambar 1. Alur CRISP DM

Berdasarkan Gambar 1, tahapan dalam penelitian ini dijelaskan sebagai berikut.

2.1. Business Understanding

Tahap business understanding difokuskan pada permasalahan prediksi harga rumah yang akurat untuk mendukung pengambilan keputusan pada sektor properti. Harga rumah dipengaruhi oleh berbagai faktor seperti lokasi, luas tanah, luas bangunan, jumlah kamar, dan

fasilitas lainnya sehingga diperlukan model yang mampu menangkap hubungan kompleks antar variabel. Tujuan penelitian ini adalah membangun model prediksi harga rumah yang optimal menggunakan algoritma *XGBoost Regressor* dengan optimasi hyperparameter menggunakan *Grid Search Cross Validation (GridSearchCV)*.

2.2. Data Understanding

Dataset yang digunakan adalah *Jakarta House Price Dataset* yang diperoleh dari platform *Kaggle* dan merepresentasikan kondisi pasar properti di Indonesia. Dataset terdiri dari sekitar 10.000 data rumah dengan atribut *district*, *city*, *bed_rooms*, *bath_rooms*, *carport*, *land_area*, *building_area*, dan *price* sebagai variabel target.

Pada tahap ini dilakukan eksplorasi data untuk memahami karakteristik dataset, mengidentifikasi tipe data, distribusi variabel, serta keberadaan missing value yang dapat memengaruhi performa model.

2.3. Data Preparation

Tahap data preparation dilakukan melalui beberapa proses *preprocessing* untuk meningkatkan kualitas data sebelum digunakan pada tahap pemodelan.

1. Penanganan Missing Value

Missing value pada atribut *land_area* dan *building_area* diisi menggunakan nilai median karena lebih robust terhadap *outlier* dibandingkan nilai rata-rata.

2. Penghapusan Fitur Tidak Relevan

Beberapa atribut yang tidak memiliki kontribusi langsung terhadap proses prediksi dihapus untuk mengurangi noise pada data.

3. Transformasi Logaritmik Target

Variabel target *price* ditransformasi menggunakan fungsi logaritmik untuk mengurangi *skewness* dan meningkatkan stabilitas model regresi.

4. One-Hot Encoding

Fitur kategorikal seperti *district* dan *city* dikonversi menjadi format numerik menggunakan *One-Hot Encoding* melalui *ColumnTransformer*.

5. Pembagian Dataset

Dataset dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% menggunakan *train_test_split*. Pembagian ini dilakukan untuk mengevaluasi kemampuan generalisasi model

terhadap data yang belum pernah dipelajari sebelumnya.

2.4. Modeling

Model prediksi dibangun menggunakan algoritma *XGBoost Regressor* yang diintegrasikan dengan tahapan *preprocessing* dalam sebuah *pipeline*. *Integrasi pipeline* dilakukan untuk memastikan seluruh proses transformasi data dilakukan secara konsisten pada data pelatihan maupun data pengujian.

Untuk memperoleh performa terbaik, dilakukan optimasi *hyperparameter* menggunakan *Grid Search Cross Validation* (*GridSearchCV*). Teknik ini mengevaluasi seluruh kombinasi parameter yang telah ditentukan untuk memperoleh konfigurasi model terbaik berdasarkan nilai *Negative Mean Squared Error* (*Neg-MSE*).

Tabel 1. Parameter GridSearch CV

<i>Hyperparameter</i>	<i>Candidate Value</i>
<i>n_estimators</i>	[200, 300, 500]
<i>max_depth</i>	[3, 5, 7]
<i>learning_rate</i>	[0.01, 0.05, 0.1]
<i>subsample</i>	[0.8, 1.0]
<i>colsample_bytree</i>	[0.8, 1.0]
<i>gamma</i>	[0, 0.1, 0.2]
<i>min_child_weight</i>	[1, 3, 5]

Berdasarkan Tabel 1 proses optimasi dilakukan menggunakan *3-Fold Cross Validation*. Pemilihan *3-fold* dilakukan untuk memperoleh keseimbangan antara stabilitas evaluasi model dan efisiensi waktu komputasi mengingat ukuran dataset yang relatif besar.

Eksperimen dalam penelitian ini dilakukan menggunakan bahasa pemrograman *Python* dengan memanfaatkan beberapa pustaka, yaitu *Pandas*, *NumPy*, *Scikit-Learn*, *XGBoost*, *Joblib*, dan *Streamlit*. Seluruh proses *preprocessing* data, pelatihan model, optimasi *hyperparameter* menggunakan *GridSearchCV*, serta evaluasi model dijalankan pada laptop dengan sistem operasi *Windows 11* dan memori (RAM) sebesar 4 GB. Lingkungan eksperimen tersebut digunakan untuk memastikan seluruh tahapan penelitian dapat dijalankan secara konsisten mulai dari pengolahan data hingga implementasi model ke dalam aplikasi berbasis web.

2.5. Evaluation

Evaluasi model dilakukan menggunakan beberapa metrik regresi yaitu *Mean Absolute*

Error (*MAE*), *Mean Squared Error* (*MSE*), *Root Mean Squared Error* (*RMSE*), dan *Coefficient of Determination* (*R² Score*).

MAE digunakan untuk mengukur rata-rata kesalahan absolut prediksi, *MSE* dan *RMSE* digunakan untuk mengukur besar kesalahan prediksi dengan mempertimbangkan penalti terhadap error yang besar, sedangkan *R²* digunakan untuk mengukur kemampuan model dalam menjelaskan variasi data target.

Selain evaluasi *numerik*, dilakukan pula evaluasi visual menggunakan *grafik Actual vs Predicted* dan *Residual Plot* untuk melihat pola kesalahan prediksi yang dihasilkan model.

2.6. Deployment

Model terbaik hasil optimasi disimpan dalam format *.pkl* dan diimplementasikan ke dalam aplikasi berbasis web menggunakan *Streamlit*. Aplikasi memungkinkan pengguna memasukkan karakteristik rumah seperti lokasi, jumlah kamar, luas tanah, dan luas bangunan untuk memperoleh estimasi harga rumah secara otomatis berdasarkan model yang telah dilatih.

3. HASIL DAN PEMBAHASAN

Model prediksi harga rumah dibangun menggunakan algoritma *XGBoost Regressor* yang diintegrasikan dengan *pipeline preprocessing*. Untuk memperoleh performa yang optimal, dilakukan optimasi *hyperparameter* menggunakan *GridSearchCV* sehingga diperoleh kombinasi parameter terbaik. Pendekatan ini memungkinkan model bekerja secara lebih efektif dalam menangkap pola data yang kompleks.

Tabel 2. Hasil Hyperparameter

<i>Hyperparameter</i>	<i>Nilai Terbaik</i>
<i>n_estimators</i>	500
<i>max_depth</i>	3
<i>learning_rate</i>	0,1
<i>subsample</i>	1,0
<i>colsample_bytree</i>	0,8
<i>gamma</i>	0
<i>min_child_weight</i>	3

Berdasarkan proses optimasi menggunakan *GridSearchCV*, diperoleh kombinasi *hyperparameter* terbaik yang ditunjukkan pada Tabel 2. Hasil optimasi menunjukkan bahwa model *XGBoost* bekerja optimal dengan nilai *n_estimators* sebesar 500, *max_depth* sebesar 3,

learning_rate sebesar 0,1, *subsample* sebesar 1,0, *colsample_bytree* sebesar 0,8, *gamma* sebesar 0, dan *min_child_weight* sebesar 3. Kombinasi parameter tersebut dipilih berdasarkan hasil evaluasi menggunakan *3-Fold Cross Validation* dengan mempertimbangkan nilai *Negative Mean Squared Error (Neg-MSE)* terbaik.

Nilai *max_depth* yang relatif rendah menunjukkan bahwa model cenderung menggunakan struktur pohon yang lebih sederhana untuk mengurangi risiko *overfitting*. Sementara itu, penggunaan *n_estimators* yang cukup besar memungkinkan proses *boosting* dilakukan secara bertahap sehingga model dapat mempelajari pola data dengan lebih baik. Selain itu, nilai *learning_rate* sebesar 0,1 memberikan keseimbangan antara kecepatan pembelajaran dan stabilitas model selama proses pelatihan. Hasil optimasi ini menunjukkan bahwa pemilihan *hyperparameter* yang tepat berperan penting dalam meningkatkan kemampuan model *XGBoost* dalam melakukan prediksi harga rumah.

Model yang telah dioptimasi kemudian dievaluasi menggunakan beberapa metrik regresi untuk mengukur kinerjanya.

```

*** Performance of the Tuned Model:
Mean Absolute Error: 0.24189826922988403
Mean Squared Error: 0.13007322338135857
Root Mean Squared Error: 0.3606566558118103
R2 Score: 0.8889843619954766

```

Gambar 2. Hasil modeling

Berdasarkan Gambar 2, diperoleh nilai *Mean Absolute Error (MAE)* sebesar 0.2418, *Mean Squared Error (MSE)* sebesar 0.1300, *Root Mean Squared Error (RMSE)* sebesar 0.3606, serta nilai *R² Score* sebesar 0.8889. Nilai error yang relatif kecil menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang rendah, sedangkan nilai *R²* yang mendekati 1 menunjukkan bahwa model mampu menjelaskan sebagian besar variasi data harga rumah dengan baik.

Nilai *R²* sebesar 0,8889 menunjukkan bahwa model mampu menjelaskan sekitar 88,89% variasi harga rumah berdasarkan fitur yang digunakan. Sementara itu, nilai *MAE*, *MSE*, dan *RMSE* yang relatif rendah menunjukkan bahwa kesalahan prediksi yang dihasilkan model berada pada tingkat yang dapat diterima. Hasil ini mengindikasikan bahwa kombinasi

preprocessing data dan optimasi *hyperparameter* menggunakan *GridSearchCV* mampu meningkatkan kemampuan model dalam menangkap hubungan non-linear antar variabel pada dataset properti. Temuan ini sejalan dengan penelitian [11] yang menyatakan bahwa optimasi *hyperparameter* pada *XGBoost* dapat meningkatkan performa model secara signifikan dibandingkan penggunaan parameter bawaan.

Hal ini menunjukkan bahwa model memiliki tingkat akurasi yang tinggi dalam memprediksi harga rumah berdasarkan fitur yang digunakan. Selain itu, kombinasi nilai error yang rendah dan nilai *R²* yang tinggi mengindikasikan bahwa model tidak hanya mampu memberikan prediksi yang mendekati nilai aktual, tetapi juga memiliki kemampuan yang baik dalam menangkap pola hubungan antar variabel dalam dataset.

Selanjutnya dilakukan perbandingan kinerja antara model *XGBoost* yang dioptimasi menggunakan *GridSearchCV* dengan model *Random Forest* sebagai pembanding. Perbandingan ini bertujuan untuk mengetahui model mana yang memiliki performa lebih baik dalam memprediksi harga rumah.

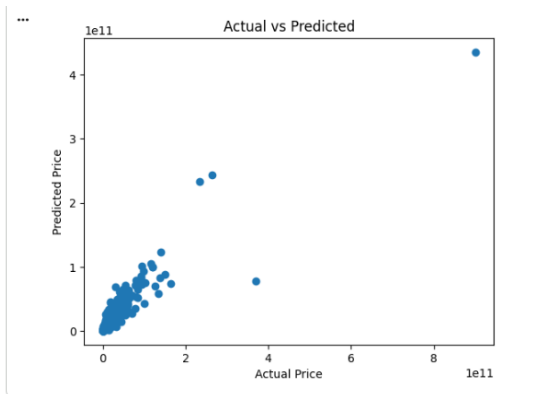
Tabel 3. Tabel Perbandingan

Metrik	XGBoost GridSearchCV	+ Random Forest + Optuna
MAE	0.2418	0.2428
MSE	0.1300	0.1381
RMSE	0.3606	0.3716
R ² Score	0.8889	0.8822

Berdasarkan Tabel 3, hasil perbandingan model *XGBoost* yang dioptimasi menggunakan *GridSearchCV* menghasilkan performa yang lebih baik dibandingkan *Random Forest* pada seluruh metrik evaluasi. Nilai *MAE* menurun dari 0,2428 menjadi 0,2418, nilai *MSE* menurun dari 0,1381 menjadi 0,1300, dan nilai *RMSE* menurun dari 0,3716 menjadi 0,3606. Selain itu, nilai *R²* meningkat dari 0,8822 menjadi 0,8889. Meskipun peningkatan tersebut relatif tidak terlalu besar, hasil ini menunjukkan bahwa *XGBoost* memiliki kemampuan yang lebih baik dalam menangkap pola hubungan yang kompleks melalui mekanisme *gradient boosting* yang bekerja secara bertahap untuk memperbaiki kesalahan model sebelumnya.

Peningkatan performa yang tidak terlalu jauh juga menunjukkan bahwa kedua algoritma sama-sama memiliki kemampuan yang baik dalam menangani data harga rumah. Oleh karena itu, keunggulan *XGBoost* dalam penelitian ini perlu dipahami sebagai peningkatan performa yang bersifat *incremental* namun konsisten pada seluruh metrik evaluasi, bukan sebagai perbedaan yang sangat signifikan. Hasil ini sejalan dengan penelitian [5] yang menunjukkan bahwa model berbasis *ensemble* cenderung menghasilkan performa yang kompetitif dalam permasalahan prediksi harga rumah.

Untuk memperkuat hasil evaluasi yang telah diperoleh, dilakukan analisis visual menggunakan grafik perbandingan nilai aktual dan hasil prediksi.

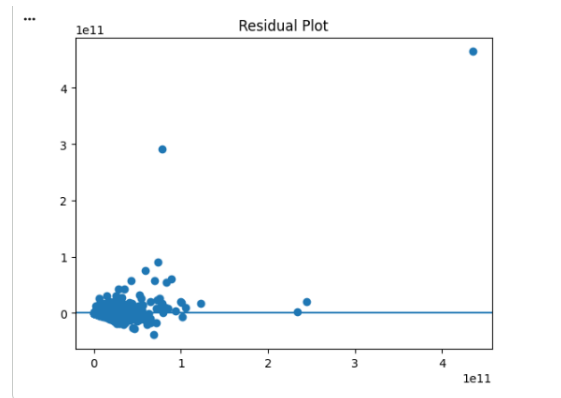


Gambar 3. Grafik Actual vs Predicted

Berdasarkan Gambar 3, terlihat bahwa sebagian besar titik data berada di sekitar garis diagonal, yang menunjukkan bahwa nilai prediksi yang dihasilkan model cukup mendekati nilai aktual. Hal ini menandakan bahwa model memiliki kemampuan yang baik dalam mengikuti pola data.

Meskipun sebagian besar titik data berada di sekitar garis diagonal, terdapat beberapa titik yang menyimpang, terutama pada nilai harga yang lebih tinggi. Hal ini menunjukkan bahwa model masih mengalami keterbatasan dalam memprediksi data dengan nilai ekstrem, yang umumnya dipengaruhi oleh variasi fitur yang lebih kompleks. Selain itu, penyebaran titik yang relatif merata di sepanjang garis diagonal menunjukkan bahwa model mampu menangkap pola hubungan antara variabel dengan cukup baik, meskipun masih terdapat ruang untuk peningkatan akurasi pada beberapa kondisi tertentu.

Selanjutnya, dilakukan analisis residual untuk melihat distribusi kesalahan prediksi.

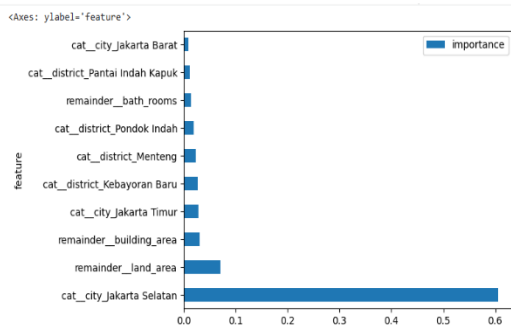


Gambar 4. Residual Plot

Berdasarkan Gambar 4, residual tersebar di sekitar garis nol tanpa membentuk pola tertentu. Hal ini menunjukkan bahwa model tidak mengalami bias yang signifikan dan mampu melakukan prediksi dengan cukup baik, meskipun terdapat beberapa nilai residual yang cukup besar yang mengindikasikan adanya data ekstrem.

Selain itu, terdapat beberapa nilai residual yang cukup besar pada titik tertentu, yang menunjukkan adanya kemungkinan data ekstrem atau outlier dalam dataset. Kondisi ini dapat memengaruhi tingkat akurasi model pada beberapa kasus tertentu. Meskipun demikian, tidak adanya pola tertentu pada distribusi residual menunjukkan bahwa model telah mampu menangkap hubungan antara variabel dengan baik dan tidak mengalami kesalahan sistematis dalam prediksi. Hal ini mengindikasikan bahwa model memiliki kemampuan generalisasi yang cukup baik terhadap data uji, meskipun masih terdapat peluang untuk peningkatan performa melalui penanganan *outlier* atau penambahan fitur yang lebih representatif.

Untuk memperoleh pemahaman yang lebih baik mengenai faktor-faktor yang memengaruhi harga rumah, dilakukan analisis feature importance terhadap model *XGBoost* yang telah dioptimasi.



Gambar 5. Feature Importance

Gambar 6. *Feature Importance XGBoost* menunjukkan tingkat kontribusi masing-masing fitur terhadap proses prediksi harga rumah yang dilakukan oleh model *XGBoost*. Berdasarkan grafik tersebut, fitur *cat_city_Jakarta Selatan* memiliki nilai *importance* tertinggi dibandingkan fitur lainnya, dengan nilai sekitar 0,60. Hasil ini menunjukkan bahwa lokasi rumah yang berada di wilayah Jakarta Selatan merupakan faktor yang paling dominan dalam memengaruhi prediksi harga rumah pada dataset yang digunakan. Tingginya kontribusi fitur ini mengindikasikan bahwa karakteristik pasar properti di Jakarta Selatan memiliki perbedaan yang cukup signifikan dibandingkan wilayah lain, sehingga sering digunakan model sebagai pembeda utama dalam menentukan nilai properti.

Selain faktor lokasi, fitur *land_area* dan *building_area* juga menunjukkan kontribusi yang cukup besar terhadap hasil prediksi. Hal ini menunjukkan bahwa luas tanah dan luas bangunan merupakan atribut fisik yang berperan penting dalam pembentukan harga rumah. Semakin besar luas tanah dan bangunan yang dimiliki suatu properti, semakin tinggi pula kecenderungan nilai jual rumah tersebut. Temuan ini sejalan dengan konsep penilaian properti yang menyatakan bahwa ukuran properti merupakan salah satu faktor utama yang memengaruhi harga pasar.

Beberapa fitur lain seperti *city Jakarta Timur*, *district Kebayoran Baru*, *district Menteng*, *district Pondok Indah*, *bath_rooms*, dan *district Pantai Indah Kapuk* juga memberikan kontribusi terhadap model, meskipun nilainya relatif lebih kecil. Kontribusi yang lebih rendah tidak menunjukkan bahwa fitur tersebut tidak penting, melainkan pengaruhnya terhadap variasi harga rumah pada dataset lebih kecil dibandingkan fitur-fitur utama yang telah disebutkan sebelumnya.

Secara keseluruhan, hasil analisis *feature importance* menunjukkan bahwa faktor lokasi dan ukuran properti merupakan variabel yang paling berpengaruh dalam prediksi harga rumah. Temuan ini mendukung hasil evaluasi model sebelumnya yang menunjukkan bahwa *XGBoost* mampu menangkap hubungan kompleks antara karakteristik properti dan harga rumah dengan baik. Selain meningkatkan interpretabilitas model, analisis ini juga memberikan informasi yang bermanfaat bagi pembeli, penjual, maupun investor dalam memahami faktor-faktor utama yang memengaruhi harga rumah pada pasar properti Jakarta.

Sebagai implementasi dari model yang telah dikembangkan, dilakukan deployment dalam bentuk aplikasi berbasis web menggunakan Streamlit. Aplikasi ini dirancang untuk mempermudah pengguna dalam melakukan prediksi harga rumah secara interaktif dengan memasukkan data properti.

Gambar 6. Tampilan Aplikasi

Berdasarkan Gambar 6, aplikasi menyediakan antarmuka yang sederhana dan mudah digunakan. Pengguna dapat memasukkan data seperti wilayah, kota, jumlah kamar tidur, kamar mandi, kapasitas mobil, luas tanah, dan luas bangunan. Setelah data dimasukkan, sistem akan menampilkan hasil prediksi harga rumah secara otomatis. Implementasi ini menunjukkan bahwa model yang dikembangkan tidak hanya memiliki performa yang baik secara teoritis, tetapi juga dapat digunakan secara praktis dalam membantu pengambilan keputusan di bidang properti.

Hasil ini menunjukkan bahwa kombinasi antara preprocessing yang tepat, pemilihan algoritma yang sesuai, serta optimasi hyperparameter memiliki peran penting dalam meningkatkan performa model secara signifikan.

Secara keseluruhan, hasil yang diperoleh menunjukkan bahwa model yang dikembangkan memiliki performa yang baik serta mampu memberikan prediksi yang akurat dan dapat diimplementasikan dalam sistem berbasis web untuk mendukung pengambilan keputusan di bidang properti.

4. PENUTUP

4.1. Kesimpulan

Penelitian ini berhasil membangun model prediksi harga rumah menggunakan algoritma *XGBoost Regressor* yang dioptimasi dengan *Grid Search Cross Validation (GridSearchCV)*. Hasil penelitian menunjukkan bahwa pendekatan yang digunakan mampu menghasilkan performa prediksi yang baik serta memberikan hasil yang lebih unggul dibandingkan model *Random Forest* pada seluruh metrik evaluasi yang digunakan. Integrasi preprocessing data berupa penanganan *missing value*, transformasi logaritmik pada variabel target, dan *One-Hot Encoding* dengan optimasi hyperparameter dalam satu pipeline terbukti mampu meningkatkan kemampuan model dalam mempelajari pola data harga rumah.

Kontribusi utama penelitian ini terletak pada penerapan *pipeline* terintegrasi yang menggabungkan *preprocessing* data dan optimasi hyperparameter *XGBoost* pada *Jakarta House Price Dataset* sebagai representasi pasar properti Indonesia. Selain menghasilkan model yang akurat, penelitian ini juga memberikan pemahaman mengenai faktor-faktor yang paling berpengaruh terhadap harga rumah melalui analisis *feature importance*, di mana lokasi dan ukuran properti menjadi variabel dominan dalam proses prediksi.

4.2. Saran

Penelitian ini masih memiliki keterbatasan karena menggunakan dataset yang hanya mencakup wilayah Jakarta sehingga kemampuan generalisasi model terhadap wilayah lain di Indonesia belum dapat dievaluasi secara menyeluruh. Oleh karena itu, penelitian selanjutnya disarankan untuk

menggunakan dataset dengan cakupan wilayah yang lebih luas agar model memiliki kemampuan generalisasi yang lebih baik. Selain itu, penambahan fitur yang lebih kompleks seperti akses fasilitas publik, kondisi lingkungan, dan faktor geografis dapat dipertimbangkan untuk meningkatkan akurasi prediksi. Penelitian berikutnya juga dapat membandingkan performa *XGBoost* dengan algoritma boosting lainnya seperti *LightGBM* dan *CatBoost* maupun menerapkan teknik optimasi hyperparameter yang berbeda untuk memperoleh model prediksi yang lebih optimal.

5. DAFTAR PUSTAKA

- [1] F. H. Balqis *et al.*, “Prediksi Harga Rumah menggunakan Algoritma Random Forest Regression House Price Prediction using the Random Forest Regression Algorithm,” vol. 15, pp. 433–445, 2026.
- [2] H. Li, “House Price Prediction and Analysis Based on Random Forest and XGBoost Models,” vol. 21, pp. 934–938, 2023.
- [3] N. Hajiyianni and V. Stylianou, “Machine Learning for Real Estate Price Prediction,” vol. 1, pp. 1–11, 2025.
- [4] D. J. Yantono, N. R. Puspaningrum, and S. A. Zein, “Comparative Analysis of Machine Learning Algorithms for Predicting Commercial Property Rental Prices Analisis Komparasi Algoritma Machine Learning untuk Prediksi Harga Sewa Properti Komersial,” vol. 6, no. January, pp. 149–155, 2026.
- [5] R. M. Fadzryan, E. A. Laksana, T. Informatika, and U. Widyatama, “Optimasi Prediksi Harga Rumah dengan Random Forest dan Optuna Hyperparameter Tuning Implementation of Random Forest and Optuna Hyperparameter Tuning Combination to Improve House Price Prediction Accuracy,” vol. 5, no. 6, pp. 1663–1671, 2025.
- [6] H. Sharma, H. Harsora, and B. Ogunleye, “An Optimal House Price Prediction Algorithm : XGBoost,” pp. 30–45, 2024.
- [7] A. Nugroho and D. Harini, “Teknik Random Forest untuk Meningkatkan

- Akurasi Data Tidak Seimbang,” vol. 2, no. 2, pp. 128–140, 2024.
- [8] A. Ristyan, A. Nugroho, and T. K. Amarya, “Optimasi Preprocessing Model Random Forest Untuk Prediksi Stroke,” vol. 12, no. 1, 2025.
- [9] E. Tr, A. Ristyan, and A. Nugroho, “Pengaruh Normalisasi Fitur Pada Perbandingan Algoritma Random Forest Regressor dan KNN Regressor dalam Menentukan Harga Smartphone Effect of Feature Normalization on Comparison of Random Forest Regressor and KNN Regressor Algorithms in Determining Smartphone Prices,” vol. 5, pp. 74–83, 2025.
- [10] R. M. Syafei and D. A. Efrilianda, “Machine Learning Model Using Extreme Gradient Boosting (XGBoost) Feature Importance and Light Gradient Boosting Machine (LightGBM) to Improve Accurate Prediction of Bankruptcy,” *Recursive J. Informatics*, vol. 1, no. 2, pp. 64–72, 2023, doi: 10.15294/rji.v1i2.71229.
- [11] H. Wijaya, D. P. Hostiadi, and E. Triandini, “Optimization XGBoost Algorithm Using Parameter Tuning in Retail Sales Prediction,” *J. Nas. Pendidik. Tek. Inform.*, vol. 13, no. 3, pp. 769–786, 2024, doi: 10.23887/janapati.v13i3.82214.
- [12] B. M. Islami and A. Nugroho, “Analisis Algoritma KNN dan Penerapan SMOTE dalam Deteksi Dini Kanker Paru-Paru,” vol. 15, no. 2, pp. 38–50, 2025.