



House Price Prediction Using XGBoost Based on Optuna Hyperparameter Optimization

Prediksi Harga Rumah Menggunakan XGBoost Berbasis Optuna Hyperparameter Optimization

Nandana Ayudya Natasaskara^{1*}, Yoga Pristyanto², Ika Nur Fajri³

^{1,2,3}Program Studi Sistem Informasi, Fakultas Ilmu Komputer,
Universitas Amikom Yogyakarta, Indonesia

E-Mail: ¹nandana@students.amikom.ac.id, ²yoga.pristyanto@amikom.ac.id, ³fajri@amikom.ac.id

Received Apr 11th 2026; Revised May 03rd 2026; Accepted Jun 03rd 2026; Available Online Jun 20th 2026

Corresponding Author: Nandana Ayudya Natasaskara

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

Fluctuations and the complexity of property market attributes make house price prediction difficult and less accurate when relying solely on default algorithm parameters. The optimal performance of Machine learning algorithms, such as Extreme Gradient Boosting (XGBoost), highly depends on proper hyperparameter tuning; however, many previous studies have either neglected this Optimization or relied on inefficient conventional methods. To address this issue, this study proposes an XGBoost model dynamically optimized with the Optuna Hyperparameter Optimization framework. Optuna, which uses Bayesian Optimization, intelligently and efficiently explores the parameter space to achieve faster convergence. Experimental results demonstrate that integrating Optuna significantly improves prediction accuracy. The Optuna-based XGBoost model achieves superior performance, with the R^2 score increasing from 0.8204 to 0.8263, while successfully reducing error rates, with RMSE decreasing to Rp 301,132,090.80, MAE decreasing to Rp 196,869,100.79, and MAPE decreasing to 16.53%. This approach proves more robust, stable, and precise than the baseline (non-optimized) model at capturing nonlinear price patterns. Despite the improvement in accuracy, this method requires substantially more training computation time, increasing to 3,550.34 seconds. The final dataset used in this study consists of 16,674 records, obtained after extensive preprocessing and outlier removal from an initial pool of 40,200 records.

Keyword: *Extreme Gradient Boosting, House Price, Hyperparameter Optimization, Optuna, Prediction*

Abstrak

Fluktuasi dan kompleksitas atribut pasar properti membuat prediksi harga rumah sulit dan kurang akurat jika hanya mengandalkan parameter algoritma bawaan. Kinerja optimal algoritma *Machine learning* seperti Extreme Gradient Boosting (XGBoost) sangat bergantung pada pengaturan *hyperparameter* yang tepat, namun banyak penelitian sebelumnya mengabaikan optimasi atau menggunakan metode konvensional yang tidak efisien. Untuk mengatasi masalah tersebut, penelitian ini mengusulkan model XGBoost yang dioptimasi secara dinamis menggunakan kerangka kerja *Optuna Hyperparameter Optimization*. Optuna, yang bekerja berdasarkan optimasi Bayesian, secara cerdas dan efisien mengeksplorasi ruang parameter guna menemukan konvergensi yang lebih cepat. Hasil eksperimen membuktikan bahwa integrasi Optuna berhasil meningkatkan keakurasian prediksi secara signifikan. Model XGBoost berbasis Optuna menghasilkan performa yang lebih unggul dengan peningkatan skor R^2 dari 0.8204 menjadi 0.8263, serta berhasil menekan tingkat kesalahan di mana RMSE turun menjadi Rp 301.132.090,80, MAE menjadi Rp 196.869.100,79, dan MAPE menyusut menjadi 16,53%. Pendekatan ini terbukti lebih tangguh, stabil, dan presisi dibandingkan model tanpa optimasi (*baseline*) dalam memetakan pola harga yang *non-linear*. Meskipun akurasi meningkat, hal ini menuntut waktu komputasi pelatihan yang jauh lebih tinggi, yakni melonjak drastis menjadi 3.550,34 detik. *Dataset* akhir yang digunakan dalam penelitian ini berjumlah 16.674 catatan, yang diperoleh setelah proses *preprocessing* dan eliminasi *outlier* secara ekstensif dari 40.200 catatan awal.

Kata Kunci: *Extreme Gradient Boosting, Harga Rumah, Hyperparameter Optimization, Optuna, Prediksi*

1. PENDAHULUAN

Perumahan merupakan salah satu kebutuhan dasar manusia, dan memprediksi harga rumah secara akurat memiliki peranan yang sangat esensial bagi berbagai sektor, seperti agen *real estate*, lembaga pemberi pinjaman hipotek (KPR), serta pembeli maupun investor [1]. Fluktuasi harga rumah dapat berdampak signifikan terhadap pengeluaran konsumen, tingkat pinjaman, dan stabilitas perekonomian secara keseluruhan [2]. Namun, pasar *real estate* sangatlah dinamis dan harga sebuah properti dipengaruhi oleh berbagai variabel, termasuk atribut fisik bangunan dan faktor lingkungan sekitar [1]. *Dataset* pada penelitian ini difokuskan pada wilayah Depok karena kota ini merupakan salah satu kawasan penyangga utama ibu kota dengan dinamika pasar properti yang sangat fluktuatif, didorong oleh ekspansi infrastruktur, pertumbuhan penduduk, dan tingginya permintaan hunian dari pekerja komuter Jakarta [3]. Interaksi dari berbagai variabel ini bersifat kompleks dan *non-linear* [4]. Mengingat volatilitas pasar saat ini, diperlukan model prediksi yang handal untuk menjembatani kesenjangan antara penawaran dan permintaan serta mengurangi risiko kerugian finansial [5].

Dalam memprediksi harga rumah, banyak studi telah beralih dari model regresi linier tradisional ke algoritma *Machine learning* (ML) karena kemampuannya yang unggul dalam memodelkan hubungan data yang kompleks. Algoritma Extreme Gradient Boosting (XGBoost) telah mendapatkan perhatian luas dan terbukti sangat efektif karena menawarkan komputasi yang dapat diskalakan, interpretasi yang baik, dan akurasi prediktif yang tinggi [6], [7]. XGBoost juga dirancang dengan teknik regularisasi yang membuatnya sangat tahan terhadap masalah *overfitting* [6], [7], [8]. Meskipun demikian, kinerja optimal dari XGBoost dan model *machine learning* lainnya sangat bergantung pada pengaturan *hyperparameter* [9], [10]. Banyak penelitian sebelumnya tentang prediksi harga rumah yang hanya menggunakan parameter bawaan (*default*) tanpa upaya pencapaian nilai optimal [1], atau menggunakan metode penalaan konvensional seperti *Grid Search* yang memakan biaya komputasi besar dan sangat lambat untuk memproses parameter dalam dimensi tinggi [11]. Situasi ini menciptakan *research gap* yang jelas dan mendorong perlunya eksplorasi terhadap metode optimasi *hyperparameter* yang lebih efisien dan dinamis. Berdasarkan permasalahan tersebut, diperlukan pendekatan yang mampu mengatasi keterbatasan metode penalaan konvensional secara lebih efektif.

Penelitian ini dilakukan untuk mengatasi masalah inefisiensi pada penalaan parameter model ML dengan mengusulkan kerangka kerja optimasi yang lebih canggih. Penelitian ini menggunakan Optuna, sebuah kerangka optimasi *hyperparameter* yang bekerja secara otomatis dan dinamis. Berbeda dengan pencarian konvensional, Optuna memanfaatkan *Tree-structured Parzen Estimator* (TPE) berdasarkan optimasi Bayesian, yang secara cerdas dan efisien mengeksplorasi ruang *hyperparameter* dengan mengevaluasi hasil percobaan sebelumnya untuk menemukan konvergensi secara lebih cepat [12]. Selain itu, mekanisme pruning pada Optuna mampu menghentikan percobaan yang kurang menjanjikan secara dini, sehingga menghemat waktu dan sumber daya komputasi [11], [13]. Dengan mengintegrasikan algoritma XGBoost dan kerangka kerja Optuna, penelitian ini diharapkan dapat merancang model prediksi harga rumah yang jauh lebih akurat, stabil, dan terhindar dari bias.

Oleh karena itu, penelitian ini bertujuan untuk menganalisis bagaimana pengaruh integrasi optimasi *hyperparameter* berbasis Optuna terhadap tingkat akurasi dan pengurangan tingkat kesalahan, seperti RMSE dan MAE, pada model XGBoost dalam memprediksi harga rumah. Selain itu, penelitian ini juga berupaya membuktikan apakah model XGBoost yang dioptimasi dengan Optuna mampu menghasilkan performa prediksi yang lebih unggul dan tangguh jika dibandingkan dengan model XGBoost yang hanya menggunakan parameter bawaan atau algoritma *machine learning* konvensional lainnya. Kontribusi utama penelitian ini adalah menyajikan bukti empiris penggunaan Optuna sebagai *framework* optimasi *hyperparameter* pada model XGBoost untuk prediksi harga properti, dilengkapi dengan dokumentasi *search space* yang lengkap dan *reproducible* sehingga dapat dijadikan acuan bagi penelitian selanjutnya di bidang prediksi harga properti Indonesia.

2. METODOLOGI PENELITIAN

Tahapan penelitian dapat dilihat pada Gambar 1.

2.1. Dataset dan Pembersihan Data

Penelitian ini menggunakan *dataset* harga rumah sekunder yang bersumber dari *platform* Kaggle, secara spesifik difokuskan pada Data Harga Rumah di wilayah Depok. *Dataset* awal terdiri dari 40.200 catatan data dengan delapan atribut fitur, yang mencakup tujuh variabel bertipe *Object* dan satu variabel bertipe *Int64*. Karakteristik data sebelum dilakukan proses pembersihan dapat dilihat pada Tabel 1.

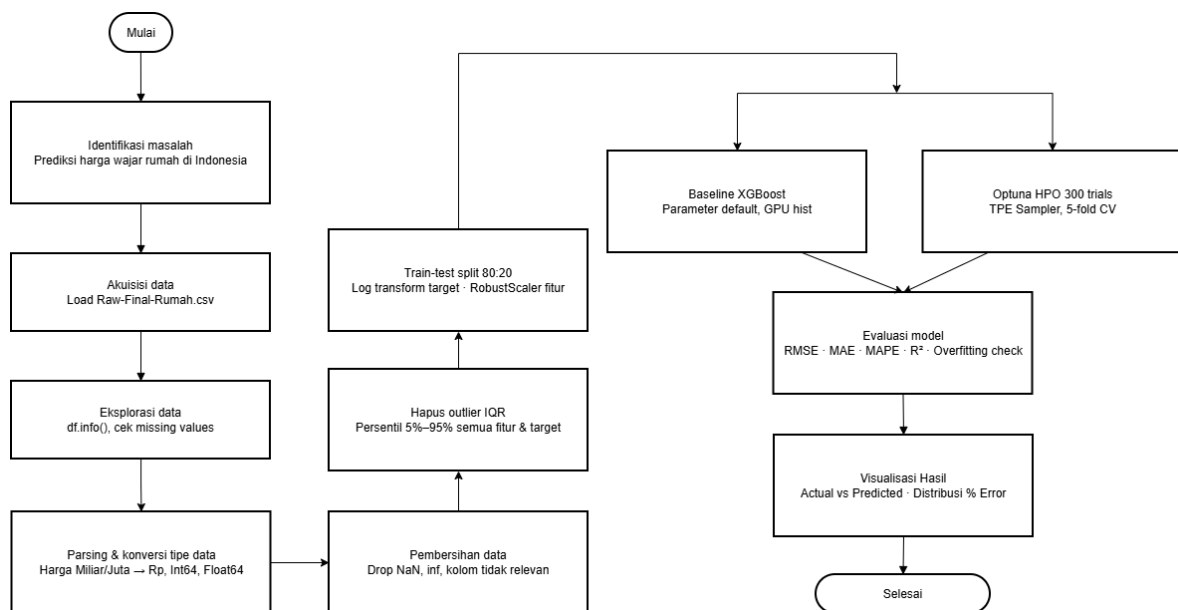
Tahap pra-pemrosesan (*preprocessing*) dilakukan secara ekstensif untuk mempersiapkan data mentah menjadi format yang dapat diproses oleh model. Langkah-langkah pembersihan meliputi:

1. Konversi Tipe Data: Nilai harga rumah yang direpresentasikan dalam format teks ("miliar" atau "juta") diekstraksi dan dikonversi ke dalam tipe data numerik (*Float64*). Kolom numerik diskrit seperti Kamar Tidur dan Kamar Mandi diubah ke format *Int64*, sedangkan nilai yang kosong pada kolom Garasi diisi

dengan angka 0 sebelum dikonversi. Untuk Luas Tanah dan Luas Bangunan, metode ekspresi reguler (*regex*) diaplikasikan guna mengekstrak nilai numerik yang relevan menjadi *Float64*.

2. Penanganan Nilai Hilang (*Missing Values*): Nilai tak terhitung diformat menjadi *Not a Number* (NaN) dan seluruh baris yang memiliki nilai kosong pada fitur-fitur krusial dihapus dari *dataset* menggunakan fungsi *dropna*. Atribut wilayah (Kecamatan) juga dihapus karena tidak digunakan dalam analisis lebih lanjut. Keputusan untuk menghapus atribut wilayah ini diambil untuk membatasi ruang lingkup penelitian hanya pada pengaruh fitur-fitur fisik (lahan dan bangunan) terhadap harga, serta menghindari kompleksitas feature engineering berbasis spasial pada data sekunder yang digunakan. Meskipun demikian, penelitian mendatang disarankan untuk mempertahankan fitur ini dengan teknik seperti *Target Encoding* atau *Label Encoding*, guna mengeksplorasi kontribusi faktor lokasi terhadap akurasi prediksi secara lebih komprehensif.

Penghapusan Data Pencilan (*Outlier*): Untuk mengatasi anomali, pencilan pada fitur *input* (Kamar Tidur, Kamar Mandi, Garasi, Luas Tanah, Luas Bangunan) maupun variabel target (Harga) dieliminasi menggunakan metode statistik Interquartile Range (IQR). Data dibatasi secara ketat dalam rentang persentil ke-5 hingga ke-95 untuk menghindari bias pada model [14].



Gambar 1. Tahapan Penelitian

Tabel 1. Karakteristik Data sebelum dilakukan Proses Pembersihan

Nama Atribut	Tipe Data	Jumlah <i>Missing Value</i>	Contoh Nilai
Harga	<i>Object</i>	0	2,5 Miliar
Kamar Tidur	<i>Object</i>	9	4
Kamar Mandi	<i>Object</i>	4.403	3
Garasi	<i>Object</i>	5.189	10
Luas Tanah	<i>Object</i>	7.699	438 m ²
Luas Bangunan	<i>Object</i>	15.602	110 m ²
Kecamatan	<i>Object</i>	0	Tapos
Page	<i>Int64</i>	0	129

2.2. Transformasi dan Pembagian Data

Setelah proses pembersihan, total *dataset* yang layak digunakan berjumlah 16.674 catatan. Data ini kemudian dipartisi menjadi dua kelompok dengan rasio 80:20, yaitu 13.339 baris (80%) sebagai data pelatihan (*training*) dan 3.335 baris (20%) sebagai data pengujian (*testing*).

Untuk menangani masalah kemiringan distribusi data (*skewness*) yang terjadi pada harga properti, variabel target pelatihan (y_{train}) ditransformasikan ke dalam skala logaritmik menggunakan fungsi np.log1p . Normalisasi fitur dilakukan menggunakan algoritma *RobustScaler*. Metode ini dipilih karena ketahanannya terhadap pencilan residual; *scaler di-fit* hanya pada data pelatihan (X_{train}) dan kemudian diaplikasikan untuk menskalakan fitur pada data pelatihan (X_{train_sc}) maupun data pengujian (X_{test_sc}) [15].

2.3. Pemodelan XGBoost dan Optimasi Hyperparameter Optuna

Pemodelan utama dilakukan menggunakan eXtreme Gradient Boosting (XGBoost), sebuah algoritma ensemble berbasis pohon (tree-based) yang sangat efisien, tahan terhadap *overfitting*, dan mampu menangani hubungan data yang kompleks [4], [16], [17]. Dalam eksperimen ini, dua pendekatan pemodelan XGBoost dilakukan untuk melihat peningkatan performa:

1. *Baseline* XGBoost: Sebuah model regresi dasar (XGBRegressor) dilatih dengan menggunakan konfigurasi parameter standar bawaan, dengan akselerasi komputasi yang memanfaatkan Graphics Processing Unit (GPU) melalui pengaturan `device='cuda'` [18].
2. Optimasi Optuna: Untuk memaksimalkan keakuratan prediksi, kerangka kerja Optuna diaplikasikan [19], [20]. Optuna adalah *framework* yang secara dinamis dan efisien mencari ruang *hyperparameter* optimal menggunakan algoritma seperti *Tree-structured Parzen Estimator* (TPE) [21]. Proses pencarian dilakukan selama 300 percobaan (trials) yang divalidasi menggunakan teknik 5-fold cross-validation. Sebagai keterbatasan penelitian, mekanisme pruning pada Optuna seperti MedianPruner dan HyperbandPruner tidak diaktifkan dalam studi ini. Pada penelitian mendatang, pengaktifan pruning diharapkan dapat memangkas waktu komputasi yang mencapai 3.550,34 detik secara signifikan, tanpa mengorbankan akurasi yang telah dicapai.

Agar eksperimen ini dapat direproduksi oleh peneliti lain, ruang pencarian (*search space*) dari fungsi *Objective* Optuna ditetapkan seperti pada Tabel 2.

Tabel 2. Ruang Pencarian dan *Objective* Optuna

Parameter	XGBoost <i>Baseline</i>	Optuna HPO Range
n_estimators	-	300 – 1000
max_depth	-	3 – 10
learning_rate	-	0.005 – 0.1 (log scale)
subsample	-	0.6 – 1.0
colsample_bytree	-	0.5 – 1.0
min_child_weight	-	1 – 20
gamma	-	0 – 5
reg_alpha	-	1e-4 – 10 (log scale)
reg_lambda	-	1e-4 – 10 (log scale)
random_state	42	42
tree_method	hist	hist
device	cuda	cuda

Setelah konfigurasi terbaik ditemukan melalui Optuna, model XGBoost final dilatih kembali menggunakan parameter-parameter tersebut. Pada tahap ini, fungsi early stopping (diatur pada 20 ronde) juga ditambahkan guna menghentikan pelatihan secara otomatis apabila performa model tidak menunjukkan perbaikan, sehingga mencegah terjadinya *overfitting* [22].

2.4. Evaluasi Kinerja Model

Metode karakterisasi kinerja model dilakukan dengan cara membandingkan nilai prediksi yang dihasilkan terhadap nilai transaksi aktual (ground truth). Sebelum metrik evaluasi dihitung, prediksi dari model (yang masih berada dalam skala logaritmik) ditransformasikan kembali ke bentuk aslinya menggunakan fungsi eksponensial.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

Kinerja model dinilai berdasarkan lima parameter evaluasi regresi standar: *Root Mean Square Error* (RMSE) untuk melihat deviasi kuadratik, *Mean Absolute Error* (MAE) untuk menghitung rata-rata kesalahan absolut, *Mean Absolute Percentage Error* (MAPE) untuk menilai persentase simpangan error, Koefisien Determinasi (R^2) untuk mengetahui seberapa baik model menjelaskan varians data, serta waktu komputasi pelatihan model dalam satuan detik [23]. Metrik evaluasi ini digunakan untuk membandingkan apakah metode XGBoost yang dioptimalkan dengan Optuna mampu mengungguli XGBoost *baseline*.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Pemrosesan Data

Pemrosesan data diawali dengan evaluasi kualitas *dataset* mentah yang berjumlah 40.200 catatan. Terdapat temuan nilai yang hilang (*Missing Values*) secara signifikan pada beberapa fitur struktural utama, terutama pada fitur Luas Bangunan (hilang 15.602 data), Luas Tanah (7.699 data), dan Garasi (5.189 data). Setelah itu, dilakukan tahap pembersihan secara ekstensif termasuk konversi ke tipe data numerik, pengisian nilai kosong, dan penghapusan anomali (*outlier*) menggunakan batas persentil 5% hingga 95% *dataset* menyusut menjadi 16.674 catatan yang solid dan representatif. Kumpulan data ini kemudian didistribusikan secara proporsional menjadi data pelatihan dan data pengujian untuk proses pemodelan. Karakteristik data setelah dilakukan proses pembersihan dapat dilihat pada Tabel 3.

Tabel 3. Karakteristik Data setelah dilakukan Proses Pembersihan

Nama Atribut	Tipe Data	Jumlah Missing Value	Contoh Nilai
Harga	<i>Float64</i>	0	2500000000.0
Kamar Tidur	<i>Int64</i>	0	4
Kamar Mandi	<i>Int64</i>	0	3
Garasi	<i>Int64</i>	0	10
Luas Tanah	<i>Float64</i>	0	438.0
Luas Bangunan	<i>Float64</i>	0	110.0

3.2. Optimasi Hyperparameter berbasis Optuna

Berbeda dengan model dasar yang menggunakan parameter bawaan (*baseline*), optimasi menggunakan Optuna mengeksplorasi ruang pencarian secara dinamis. Optuna mengonvergensi parameter optimal yang meminimalkan tingkat kesalahan pada model. Konfigurasi terbaik yang dihasilkan adalah kombinasi dari $n_estimators$: 483, max_depth : 8, dan $learning_rate$: 0.0346. Selain itu, parameter regularisasi L1 (reg_alpha : 0.011) dan L2 (reg_lambda : 0.181) yang ditemukan berperan penting untuk mengontrol bobot pada pohon regresi, sehingga mencegah kompleksitas berlebih pada algoritma XGBoost.

3.3. Evaluasi Kinerja dan Komparasi Model

Karakterisasi kinerja dari kedua pendekatan XGBoost dinilai berdasarkan pengujian pada data uji yang belum pernah dilihat oleh model. Hasil komparasi presisi ini dirangkum pada Tabel 4.

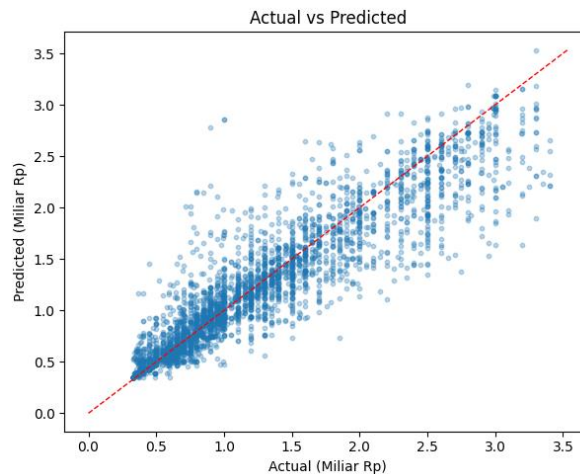
Tabel 4. Perbandingan Kinerja *Baseline* XGBoost dan XGBoost-Optuna

Metode	RMSE	MAE	MAPE (%)	R^2	Waktu Komputasi (detik)
<i>Baseline</i>	306.258.500,92	201.395.200,82	16,76	0,8204	2,94
Optuna	301.132.090,80	196.869.100,79	16,53	0,8263	3.550,34

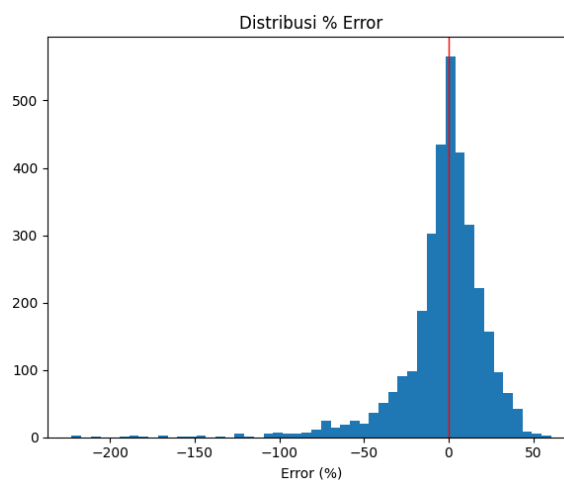
Data pada Tabel 4 menegaskan bahwa penerapan Optuna meningkatkan seluruh metrik evaluasi akurasi. Semakin rendah skor kesalahan absolut (MAE), semakin superior model prediksinya. Pada eksperimen ini, nilai MAE berhasil ditekan dari Rp 201.395.200,82 menjadi Rp 196.869.100,79, yang sejalan dengan penurunan deviasi kuadratik (RMSE) dan persentase simpangan error (MAPE). Hal ini diiringi dengan peningkatan kemampuan model dalam menjelaskan varians data, di mana koefisien determinasi (R^2) naik dari 0.8204 menjadi 0.8263.

Selain itu, model yang dioptimasi menunjukkan ketahanan yang baik terhadap *overfitting*. Pemeriksaan internal menunjukkan bahwa model mencapai skor R^2 sebesar 0.8823 pada data pelatihan dan 0.8284 pada data pengujian. Jarak (gap) performa yang sangat sempit, yakni hanya 0.0539, membuktikan bahwa kemampuan generalisasi model pada data baru berfungsi dengan baik dan stabil.

Distribusi akurasi prediksi ini divisualisasikan lebih lanjut pada Gambar 2. Sebaran titik pada *plot Actual vs Predicted* terkonsentrasi secara proporsional di sepanjang garis diagonal ideal, mengindikasikan prediksi yang searah dengan harga aktual. Selanjutnya, pada Gambar 3. Histogram persentase kesalahan menunjukkan distribusi normal dengan residu yang terpusat erat di angka nol, memvalidasi ketiadaan bias sistematis yang parah pada model.



Gambar 2. Distribusi Akurasi Prediksi



Gambar 3. Distribusi Persentase Error

3.4. Diskusi

Hasil dari penelitian ini secara langsung menjawab pertanyaan tujuan penelitian yang dirumuskan di awal. Pertama, mengenai pengaruh integrasi Optuna, hasil menunjukkan bahwa pencarian *hyperparameter* pada Optuna mampu mereduksi *error* secara konsisten dan terbukti sebagai pilihan berharga untuk meningkatkan akurasi. RMSE berhasil diturunkan secara spesifik sebesar Rp 5.126.410,12 dan MAE berhasil dipangkas sebesar Rp 4.526.100,03 dibandingkan model bawaan. Ini membuktikan bahwa pengaturan parameter tree dan regularisasi oleh Optuna sukses memetakan hubungan yang rumit (*non-linear*) antara fitur *input* dengan harga akhir properti.

Kedua, model XGBoost yang dioptimasi terbukti menghasilkan performa prediksi yang lebih unggul dan tangguh dibandingkan *baseline*. Kecilnya gap antara data latih dan data uji mengonfirmasi bahwa model tidak menghafal data secara buta (*overfitting*), sehingga sangat stabil meskipun menghadapi data transaksi properti sekunder yang sangat fluktuatif. Meskipun demikian, peningkatan akurasi sebesar 0,59% pada R^2 perlu dikaji dalam konteks biaya komputasi yang dikeluarkan. Waktu pelatihan meningkat drastis dari 2,94 detik pada model *baseline* menjadi 3.550,34 detik pada model Optuna, atau sekitar 1.000 kali lebih lama. Dalam konteks penelitian dan pengembangan model (*one-time training*), trade-off ini masih dapat diterima karena proses pelatihan hanya dilakukan sekali dan tidak mempengaruhi kecepatan inferensi. Namun, untuk skenario yang membutuhkan pelatihan ulang secara berkala, efisiensi komputasi menjadi pertimbangan kritis yang perlu diatasi, misalnya dengan mengaktifkan mekanisme pruning atau membatasi jumlah trial pada Optuna.

4. KESIMPULAN

Berdasarkan hasil analisis dan evaluasi yang telah dilakukan, penelitian ini menyimpulkan bahwa integrasi optimasi *hyperparameter* berbasis Optuna pada algoritma Extreme Gradient Boosting (XGBoost) berhasil meningkatkan keakurasian prediksi harga rumah. Fakta empiris menunjukkan bahwa model XGBoost-Optuna mengungguli model *baseline* (bawaan) pada seluruh metrik evaluasi keakurasian; dengan penurunan tingkat kesalahan *Root Mean Square Error* (RMSE) menjadi Rp 301.132.090,80, penurunan *Mean Absolute*

Error (MAE) menjadi Rp 196.869.100,79, dan persentase kesalahan absolut (MAPE) yang menyusut menjadi 16,53%. Kemampuan model dalam menjelaskan varians data juga meningkat, ditandai dengan kenaikan nilai koefisien determinasi (R^2) dari 0.8204 menjadi 0.8263. Selain itu, model yang dioptimasi terbukti stabil dan tidak mengalami *overfitting* yang parah, dibuktikan dengan selisih (*gap*) performa R^2 yang sangat kecil (0.0539) antara data pelatihan dan data pengujian. Meskipun demikian, terdapat fakta bahwa peningkatan akurasi ini menuntut biaya komputasi yang sangat masif, di mana waktu pelatihan model melonjak drastis dari 2,94 detik menjadi 3.550,34 detik.

Secara praktis, model prediksi ini dapat diaplikasikan oleh agen *real estate*, platform teknologi properti (*proptech*), perusahaan penilai aset, maupun lembaga perbankan (untuk KPR) sebagai alat bantu penaksiran harga properti sekunder yang objektif dan berbasis data (*data-driven*). Implikasi dari penelitian ini menegaskan bahwa pencarian *hyperparameter* secara dinamis sangat krusial untuk memetakan pola harga properti yang *non-linear*. Sebagai spekulasi, jika model ini diintegrasikan ke dalam sistem aplikasi harian, hal ini berpotensi besar dalam menekan asimetri informasi antara penjual dan pembeli, sehingga menciptakan ekosistem pasar perumahan yang lebih transparan dan efisien.

Untuk penelitian lebih lanjut, disarankan untuk mengevaluasi keseimbangan (*trade-off*) antara durasi komputasi dan perolehan akurasi. Peneliti selanjutnya dapat membatasi jumlah percobaan (*trials*) pada Optuna atau menerapkan metode pemangkasan (*pruning*) yang lebih agresif guna mempercepat waktu pelatihan tanpa mengorbankan performa secara signifikan. Selain itu, disarankan untuk memperluas cakupan *dataset* di luar wilayah Depok (seperti Jabodetabek) serta mengintegrasikan variabel eksternal tambahan seperti jarak ke fasilitas transportasi, tingkat suku bunga, maupun kondisi makroekonomi untuk menghasilkan model prediksi yang lebih komprehensif dan berdaya guna luas. Selain itu, integrasi fitur spasial seperti jarak ke stasiun MRT/KRL, akses jalan tol, atau kedekatan dengan pusat perbelanjaan menggunakan data dari OpenStreetMap atau Google Maps API sangat direkomendasikan, mengingat faktor aksesibilitas memiliki korelasi kuat terhadap nilai properti dan berpotensi meningkatkan akurasi prediksi secara lebih signifikan dibandingkan optimasi *hyperparameter* semata.

REFERENSI

- [1] H. Sharma, H. Harsora, and B. Ogunleye, "An Optimal House Price Prediction Algorithm: XGBoost," *Analytics*, vol. 3, no. 1, pp. 30–45, Jan. 2024, doi: 10.3390/analytics3010003.
- [2] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *J. Appl. Comput. Sci. Technol.*, vol. 4, no. 1, pp. 58–64, Jul. 2023, doi: 10.52158/jacost.v4i1.491.
- [3] N. V. Setia, P. A. Bella, and L. J. Tjung, "Faktor – Faktor Lokasi Yang Mempengaruhi Harga Jual Rumah Di Kecamatan Sawangan Kota Depok," *j.sains.teknol.n.a.arsit.indones.*, vol. 6, no. 1, pp. 669–684, Apr. 2024, doi: 10.24912/stupa.v6i1.27509.
- [4] F. Trindade Neves, M. Aparicio, and M. De Castro Neto, "The Impacts of Open Data and eXplainable AI on Real estate Price Predictions in Smart Cities," *Applied Sciences*, vol. 14, no. 5, p. 2209, Mar. 2024, doi: 10.3390/app14052209.
- [5] R. Muhammad Fadzryan and E. Angga Laksana, "Optimasi Prediksi Harga Rumah dengan Random Forest dan Optuna Hyperparameter Tuning," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 5, no. 6, pp. 1663–1671, Jun. 2025, doi: 10.52436/1.jpti.846.
- [6] W. Kim and J. Hong, "Stacked Ensemble Model for the Automatic Valuation of Residential Properties in South Korea: A Case Study on Jeju Island," *Land*, vol. 13, no. 9, p. 1436, Sep. 2024, doi: 10.3390/land13091436.
- [7] K. Shubham, S. Metya, A. K. Sinha, M. Simon, and A. K. Yadav, "XGBoost based surrogate technique for system reliability analysis of foundation over cavity aided with bootstrapping," *Sci Rep*, vol. 16, no. 1, p. 7113, Feb. 2026, doi: 10.1038/s41598-026-37058-0.
- [8] K. Xing, M. Feng, I. A. Bonev, H. Champlaud, M. Cheriet, and Z. Liu, "XGBoost-Based Digital Twin Model for Predicting Trajectory Errors in a Hexapod Coordinated Machining System Using Positioning Accuracy and Vibration Data," *Sensors*, vol. 25, no. 23, p. 7142, Nov. 2025, doi: 10.3390/s25237142.
- [9] A. Mehdary, A. Chehri, A. Jakimi, and R. Saadane, "Hyperparameter Optimization with Genetic Algorithms and XGBoost: A Step Forward in Smart Grid Fraud Detection," *Sensors*, vol. 24, no. 4, p. 1230, Feb. 2024, doi: 10.3390/s24041230.
- [10] C.-Y. Ku, C.-Y. Liu, and T.-Y. Wu, "AI-Driven Optimization for Efficient Public Bus Operations," *Mathematics*, vol. 13, no. 20, p. 3249, Oct. 2025, doi: 10.3390/math13203249.
- [11] L.-H. Lai *et al.*, "The Use of Machine learning Models with Optuna in Disease Prediction," *Electronics*, vol. 13, no. 23, p. 4775, Dec. 2024, doi: 10.3390/electronics13234775.
- [12] R. M. Munshi *et al.*, "Optimising hyperparameters with a tree structured Parzen estimator to improve diabetes prediction," *Sci Rep*, vol. 15, no. 1, p. 35430, Oct. 2025, doi: 10.1038/s41598-025-19295-x.
- [13] G. Hariharan, R. Mallya, N. Kumar, D. Doreswamy, G. M. Chennegowda, and S. K. Bhat, "Optuna-Optimized Ensemble and Neural Network Models for Static Characteristics Prediction of Active

- Bearings with Geometric Adjustments,” *Modelling*, vol. 6, no. 3, p. 98, Sep. 2025, doi: 10.3390/modelling6030098.
- [14] A. S. AlSalehy and M. Bailey, “Improving Time Series Data Quality: Identifying *Outliers* and Handling *Missing Values* in a Multilocation Gas and Weather *Dataset*,” *Smart Cities*, vol. 8, no. 3, p. 82, May 2025, doi: 10.3390/smartcities8030082.
- [15] A. Neftissov, A. Biloshchytskyi, I. Kazambayev, S. Dolhopolov, and T. Honcharenko, “An Advanced Ensemble *Machine learning Framework* for Estimating Long-Term Average Discharge at Hydrological Stations Using Global Metadata,” *Water*, vol. 17, no. 14, p. 2097, Jul. 2025, doi: 10.3390/w17142097.
- [16] D. Ashourloo *et al.*, “*Machine learning* approaches for wheat yield prediction integrating biophysical modeling and remote sensing: Effects of sample size, dimensionality, and transferability,” *Smart Agricultural Technology*, vol. 13, p. 101936, Mar. 2026, doi: 10.1016/j.atech.2026.101936.
- [17] V. Swarnkar, S. Ralhan, M. Singh, D. Parashar, and M. Singh, “Correlation based feature importance analysis for improving *machine learning* stability predictions in hybrid PV systems,” *Sci Rep*, vol. 16, no. 1, p. 10041, Feb. 2026, doi: 10.1038/s41598-026-37270-y.
- [18] M. D. Gajewski, P. Xia, B. Gajewska, J. Pais, and M. Miecznikowski, “Estimation of the XGBoost Regression Model Used in the Prediction of Pavement’s Mechanical and Geometrical Parameters Based on Static Interpretation of the FWD Test,” *Applied Sciences*, vol. 15, no. 24, p. 12943, Dec. 2025, doi: 10.3390/app152412943.
- [19] F. A. P. Sadewa and D. Kurniawan, “Application of ADASYN and Optuna in the XGBoost Algorithm for Stunting Detection,” vol. 10, no. 1.
- [20] Y. Cai, J. Feng, Y. Wang, Y. Ding, Y. Hu, and H. Fang, “The Optuna–LightGBM–XGBoost Model: A Novel Approach for Estimating Carbon Emissions Based on the Electricity–Carbon Nexus,” *Applied Sciences*, vol. 14, no. 11, p. 4632, May 2024, doi: 10.3390/app14114632.
- [21] M. Imani and H. R. Arabnia, “*Hyperparameter Optimization* and Combined Data Sampling Techniques in *Machine learning* for Customer Churn Prediction: A Comparative Analysis,” *Technologies*, vol. 11, no. 6, p. 167, Nov. 2023, doi: 10.3390/technologies11060167.
- [22] T. Dong *et al.*, “Enhancing Cardiovascular Risk Prediction: Development of an Advanced Xgboost Model with Hospital-Level Random Effects,” *Bioengineering*, vol. 11, no. 10, p. 1039, Oct. 2024, doi: 10.3390/bioengineering11101039.
- [23] J. Sun *et al.*, “Construction and *Optimization* of Integrated Yield Prediction Model Based on Phenotypic Characteristics of Rice Grown in Small–Scale Plantations,” *Agriculture*, vol. 15, no. 2, p. 181, Jan. 2025, doi: 10.3390/agriculture15020181.