



Optimasi *Random Forest* untuk Prediksi Harga Rumah menggunakan Seleksi Fitur dan Penyetelan Hyperparameter

Roberto Reyna Wijaya¹, Muhammad Eka Purbaya^{2*}, Rozzan Zhofrul Islam³

¹⁻³ Program Studi Bisnis Digital, Universitas Telkom, Indonesia

Email: robertreyw@student.telkomuniversity.ac.id¹, mekapur@telkomuniversity.ac.id², rozzanzhofrulislam@student³

*Korespondensi penulis: mekapur@telkomuniversity.ac.id

Abstract. House price prediction has become an important need in the property sector because selling prices are influenced by various factors, such as building quality, living area size, garage capacity, and physical property conditions. The complexity of relationships among these variables makes machine learning approaches, particularly Random Forest Regression, relevant for developing accurate and stable prediction models. This article discusses the optimization of Random Forest Regression for house price prediction through feature selection and hyperparameter tuning. The dataset used in this study is the House Prices: Advanced Regression Techniques dataset from Kaggle, with SalePrice as the target variable. The analysis process includes data exploration, data preprocessing, feature selection, train-test data splitting, Random Forest model development, and hyperparameter optimization using Randomized Search. Model performance was evaluated using R-Squared, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The results show that the optimized model achieved an R-Squared value of 0.8927 and an RMSE value of 28,687.08. The most dominant features influencing house prices were OverallQual, GrLivArea, and GarageCars. Although the performance improvement after optimization was relatively modest, Random Forest proved capable of producing stable predictions and can be used as a decision-support tool in the property business. However, the model still has limitations in predicting extreme prices in the luxury housing segment.

Keywords: Feature Selection; House Price; Hyperparameter Tuning; Machine Learning; Property Predictio.

Abstrak. Prediksi harga rumah menjadi salah satu kebutuhan penting dalam bidang properti karena harga jual dipengaruhi oleh banyak faktor, seperti kualitas bangunan, luas area hunian, kapasitas garasi, dan kondisi fisik rumah. Kompleksitas hubungan antar variabel tersebut membuat pendekatan *machine learning*, khususnya *Random Forest Regression*, relevan digunakan untuk membangun model prediksi yang akurat dan stabil. Artikel ini membahas optimasi Random Forest Regression dalam memprediksi harga rumah melalui seleksi fitur dan penyetelan hyperparameter. Dataset yang digunakan adalah House Prices: Advanced Regression Techniques dari Kaggle, dengan variabel target berupa SalePrice. Tahapan analisis meliputi eksplorasi data, praproses data, pemilihan fitur, pembagian data latih dan data uji, pembangunan model Random Forest, serta optimasi hyperparameter menggunakan Randomized Search. Evaluasi model dilakukan menggunakan metrik R-Squared, Root Mean Squared Error (RMSE), dan Mean Absolute Error (MAE). Hasil pengujian menunjukkan bahwa model setelah optimasi menghasilkan nilai R-Squared sebesar 0,8927 dan RMSE sebesar 28.687,08. Fitur yang paling dominan dalam memengaruhi harga rumah adalah OverallQual, GrLivArea, dan GarageCars. Meskipun peningkatan performa setelah optimasi tidak terlalu besar, Random Forest terbukti mampu menghasilkan prediksi yang stabil dan dapat digunakan sebagai alat bantu pengambilan keputusan dalam bisnis properti. Namun, model masih memiliki keterbatasan dalam memprediksi harga ekstrem pada segmen rumah mewah.

Kata Kunci: Harga Rumah; Pembelajaran Mesin; Penyetelan Hiperparameter; Prediksi Harga Properti; Seleksi Fitur.

1. LATAR BELAKANG

Perkembangan sektor properti dalam beberapa tahun terakhir menunjukkan bahwa harga rumah semakin dipengaruhi oleh kombinasi faktor fisik, spasial, dan karakteristik lingkungan yang kompleks (Sa'diyah et al., 2025). Dalam konteks ini, prediksi harga rumah menjadi persoalan penting karena tidak hanya berkaitan dengan nilai investasi, tetapi juga dengan pengambilan keputusan bagi pembeli, pengembang, dan pembuat kebijakan. Sejumlah

penelitian terbaru menunjukkan bahwa penilaian harga properti kini banyak bergeser dari pendekatan *hedonic* tradisional menuju pendekatan *machine learning* yang lebih adaptif terhadap pola data non-linier dan interaksi antar-fitur (Liu, 2024; Zaki, Nayyar, Dalal, & Ali, 2022).

Salah satu pendekatan yang banyak digunakan dalam prediksi harga rumah adalah *Random Forest Regression*. Metode ini dipilih karena memiliki kemampuan yang baik dalam menangani data tabular, hubungan nonlinier, serta interaksi antar-variabel yang kompleks (Hapsari & Fauzani, 2025). Berbagai studi juga menunjukkan bahwa model berbasis *ensemble*, termasuk *Random Forest* dan variasinya, mampu memberikan performa yang kompetitif dalam tugas prediksi harga properti (Erbulut & Çolak, 2025; Mishra et al., 2024; Tepe, 2024). Meski demikian, keberhasilan model sangat dipengaruhi oleh kualitas fitur yang digunakan dan penyesuaian parameter agar model dapat bekerja secara optimal pada data yang bervariasi.

Tantangan utama dalam prediksi harga rumah muncul pada segmen harga ekstrem, khususnya properti mewah. Beberapa penelitian menunjukkan bahwa model prediksi sering mengalami *underprediction* pada rumah berharga tinggi karena fitur premium tidak terdistribusi secara merata dalam data pelatihan (Dsilva, Kadam, Kadam, & Deshpande, 2025). Selain itu, distribusi harga yang miring dan adanya *outlier* dapat memengaruhi stabilitas model serta menyebabkan perbedaan antara hasil evaluasi berbasis RMSE dan MAE. RMSE lebih sensitif terhadap kesalahan besar, sedangkan MAE cenderung lebih stabil dalam menggambarkan rata-rata kesalahan prediksi (Hodson, 2022; Karunasingha, 2022). Oleh karena itu, evaluasi model tidak cukup hanya melihat akurasi rata-rata, tetapi juga perlu memperhatikan perilaku model pada nilai ekstrem (Fauziyyah et al., 2026).

Di sisi lain, penelitian-penelitian terdahulu lebih banyak berfokus pada perbandingan algoritma atau peningkatan akurasi umum, tanpa secara mendalam membahas stabilitas model pada distribusi harga yang kompleks. Padahal, dalam konteks prediksi rumah, aspek seperti seleksi fitur, optimasi *hyperparameter*, dan analisis *residual* memiliki peran penting untuk memastikan model benar-benar andal pada berbagai segmen pasar (Andrade-Girón, Marin-Rodriguez, & Zuñiga-Rojas, 2025; Man & Chan, 2021; Pfeifer, Holzinger, & Schimek, 2022). Kesenjangan ini menunjukkan perlunya pendekatan yang tidak hanya akurat secara statistik, tetapi juga lebih *robust* terhadap variasi data properti (Chantika et al., 2025).

Berdasarkan latar belakang tersebut, penelitian ini mengembangkan model *Random Forest Regression* untuk prediksi harga rumah dengan mengombinasikan *feature selection* dan *hyperparameter tuning*. Pendekatan ini dipilih untuk memperoleh model yang lebih efisien, lebih terfokus pada fitur-fitur paling relevan, dan lebih optimal dalam memprediksi harga

rumah pada data dengan karakteristik *nonlinier* dan sebaran yang tidak seimbang. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi baik secara metodologis maupun praktis dalam pengembangan model prediksi harga properti yang lebih andal (Bili et al., 2026).

Adapun tujuan penelitian ini adalah mengidentifikasi variabel-variabel yang paling berpengaruh terhadap harga jual rumah, mengoptimalkan performa *Random Forest Regression* melalui seleksi fitur dan penyetelan parameter, serta menganalisis kemampuan model dalam menangani variasi harga pada rentang yang berbeda. Hasil penelitian diharapkan dapat menjadi dasar pengembangan model prediksi harga rumah yang lebih akurat dan dapat diterapkan pada konteks analisis properti yang lebih luas.

Struktur artikel ini disusun untuk memberikan gambaran menyeluruh mengenai proses penelitian. Setelah bagian Pendahuluan, bagian Kajian Teoritis membahas landasan teori serta penelitian terdahulu terkait estimasi harga properti. Bagian Metode Penelitian menjelaskan dataset *Ames Housing*, tahapan pra-pemrosesan data, serta skenario optimasi model. Bagian Hasil dan Pembahasan menyajikan temuan eksperimen, perbandingan metrik evaluasi, serta analisis signifikansi fitur. Terakhir, penelitian ini ditutup dengan bagian Kesimpulan dan Rekomendasi, yang merangkum kontribusi ilmiah dan memberikan saran untuk pengembangan model di masa depan.

2. KAJIAN TEORITIS

Analisis Regresi dalam Prediksi Properti

Analisis regresi merupakan metode statistik yang digunakan untuk memodelkan hubungan fungsional antara variabel dependen, yaitu harga jual, dengan sejumlah variabel independen yang merepresentasikan karakteristik properti. Dalam konteks harga rumah, hubungan antarvariabel sering kali bersifat non-linear dan dipengaruhi oleh distribusi data yang asimetris atau condong ke kanan (*right-skewed*) (Lin, Liu, & Sun, 2024). Secara matematis, model regresi bertujuan untuk meminimalkan selisih antara nilai aktual (y) dan nilai prediksi ($\hat{y}$), yang direpresentasikan melalui fungsi berikut:

$$y = \beta_0 + \sum_{i=1}^n \beta_i X_i + \dots\dots\dots(i)$$

Variabel X_i mencakup fitur-fitur dominan seperti General Living Area (GrLivArea) dan Material Quality (Overall Qual), yang secara empiris telah terbukti memiliki korelasi kuat dengan nilai pasar properti (Ren, 2024).

Mekanisme Regresi *Random Forest*

Random Forest merupakan algoritma berbasis *ensemble learning* yang bekerja dengan membangun banyak pohon keputusan secara paralel melalui teknik *bagging* atau *bootstrap aggregating* (Ren, 2024). Algoritma ini dipilih karena memiliki ketahanan yang baik dalam menangani interaksi fitur yang kompleks serta mampu mengurangi dampak *outlier* pada dataset *real estate* (Yu, 2023). Prediksi akhir model ($\hat{y}$) diperoleh dengan menghitung rata-rata hasil prediksi dari seluruh pohon individu $\{T_1, T_2, \dots, T_k\}$:

$$\hat{y} = (1/k) \sum T_i(x) \dots \dots \dots (ii)$$

Metrik Evaluasi Statistik

Keandalan model dalam mengestimasi harga rumah diukur menggunakan dua metrik evaluasi utama yang memberikan sudut pandang berbeda terhadap distribusi kesalahan (Dritsas, Trigka, & Mylonas, 2025). *Root Mean Squared Error* (RMSE): Metrik ini memberikan penalti kuadratik terhadap kesalahan yang besar, sehingga sangat efektif untuk mendeteksi penyimpangan prediksi pada harga rumah mewah yang sering menjadi *outlier*.

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \dots \dots \dots (iii)$$

Mean Absolute Error (MAE): Metrik ini memberikan rata-rata linear dari nilai kesalahan absolut, sehingga lebih stabil ketika menghadapi variansi *error* yang tidak konstan atau heteroskedastisitas.

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \dots \dots \dots (iv)$$

Penelitian Terdahulu

Penelitian mengenai estimasi harga properti telah mengalami transformasi yang signifikan seiring dengan perkembangan teknik *machine learning* dan pemanfaatan dataset publik seperti *Ames Housing* sumber data. Berbagai studi terbaru berupaya menguraikan efektivitas algoritma *ensemble* untuk mengidentifikasi model prediksi yang tidak hanya akurat secara statistik, tetapi juga stabil dalam menghadapi fluktuasi pasar. **Tabel 1** merangkum beberapa kontribusi ilmiah utama yang menjadi dasar sekaligus titik awal pengembangan penelitian ini.

Tabel 1. Dinamika Penelitian Prediksi Harga Properti (2021–2026).

Penulis & Tahun	Pendekatan Utama	Temuan Utama	Kesenjangan Penelitian yang Tersisa
(Thakur & Bisht, 2025)	Ensemble Stacking	Penggabungan model melalui teknik stacking terbukti efektif menurunkan nilai RMSE secara signifikan.	Fokus penelitian masih pada akurasi global dan belum membahas anomali harga pada segmen properti kelas atas.
(Andrade-Girón et al., 2025)	Seleksi Fitur Cerdas	Seleksi fitur berbasis informatika mampu meningkatkan stabilitas model pada data real estat.	Belum mengeksplorasi perbandingan mekanisme antara teknik bagging dan boosting.
(Dsilva et al., 2025)	Random Forest, XGBoost, Pembelajaran Mendalam	Teknik ensemble secara konsisten mengungguli pembelajaran mendalam pada jenis data tabular.	Fenomena prediksi rendah (underprediction) pada harga ekstrem masih menjadi tantangan yang belum terselesaikan.
(Dritsas et al., 2025)	Pendekatan Pembelajaran Mesin	Penanganan distribusi data yang miring (<i>skewness</i>) melalui prapemrosesan data menjadi faktor penentu dalam keandalan model.	Analisis mengenai pengaruh fitur tertentu terhadap stabilitas model pada rentang harga tinggi masih terbatas.
(Ren, 2024)	Penyetelan Hiperparameter	Optimasi parameter yang tepat memungkinkan akurasi Random Forest bersaing secara ketat dengan model boosting.	Belum menguji efektivitas seleksi fitur pada dataset berdimensi tinggi.
(Liu, 2024)	Regresi Nonlinier	Sifat harga properti yang nonlinier menjadikan penggunaan model berbasis pohon keputusan lebih relevan.	Pembahasan masih bersifat umum dan belum memasuki tahap optimasi lanjutan.

Sumber: Dianalisis oleh peneliti.

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif berbasis *machine learning* untuk membangun model prediksi harga rumah. Metode utama yang digunakan adalah *Random Forest Regression*, dengan dua tahap optimasi, yaitu seleksi fitur dan penyetelan *hyperparameter*, agar model dapat menghasilkan prediksi yang lebih akurat dan efisien. Seluruh proses analisis dilakukan menggunakan *Google Colaboratory* sebagai lingkungan pengolahan data dan pemrograman.

Dataset

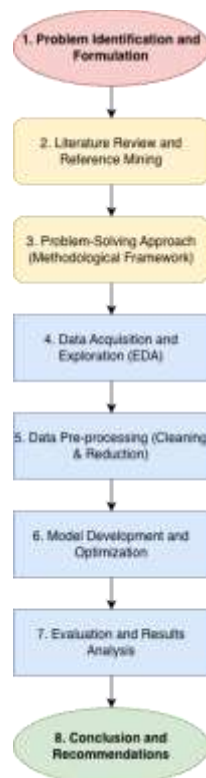
Dataset yang digunakan pada penelitian ini adalah House Prices: *Advanced Regression Techniques* yang diperoleh dari Kaggle. Dataset ini berisi data karakteristik rumah dan harga jual properti, sehingga sesuai untuk membangun model prediksi harga rumah. Data tersebut mencakup berbagai atribut seperti luas tanah, kualitas bangunan, jumlah kamar, kapasitas garasi, tahun pembangunan, kondisi *basement*, serta variabel target berupa SalePrice sebagai harga jual rumah. Dataset ini dipilih karena memiliki jumlah fitur yang cukup lengkap dan relevan untuk analisis prediktif.

Secara umum, dataset terdiri dari data numerik dan kategorikal yang menggambarkan kondisi fisik serta atribut lingkungan rumah. Oleh karena itu, sebelum pemodelan dilakukan, data perlu dibersihkan, disiapkan, dan dipilih fitur yang paling berpengaruh terhadap harga jual. Tahap ini penting untuk memastikan bahwa model yang dibangun hanya menggunakan informasi yang relevan dan tidak terdistraksi oleh fitur yang kurang informatif.

Alur Penelitian

Alur penelitian pada studi ini mengikuti tahapan yang ditunjukkan pada gambar 1. Secara umum, alur tersebut terdiri atas delapan tahap utama, yaitu: identifikasi masalah, studi pustaka, pemilihan pendekatan, pengumpulan data, pra-pemrosesan data, pengembangan model, evaluasi hasil, serta penarikan kesimpulan dan rekomendasi.

Pada tahap identifikasi masalah, peneliti merumuskan kebutuhan untuk memprediksi harga rumah secara akurat berdasarkan karakteristik properti. Tahap ini menjadi dasar dalam menentukan tujuan penelitian, yaitu menghasilkan model prediksi yang mampu menangkap hubungan antara fitur rumah dan harga jual. Tahap berikutnya adalah studi pustaka, yaitu mengkaji penelitian-penelitian terdahulu yang berkaitan dengan prediksi harga rumah, *Random Forest Regression*, seleksi fitur, dan optimasi *hyperparameter*. Studi pustaka ini digunakan untuk memperkuat landasan teori serta menentukan pendekatan metodologis yang paling sesuai.



Gambar 1. Diagram alir penelitian.

Setelah itu, dilakukan pemilihan pendekatan penelitian, yaitu penggunaan *Random Forest* sebagai algoritma utama karena mampu menangani data numerik dan kategorikal dengan baik serta memiliki ketahanan terhadap *overfitting*. Algoritma ini kemudian dioptimasi melalui seleksi fitur dan penyetelan *hyperparameter* agar performanya meningkat. Pada tahap pengumpulan data, dataset diambil dari Kaggle dan kemudian diimpor ke lingkungan Google Colaboratory untuk diproses lebih lanjut. Selanjutnya, data memasuki tahap praproses, yang meliputi pemeriksaan struktur data, pengecekan *missing value*, identifikasi duplikasi, penghapusan atribut yang tidak relevan atau memiliki terlalu banyak nilai kosong, serta pemilihan fitur yang paling berpengaruh terhadap variabel target.

Tahap pengembangan model dilakukan dengan membagi data menjadi data latih dan data uji, kemudian melatih model *Random Forest Regression* menggunakan fitur-fitur terpilih. Setelah model dasar dibangun, dilakukan penyetelan *hyperparameter* untuk memperoleh kombinasi parameter terbaik, seperti jumlah pohon, kedalaman maksimum pohon, dan jumlah minimum sampel pada setiap percabangan.

Selanjutnya, model dievaluasi menggunakan metrik kinerja seperti R-squared, RMSE, dan MAE. Hasil evaluasi model sebelum dan sesudah optimasi dibandingkan untuk mengetahui apakah seleksi fitur dan penyetelan *hyperparameter* berhasil meningkatkan performa prediksi. Tahap terakhir adalah penarikan kesimpulan dan pemberian rekomendasi berdasarkan hasil analisis. Kesimpulan difokuskan pada efektivitas model *Random Forest* dalam memprediksi harga rumah, sedangkan rekomendasi diarahkan pada pengembangan metode lebih lanjut atau penerapan model pada data properti yang lebih luas.

Alat dan Bahan

Penelitian ini menggunakan Google Colaboratory sebagai perlengkapan utama dalam proses pengolahan data, implementasi algoritma, dan evaluasi model. Pemilihan Google Colab didasarkan pada kemudahan akses, dukungan terhadap bahasa pemrograman Python, serta ketersediaan pustaka *machine learning* yang lengkap seperti pandas, NumPy, scikit-learn, dan matplotlib. Dengan demikian, seluruh tahapan analisis dapat dilakukan secara terstruktur dan efisien dalam satu lingkungan kerja berbasis *cloud*.

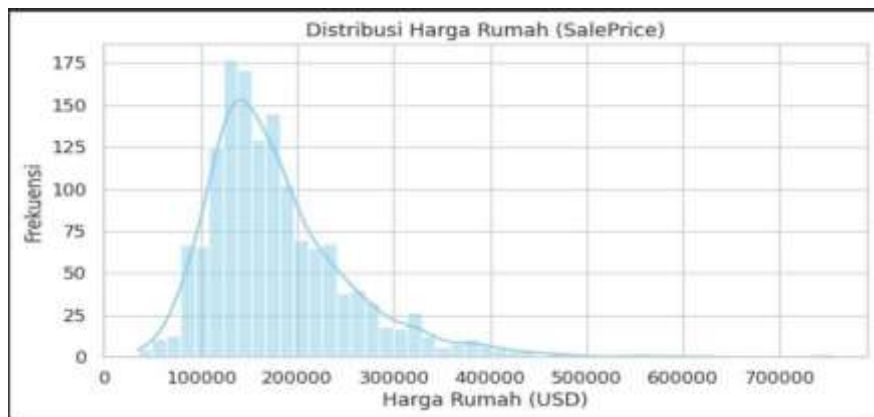
4. HASIL DAN PEMBAHASAN

Data Exploration

Tahap *data exploration* dilakukan untuk memahami karakteristik awal *dataset* sebelum masuk ke tahap *praproses* dan pemodelan. *Dataset* yang digunakan terdiri dari 1.460 baris dan 81 kolom yang memuat variabel numerik dan kategorikal terkait karakteristik properti serta

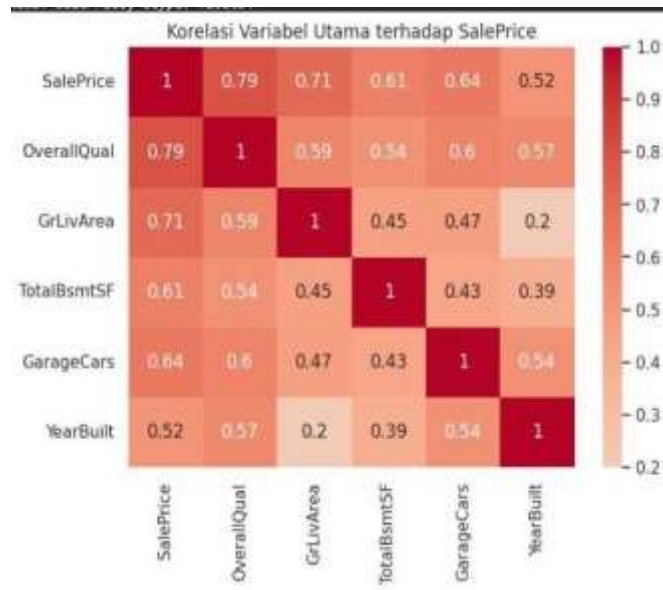
harga jual rumah. Pemeriksaan awal melalui tampilan data, struktur *dataset*, dan statistik deskriptif menunjukkan bahwa data memiliki variasi atribut yang cukup kompleks sehingga perlu dianalisis secara lebih mendalam.

Hasil eksplorasi variabel target SalePrice pada gambar 2 menunjukkan bahwa distribusi harga rumah cenderung miring ke kanan (*right-skewed*). Hal ini terlihat dari nilai rata-rata yang lebih tinggi dibandingkan median, yang mengindikasikan adanya sejumlah rumah dengan harga sangat tinggi. Kondisi ini penting karena keberadaan nilai ekstrem dapat memengaruhi hasil analisis dan performa model pada tahap berikutnya.

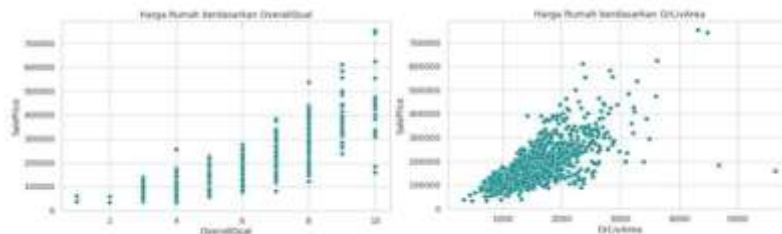


Gambar 2. Distribusi Harga Rumah (*SalePrice*).

Analisis korelasi pada gambar 3 memperlihatkan bahwa beberapa fitur memiliki hubungan yang cukup kuat dengan SalePrice, terutama OverallQual, GrLivArea, dan TotalBsmtSF. Temuan ini menunjukkan bahwa kualitas keseluruhan rumah, luas area hunian, dan luas basement merupakan faktor penting dalam menentukan harga rumah. Visualisasi *scatter plot* pada gambar 4 juga memperkuat hasil tersebut karena menunjukkan kecenderungan bahwa semakin tinggi kualitas dan luas rumah, maka harga jualnya cenderung meningkat.



Gambar 3. Korelasi Variabel Utama terhadap SalePrice.



Gambar 4. Visualisasi Hubungan Fitur Penting terhadap Harga Rumah.

Secara keseluruhan, hasil data *exploration* memberikan gambaran awal bahwa dataset memiliki potensi yang baik untuk digunakan dalam prediksi harga rumah, tetapi juga mengandung tantangan berupa distribusi yang tidak seimbang dan kemungkinan *outlier*. Oleh karena itu, tahap selanjutnya perlu difokuskan pada pembersihan data dan pemilihan fitur yang paling relevan agar model yang dibangun dapat menghasilkan prediksi yang lebih akurat dan stabil.

Data Preprocessing

Data preprocessing dilakukan untuk meningkatkan kualitas dataset sebelum digunakan dalam tahap pemodelan. Proses ini dimulai dengan pemeriksaan missing value dan data duplikat guna memastikan bahwa dataset memiliki struktur yang bersih dan konsisten. Hasil pemeriksaan menunjukkan bahwa sebagian besar variabel memiliki data lengkap, meskipun terdapat beberapa atribut dengan nilai hilang yang cukup besar.

Selanjutnya, dilakukan penghapusan beberapa kolom yang memiliki persentase *missing value* sangat tinggi, seperti PoolQC, MiscFeature, Alley, dan Fence, karena variabel tersebut dinilai kurang informatif dan sulit diimputasi secara akurat. Setelah itu, dilakukan seleksi fitur untuk mempertahankan variabel yang paling relevan terhadap SalePrice agar model lebih

efisien dan fokus pada atribut yang berpengaruh terhadap harga rumah. Pada tahap akhir, data dipisahkan menjadi variabel *input* (X) dan variabel *target* (y) sebagai persiapan untuk proses pelatihan model. Tahapan ini penting karena kualitas data yang baik akan sangat menentukan performa model dalam melakukan prediksi harga rumah secara akurat.

Model Development and Optimization

Pada tahap ini, model prediksi harga rumah dibangun menggunakan algoritma Random Forest Regression. Pemilihan algoritma ini didasarkan pada kemampuannya dalam menangani data tabular, memodelkan hubungan nonlinier antarvariabel, serta mengurangi risiko overfitting melalui kombinasi banyak pohon keputusan. Sebagai langkah awal, dibentuk model baseline untuk memperoleh gambaran performa awal sebelum dilakukan proses optimasi.

Setelah model awal dibangun, dilakukan optimasi melalui *hyperparameter tuning* menggunakan pendekatan RandomizedSearchCV. Proses ini bertujuan untuk menemukan kombinasi parameter terbaik agar model dapat belajar lebih optimal dari data. Ruang pencarian parameter yang diuji disajikan pada tabel 2 mencakup *n_estimators*, *max_depth*, *max_features*, *min_samples_split*, *min_samples_leaf*, dan *bootstrap*, dengan total 20 iterasi pencarian serta *cross-validation* sebanyak 3 kali. Evaluasi selama pencarian menggunakan *negative root mean squared error* sebagai dasar pemilihan konfigurasi terbaik.

Hasil optimasi menunjukkan bahwa model *Random Forest* yang telah disesuaikan memiliki konfigurasi parameter yang lebih sesuai dengan karakteristik dataset. Kombinasi parameter terbaik pada tabel 3 menghasilkan model yang lebih stabil dan lebih siap digunakan pada tahap evaluasi.

Tabel 2. Parameter yang Diuji.

Hyperparameter	Nilai yang Diuji
<i>n_estimators</i>	300, 600, 900
<i>max_depth</i>	None, 10, 20, 30
<i>max_features</i>	sqrt, log2
<i>min_samples_split</i>	2, 5, 10
<i>min_samples_leaf</i>	1, 2, 4
<i>bootstrap</i>	True, False

Sumber: Dianalisis oleh peneliti.

Tabel 3. Tabel Hasil Hyperparameter Terbaik.

Hyperparameter	Nilai Terbaik
<i>n_estimators</i>	900
<i>max_depth</i>	10
<i>max_features</i>	sqrt
<i>min_samples_split</i>	5
<i>min_samples_leaf</i>	2
<i>bootstrap</i>	True

Sumber: Dianalisis oleh peneliti.

Model Evaluation

Evaluasi model dilakukan untuk menilai kinerja *Random Forest Regression* dalam memprediksi harga rumah berdasarkan metrik R^2 , RMSE, dan MAE. Penggunaan ketiga metrik ini bertujuan untuk memberikan gambaran yang lebih komprehensif mengenai kemampuan model, baik dalam menjelaskan variasi data maupun dalam mengukur besar kesalahan prediksi.

Berdasarkan hasil pengujian yang disajikan pada tabel 4, model setelah optimasi menunjukkan nilai R^2 sebesar 0,89271, sedikit lebih tinggi dibandingkan model sebelum optimasi yang sebesar 0,89205. Hasil ini menunjukkan bahwa model yang telah dioptimasi memiliki kemampuan yang sedikit lebih baik dalam menjelaskan variasi harga rumah. Selain itu, nilai RMSE menurun dari 28.775,19 menjadi 28.687,08, yang mengindikasikan bahwa kesalahan prediksi model terhadap nilai aktual menjadi lebih kecil.

Di sisi lain, nilai MAE mengalami peningkatan dari 18.417,08 menjadi 18.656,25. Temuan ini menunjukkan bahwa meskipun optimasi berhasil menurunkan kesalahan besar yang tercermin pada RMSE, rata-rata kesalahan absolut model belum mengalami perbaikan. Perbedaan ini dapat terjadi karena RMSE lebih sensitif terhadap error ekstrem, sedangkan MAE menggambarkan rata-rata kesalahan secara lebih merata pada seluruh observasi.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa model setelah optimasi memiliki performa yang lebih baik secara umum, terutama dari sisi kemampuan menjelaskan variasi data dan menurunkan error besar. Namun, peningkatan tersebut masih bersifat terbatas karena tidak diikuti oleh penurunan MAE. Oleh karena itu, model hasil optimasi dapat dinilai lebih baik daripada model awal, tetapi interpretasinya tetap perlu dilakukan secara hati-hati dengan mempertimbangkan karakteristik masing-masing metrik evaluasi.

Tabel 4. Perbandingan Kinerja Model.

Metrik	Sebelum Optimasi	Setelah Optimasi	Perubahan
R^2	0.89205	0.89271	+ 0,00066
RMSE	28.775,19	28.687,08	- 88,11
MAE	18417,08	18656,249	+ 239,169

Sumber: Dianalisis oleh peneliti.

5. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma *Random Forest Regression* mampu membangun model prediksi harga rumah dengan performa yang baik. Hasil evaluasi pada data uji menunjukkan bahwa model setelah optimasi menghasilkan nilai R^2 sebesar 0,8927 dan RMSE sebesar 28.687,08, yang menandakan bahwa model mampu menjelaskan sebagian besar variasi harga rumah dengan tingkat kesalahan prediksi yang relatif rendah. Dengan demikian,

model yang dikembangkan layak digunakan sebagai pendekatan prediktif dalam estimasi harga properti berbasis data. Selain menghasilkan performa prediksi yang baik, penelitian ini juga mengidentifikasi bahwa OverallQual, GrLivArea, dan GarageCars merupakan variabel yang paling dominan dalam memengaruhi harga jual rumah. Temuan tersebut menegaskan bahwa kualitas bangunan, luas area hunian, dan kapasitas garasi merupakan faktor utama yang menentukan nilai pasar properti. Oleh karena itu, hasil penelitian ini tidak hanya memberikan kontribusi dalam pengembangan model prediktif, tetapi juga menyediakan informasi empiris yang dapat dimanfaatkan sebagai dasar pengambilan keputusan pada sektor properti.

Meskipun demikian, model masih memiliki keterbatasan dalam memprediksi rumah dengan harga ekstrem, khususnya pada segmen properti mewah, yang ditunjukkan oleh kecenderungan underprediction. Hal ini menunjukkan bahwa model lebih optimal digunakan pada segmen harga umum dibandingkan pada data dengan karakteristik yang sangat ekstrem. Oleh karena itu, penelitian selanjutnya disarankan untuk menambahkan fitur yang lebih representatif, menerapkan penanganan *outlier* yang lebih spesifik, serta membandingkan performa Random Forest Regression dengan algoritma lain agar hasil prediksi menjadi lebih akurat dan merata pada seluruh rentang harga rumah.

DAFTAR REFERENSI

- Andrade-Girón, D. C., Marin-Rodriguez, W. J., & Zuñiga-Rojas, M. G. (2025). Intelligent feature selection ensemble model for price prediction in real estate markets. *Informatics*, 12(2). <https://doi.org/10.3390/informatics12020052>
- Bili, I. K. M., Sudiarta, I. W., Wungabelen, M. Y., Rosdiana, N. K. A., & Rafiana, P. (2026). Analisis dan prediksi customer churn pada platform streaming berbasis langganan menggunakan metode random forest. *Jurnal Bisnis Inovatif dan Digital*, 3(1), 1–15. <https://doi.org/10.61132/jubid.v3i1.1226>
- Chantika, N. A. P., Herawati, Y., & Divianto. (2025). Analisis penerapan metode Altman Z-Score untuk prediksi kebangkrutan PT Kimia Farma tahun 2019–2023. *Jurnal Inovasi Manajemen, Kewirausahaan, Bisnis dan Digital*, 2(3), 246–257. <https://doi.org/10.61132/jimakebidi.v2i3.758>
- Dritsas, E., Trigka, M., & Mylonas, P. (2025). Evaluating machine learning approaches for residential property price estimation. *2025 10th South East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM 2025)*. <https://doi.org/10.1109/SEEDA-CECNSM68644.2025.11329623>
- Dsilva, M., Kadam, V., Kadam, H., & Deshpande, H. (2025). House price prediction: A comparative study of machine learning ensemble techniques and deep learning model. *Proceedings of the IEEE International Conference for Women in Innovation, Technology and Entrepreneurship (ICWITE 2025)*. <https://doi.org/10.1109/ICWITE64848.2025.11307001>

- Erbulut, Ö. G., & Çolak, Z. (2025). A hybrid machine learning approach for housing price prediction: The stacking regressor method. *International Journal of Housing Markets and Analysis*. <https://doi.org/10.1108/IJHMA-01-2025-0010>
- Fauziyyah, N. H., Sudiarsa, I. W., Sastradewi, I. A. E., Parisya, K. A. Y., & Sartika. (2026). Analisis prediksi customer churn pada sektor e-commerce berdasarkan perilaku transaksi menggunakan pendekatan machine learning. *Jurnal Manajemen Bisnis Digital Terkini*, 3(1), 300–318. <https://doi.org/10.61132/jumbidter.v3i1.1228>
- Hapsari, A. T., & Fauzani, M. M. (2025). Implementasi data mining untuk prediksi minat pembelian produk roti menggunakan metode apriori pada Toko Roti Dapur Bunda. *Jurnal Ekonomi dan Pembangunan Indonesia*, 3(1), 290–300. <https://doi.org/10.61132/jepi.v3i1.1249>
- Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, 15(14), 5481–5487. <https://doi.org/10.5194/gmd-15-5481-2022>
- Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, 585, 609–629. <https://doi.org/10.1016/j.ins.2021.11.036>
- Lin, Q., Liu, Z., & Sun, Y. (2024). To improve the students' learning motivation of vocational colleges through design thinking. <https://doi.org/10.3233/FAIA231496>
- Liu, J. (2024). Dataset analysis and house price prediction. In *Highlights in Science, Engineering and Technology* (Vol. 2023).
- Man, X., & Chan, E. P. (2021). The best way to select features? Comparing MDA, LIME, and SHAP. *Journal of Financial Data Science*, 3(1), 127–139. <https://doi.org/10.3905/jfds.2020.1.047>
- Mishra, A., Jason, Singh, Y., Rajakumar, P., Partheeban, N., & Indervati. (2024). House price prediction using machine learning. In *Proceedings of the IEEE 2024 1st International Conference on Advances in Computing, Communication and Networking (ICAC2N 2024)* (pp. 239–243). <https://doi.org/10.1109/ICAC2N63387.2024.10895110>
- Pfeifer, B., Holzinger, A., & Schimek, M. G. (2022). Robust random forest-based all-relevant feature ranks for trustworthy AI. *Studies in Health Technology and Informatics*, 294, 137–138. <https://doi.org/10.3233/SHTI220418>
- Ren, K. (2024). House price prediction based on machine learning algorithms: Taking Ames as an example. *Advances in Economics, Management and Political Sciences*, 85(1), 181–189. <https://doi.org/10.54254/2754-1169/85/20240870>
- Sa'diyah, H., Ilmi, A. P. Z., Mafikah, A. D., Amelia, R. N., Sari, S. I. N., & Latifah, E. (2025). Praktek jual beli pada marketplace dalam perspektif etika bisnis Islam. *MENAWAN: Jurnal Riset dan Publikasi Ilmu Ekonomi*, 3(1), 217–230. <https://doi.org/10.61132/menawan.v3i1.1218>
- Tepe, E. (2024). A random forests-based hedonic price model accounting for spatial autocorrelation. *Journal of Geographical Systems*, 26(4), 511–540. <https://doi.org/10.1007/s10109-024-00449-w>
- Thakur, B., & Bisht, A. (2025). Accurate house price prediction using ensemble machine learning model. *Proceedings of the 2025 IEEE 3rd International Symposium on Sustainable Energy, Signal Processing and Cybersecurity (ISSSC 2025)*. <https://doi.org/10.1109/iSSSC66652.2025.11388391>

Yu, J. (2023). Prediction on housing price based on the data on Kaggle. In *Proceedings of the 2022 3rd International Conference on E-Commerce and Internet Technology (ECIT 2022)* (pp. 627–634). Atlantis Press. https://doi.org/10.2991/978-94-6463-106-7_80