JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Prediksi Harga Rumah Menggunakan Algoritma Regresi Linier,Random Forest, Dan Gradient Boosting

Akhmadi*, Fikri Budiman

Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1,*]111202214699@mhs.dinus.ac.id, [2]fikri.budiman@dsn.dinus.ac.id
Email Penulis Korespondensi: 111202214699@mhs.dinus.ac.id
Submitted **24-11-2025;** Accepted **29-12-2025;** Published **31-12-2025**

Abstrak

Perkiraan harga rumah merupakan permasalahan penting dalam sektor properti karena dipengaruhi oleh berbagai faktor yang saling berkaitan, seperti karakteristik bangunan dan kondisi lingkungan, sehingga sering sulit diprediksi secara akurat menggunakan pendekatan konvensional dan berpotensi menimbulkan kesalahan dalam pengambilan keputusan. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan serta membandingkan kinerja model prediksi harga rumah menggunakan tiga algoritma *machine learning*, yaitu Regresi Linier, *Random Forest*, dan *Gradient Boosting*. Dataset yang digunakan berasal dari *Home Value Insights Dataset* di Kaggle yang terdiri atas 1.000 data rumah dengan delapan atribut utama. Tahapan penelitian meliputi pra-pemrosesan data, pembagian data latih dan uji, pelatihan model, optimasi parameter menggunakan *GridSearchCV*, serta evaluasi performa berdasarkan metrik *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, dan *Coefficient of Determination (R^2)* melalui metode *10-Fold Cross Validation*. Hasil pengujian menunjukkan bahwa Regresi Linier memberikan performa terbaik dengan nilai R^2 sebesar 0,8539 serta tingkat kesalahan prediksi yang lebih rendah dibandingkan *Random Forest* dan *Gradient Boosting*. Meskipun model *ensemble* menunjukkan hasil yang kompetitif, peningkatan kompleksitas model tidak menghasilkan peningkatan akurasi yang signifikan, sehingga Regresi Linier dinilai sebagai pendekatan yang paling sederhana, efisien, dan mudah diinterpretasikan untuk sistem prediksi harga rumah pada dataset dengan karakteristik hubungan yang cenderung linear.

Kata Kunci: Prediksi Harga Rumah; Regresi Linier; Random Forest; Gradient Boosting; Machine Learning; Grid Search;

Abstract

House price prediction is a crucial issue in the property sector because it is influenced by various interrelated factors, such as building characteristics and environmental conditions. Accurate prediction using conventional approaches is often difficult and can lead to errors in decision-making. Therefore, this study aims to develop and compare the performance of house price prediction models using three machine learning algorithms: Linear Regression, Random Forest, and Gradient Boosting. The dataset used is the Home Value Insights Dataset on Kaggle, which consists of 1,000 houses with eight main attributes. The research stages include data pre-processing, dividing training and test data, model training, parameter optimization using GridSearchCV, and performance evaluation based on Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) metrics using the 10-Fold Cross Validation method. The test results show that Linear Regression provides the best performance with an R^2 value of 0.8539 and a lower prediction error rate than Random Forest and Gradient Boosting. Although the ensemble model shows competitive results, increasing model complexity does not result in a significant increase in accuracy, so Linear Regression is considered the simplest, most efficient, and most easily interpreted approach for house price prediction systems on datasets with characteristics that tend to be linear.

Keywords: House Price Prediction; Linear Regression; Random Forest; Gradient Boosting; Machine Learning; Grid Search

1. PENDAHULUAN

Pasar properti mempunyai peran besar untuk mendorong pertumbuhan ekonomi dan sering dijadikan indikator tingkat kesejahteraan suatu daerah [1]. Peningkatan jumlah penduduk serta laju urbanisasi yang tinggi di kota besar seperti Jakarta, Bandung, dan Surabaya menyebabkan kebutuhan akan hunian layak dan terjangkau semakin meningkat [2]. Kondisi ini menyebabkan harga rumah mengalami fluktuasi yang signifikan, dipengaruhi oleh berbagai faktor seperti lokasi, luas bangunan, kualitas lingkungan, serta ketersediaan fasilitas umum di sekitarnya [3]. Faktor-faktor tersebut saling berinteraksi secara kompleks dan dinamis, sehingga analisis prediksi harga rumah sebagai tantangan penting pada bidang ekonomi dan teknologi [4].

Nilai suatu rumah ditentukan oleh kombinasi banyak variabel, baik yang bersifat fisik maupun nonfisik. Atribut seperti jumlah kamar tidur, jumlah kamar mandi, luas tanah, dan usia bangunan berperan besar dalam menentukan nilai properti [5]. Selain itu, aspek eksternal seperti kondisi lingkungan dan aksesibilitas turut memengaruhi harga rumah secara signifikan [6]. Kompleksitas hubungan antarvariabel ini membuat metode analisis konvensional sering kali kurang mampu menggambarkan pola hubungan yang sesungguhnya, sehingga dibutuhkan pendekatan yang lebih adaptif dan akurat dalam memprediksi harga properti [7], [8].

Kemajuan teknologi informasi telah mendorong perkembangan metode analisis berbasis machine learning yang mampu memproses data secara otomatis dan mengenali pola tersembunyi di dalamnya [9]. Pendekatan ini memungkinkan komputer untuk belajar dari data historis tanpa memerlukan instruksi eksplisit, sehingga dapat menghasilkan prediksi yang lebih presisi [10]. Berbagai studi sebelumnya menunjukkan bahwa algoritma seperti Regresi Linier, Random Forest, dan Gradient Boosting mampu memberikan tingkat akurasi lebih tinggi diperbandingkan teknik statistik konvensional [11], [12]. Selain itu, ketiga algoritma tersebut juga efektif dalam menangani hubungan non-linier antarvariabel sehingga mampu menghasilkan prediksi yang konsisten dan stabil [13]

Regresi Linier adalah salah satu pendekatan dasar yang banyak dipergunakan dalam memodelkan harga rumah karena mampu menggambarkan hubungan matematis antara variabel bebas dan variabel terikat secara sederhana [14].

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Namun, metode ini memiliki keterbatasan saat diimplementasikan terhadap data yang bersifat kompleks serta tidak linier [15]. Untuk mengatasi hal tersebut, sejumlah penelitian mengombinasikan Regresi Linier dengan algoritma lain seperti Support Vector Regression (SVR) agar dapat menangkap pola hubungan yang lebih variatif [16], [17]. Di sisi lain, metode ensemble learnifng seperti Random Forest menawarkan pendekatan yang lebih stabil karena menggabungkan banyak pohon keputusan untuk menghasilkan model prediksi yang tahan terhadap overfitting [18], [19]

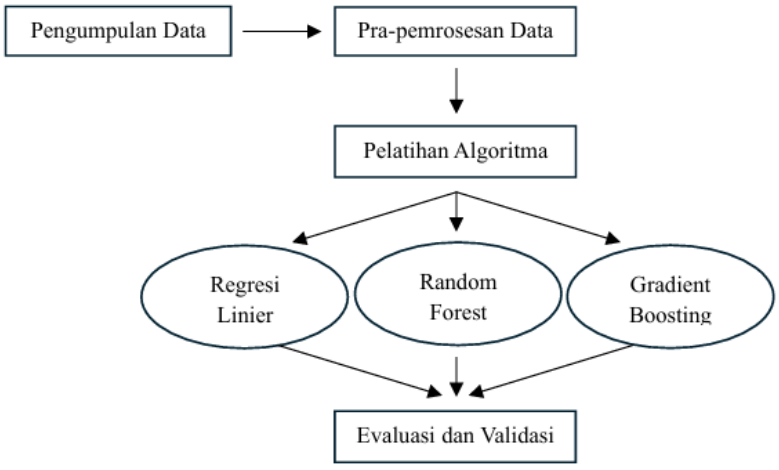
Selain Random Forest, algoritma Gradient Boosting juga banyak diterapkan karena kemampuannya dalam meningkatkan akurasi prediksi melalui proses pembelajaran bertahap [20]. Algoritma ini bekerja dengan memperbaiki kesalahan dari model sebelumnya sehingga hasil prediksi menjadi lebih presisi, terutama pada data dengan distribusi tidak merata [21]. Di samping itu, metode Regresi Linier tetap menjadi pendekatan yang efisien dan mudah diinterpretasikan untuk memprediksi harga rumah, karena mampu menghasilkan estimasi yang stabil meskipun tanpa kombinasi dengan algoritma lain [22].

Ketiga algoritma tersebut telah terbukti mampu menghasilkan prediksi yang baik pada berbagai studi sebelumnya. Namun, hingga saat ini belum terdapat studi yang langsung memperbandingkan langsung membandingkan performa Regresi Linier, Random Forest, dan Gradient Boosting dalam satu kajian yang sama. Perbandingan tersebut penting dilakukan untuk mengetahui algoritma mana yang paling tepat dipergunakan untuk memperkirakan harga rumah sesuai karakteristik data yang tersedia. Berdasarkan alasan tersebut, studi yang dilaksanakan dirancang guna membangun sistem prediksi harga rumah sekaligus menilai kinerja ketiga algoritma tersebut. Dataset yang digunakan berasal dari Home Value Insights Dataset di Kaggle [23], yang berisi 1.000 data rumah dengan delapan atribut utama, yaitu "Square_Footage, Num_Bedrooms, Num_Bathrooms, Year_Built, Lot_Size, Garage_Size, Neighborhood_Quality, dan House_Price sebagai variabel target. Evaluasi dilakukan menggunakan Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan Koefisien Determinasi (R^2) dengan validasi 10-Fold Cross Validation untuk memastikan konsistensi hasil". Melalui pendekatan ini, diharapkan diperoleh model prediksi harga rumah yang akurat, stabil, serta efisien.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian pada studi ini terdiri atas lima proses utama yang digunakan dalam pengembangan model prediksi harga rumah. Proses tersebut meliputi pengumpulan data, pra-pemrosesan data, pelatihan algoritma, evaluasi serta validasi model, dan diakhiri dengan penyimpanan model beserta analisis hasil. Setiap tahapan saling berkaitan dan disusun secara sistematis agar alur penelitian dapat berjalan secara terstruktur dan berurutan. Gambaran menyeluruh mengenai aliran proses penelitian ditampilkan pada Gambar 1 dalam bentuk *alur diagram penelitian*, yang menunjukkan tahapan mulai dari pengolahan data mentah hingga terbentuknya model prediksi yang siap digunakan.



Gambar 1. Alur diagram penelitian

2.2 Pengumpulan Dataset

Dataset yang dipergunakan pada studi yang dilaksanakan diperoleh dari Home Value Insights Dataset yang tersedia pada platform Kaggle [23]. Dataset tersebut berisi 1.000 baris data yang memuat informasi terkait karakteristik fisik rumah dan faktor lingkungan yang memengaruhi nilai properti. Atribut yang tersedia meliputi luas bangunan (Square_Footage), jumlah kamar tidur dan kamar mandi, luas lahan (Lot_Size), tahun pembangunan (Year_Built), ukuran garasi, serta kualitas lingkungan (Neighborhood_Quality). Variabel House_Price digunakan sebagai target dalam proses pemodelan. Dataset ini dipilih karena memiliki kelengkapan fitur yang mendukung penelitian prediksi harga rumah dan telah banyak digunakan dalam studi serupa [1], [2], [3], [4], sehingga dapat dijadikan acuan pembanding. Data yang diunduh dalam

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

format CSV kemudian diolah menggunakan Google Colab dengan bantuan pustaka Python seperti pandas, numpy, matplotlib, joblib, dan scikit-learn sebelum masuk ke tahap pra-pemrosesan.

2.3 Pra-pemrosesan Data

Tahap pra-pemrosesan dilaksanakan guna memastikan bahwa dataset berada dalam kondisi yang bersih, konsisten, serta siap dipergunakan pada proses pelatihan model. Tahapan ini dimulai dengan pemeriksaan adanya data duplikat serta pengecekan kesesuaian format pada setiap variabel numerik. Nilai yang hilang ditangani menggunakan metode imputasi, yaitu median untuk fitur numerik dan modus untuk fitur kategorikal agar distribusi data tetap stabil. Variabel House_Price kemudian dinormalisasi ke dalam satuan juta rupiah untuk menyamakan skala dan mempermudah interpretasi. Selanjutnya, fitur kategorikal Neighborhood_Quality diubah menjadi representasi numerik melalui teknik one-hot encoding, sedangkan seluruh fitur numerik lainnya distandarkan menggunakan StandardScaler agar memiliki rentang nilai yang seragam. Selain itu, dilakukan rekayasa fitur dengan menambahkan atribut Age, yang dihitung berdasarkan selisih antara tahun penelitian dan Year_Built, sehingga menghasilkan informasi tambahan mengenai usia bangunan. Setelah melalui rangkaian proses tersebut, dataset menjadi lebih terstruktur dan layak digunakan pada tahap pemodelan.

2.4 Pelatihan Algoritma

Tahap pelatihan algoritma merupakan inti dari penelitian ini, di mana model dibangun untuk mempelajari pola hubungan antara fitur-fitur pada dataset dan nilai harga rumah. Secara umum, proses pemodelan dalam machine learning menggambarkan hubungan antara variabel input dan output melalui suatu fungsi prediktif yang dapat dituliskan sebagai:

$$\hat{y} = f(x) + \varepsilon \tag{1}$$

Sebelum model dilatih, dataset terlebih dahulu dipisahkan ke dalam dua bagian mempergunakan fungsi train_test_split, yakni 80% sebagai data latih dan 20% menjadi data uji. Pembagian ini mempunyai tujuan supaya model mendapatkan data yang cukup untuk proses pembelajaran sekaligus memiliki data terpisah yang dapat digunakan untuk mengevaluasi kemampuan generalisasi. Untuk memastikan model memiliki performa yang stabil dan tidak bergantung pada satu pembagian data saja, diterapkan metode 10-Fold Cross Validation, yang mana data latih dipisahkan ke dalam sepuluh subset dan proses evaluasi dilakukan secara bergantian. Nilai akhir performa dihitung dari rata-rata skor keseluruhan fold, sesuai dengan rumus:

$$Score = \frac{1}{k}\sum_{i=1}^{k} Score_i\,, k = 10 \tag{2}$$

Tiga algoritma digunakan dalam penelitian ini. Regresi Linier diterapkan sebagai metode dasar untuk memodelkan hubungan linear antara variabel input dan harga rumah menggunakan persamaan:

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \tag{3}$$

Model ini dipilih karena kesederhanaannya serta kemampuannya memberikan interpretasi yang jelas pada setiap fitur. Selanjutnya, algoritma Random Forest digunakan untuk menangkap pola non-linear melalui kombinasi banyak pohon keputusan yang dibentuk dari subset data serta fitur yang dipilih dengan acak. Prediksi akhir diperoleh dari rata-rata hasil seluruh pohon keputusan dalam ensemble:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_i(x) \tag{4}$$

Terakhir, algoritma Gradient Boosting diterapkan untuk mempelajari pola data secara bertahap. Setiap model baru yang ditambahkan berfungsi mengoreksi kesalahan prediksi dari model sebelumnya sehingga akurasi meningkat seiring jumlah iterasi. Proses pembaruan model digambarkan melalui persamaan:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{5}$$

Dengan ketiga algoritma tersebut, penelitian ini memperoleh gambaran performa model baik dari sisi linearitas maupun kompleksitas hubungan antarvariabel, sehingga dapat dibandingkan secara menyeluruh.

2.5 Evaluasi dan Validasi

Evaluasi model dilaksanakan untuk melakukan penilaian seberapa jauh algoritma yang dilatih dapat memperoleh prediksi harga rumah secara akurat dan konsisten pada data yang tidak dipergunakan selama proses pelatihan. Pada tahap ini, performa model diuji mempergunakan tiga metrik utama, yakni "Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), dan Koefisien Determinasi (R^2)". Ketiga metrik ini dipilih sebab memberikan gambaran komprehensif tentang tingkat kesalahan prediksi serta kapabilitas model untuk menguraikan variasi nilai target.

RMSE digunakan untuk melakukan pengukuran rata-rata kesalahan kuadrat antara nilai aktual serta nilai prediksi, yang dikalkulasikan mempergunakan rumus berikut:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

Sementara itu, MAE menggambarkan rata-rata penyimpangan absolut antara nilai aktual serta hasil prediksi tanpa memperhitungkan arah deviasi, sehingga mudah dipahami pada satuan yang sama dengan harga rumah:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (7)$$

Koefisien Determinasi (R^2) dipergunakan dalam meninjau besarnya variasi harga rumah yang bisa diuraikan oleh model, dimana nilai yang mendekati 1 mengindikasikan kemampuan prediksi yang semakin baik. Persamaan R^2 dinyatakan sebagai:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (8)$$

Selain evaluasi numerik, penelitian ini juga menerapkan validasi visual berupa grafik perbandingan antara nilai aktual serta nilai prediksi serta analisis residual guna mendeteksi potensi bias atau pola khusus yang mungkin muncul. Melalui kombinasi evaluasi kuantitatif dan visual ini, diperoleh pemahaman menyeluruh mengenai kualitas model dan kemampuannya untuk diterapkan pada data baru secara andal.

2.6 Penyimpanan Model dan Analisis Hasil

Tahap penyimpanan model dilakukan setelah diperoleh algoritma dengan performa terbaik berdasarkan hasil evaluasi dan validasi. Model disimpan dalam format Pickle (.pkl) menggunakan pustaka joblib agar dapat digunakan kembali tanpa pelatihan ulang, sehingga meningkatkan efisiensi penerapan sistem pada analisis data baru. Selain itu, penelitian ini menghasilkan berbagai output analisis dan visualisasi, seperti hasil prediksi dalam format .csv serta grafik perbandingan nilai aktual dan prediksi dan heatmap korelasi fitur dalam format .png. Visualisasi tersebut membantu memahami hubungan antarvariabel dan kontribusi tiap fitur terhadap harga rumah. Berdasarkan hasil analisis, fitur Square_Footage, Lot_Size, dan Year_Built memiliki pengaruh terbesar terhadap variasi harga, sedangkan Neighborhood_Quality memberi pengaruh tambahan bersifat kontekstual. Proses optimasi parameter dilakukan menggunakan Grid Search pada model berbasis pohon keputusan, sementara analisis koefisien regresi digunakan untuk menilai pengaruh variabel pada model Regresi Linier. pengaruh setiap fitur. Berdasarkan hasil keseluruhan, model Regresi Linier menunjukkan performa paling stabil dan efisien, sehingga ditetapkan sebagai model terbaik. Model ini mampu menjelaskan sebagian besar variasi harga rumah dan memberikan tingkat kesalahan prediksi yang rendah, menjadikannya solusi prediktif yang andal, mudah diimplementasikan, dan relevan untuk digunakan dalam sistem pendukung keputusan di bidang properti.

3. HASIL DAN PEMBAHASAN

Bab ini membahas hasil penerapan tiga algoritma *machine learning*, yaitu Regresi Linier, *Random Forest*, dan *Gradient Boosting*, dalam menyelesaikan permasalahan prediksi harga rumah. Pembahasan tidak hanya menampilkan hasil pengujian, tetapi juga menguraikan tahapan penerapan masing-masing algoritma mulai dari pengolahan data hingga evaluasi performa model. Seluruh eksperimen dilakukan menggunakan Python 3.11 pada platform Google Colab dengan dukungan pustaka *scikit-learn*, *pandas*, *numpy*, *matplotlib*, *joblib*, dan *seaborn*.

3.1 Hasil Eksperimen

3.1.1 Pemeriksaan dan Statistik Awal Dataset

Dataset *Home Value Insights* yang digunakan dalam penelitian ini terdiri dari 1.000 baris data dengan delapan atribut prediktor yang mencerminkan karakteristik fisik bangunan serta kondisi lingkungan sekitarnya. Pada tahap awal, dilakukan pemeriksaan data untuk memastikan kualitas dataset sebelum digunakan dalam proses pemodelan. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung nilai hilang (*missing values*) maupun data duplikat, sehingga seluruh data dapat langsung digunakan tanpa perlu proses pembersihan tambahan. Untuk memberikan gambaran awal mengenai struktur dan jenis informasi yang terdapat dalam dataset, lima baris pertama data ditampilkan pada Tabel 1. Tabel tersebut menyajikan atribut utama yang digunakan dalam penelitian, yaitu luas bangunan (*Square_Footage*), jumlah kamar tidur (*Num_Bedrooms*), jumlah kamar mandi (*Num_Bathrooms*), tahun pembangunan (*Year_Built*), ukuran lahan (*Lot_Size*), ukuran garasi (*Garage_Size*), kualitas lingkungan (*Neighborhood_Quality*), serta harga rumah (*House_Price*). Penyajian cuplikan data ini bertujuan untuk memperjelas format dan karakteristik data yang menjadi dasar dalam penerapan algoritma prediksi harga rumah.

Tabel 1. Cuplikan data awal dari dataset Home Value Insights

Square_ Footage	Num_ Bedrooms	Num_ Bathrooms	Year_ Built	Lot_Size	Garage_ Size	Neighborhood_ Quality	House_ Price
1360	2	1	1981	0.599637	0	5	262382.9
4272	3	3	2016	4.753014	1	6	985260.9
3592	1	2	2016	3.634823	0	9	777977.4

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

966	1	2	1977	2.730667	1	8	22968.9
4926	2	1	1993	4.699073	0	8	1041741.0

3.1.2 Hasil Pra-pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk memastikan data berada dalam kondisi optimal sebelum diterapkan pada algoritma prediksi. Pada tahap ini, seluruh fitur numerik distandarkan untuk menyamakan skala antarvariabel, sementara fitur kategorikal *Neighborhood_Quality* dikonversi menjadi representasi numerik menggunakan teknik *one-hot encoding*. Selain itu, dilakukan rekayasa fitur dengan menambahkan variabel *Age* yang merepresentasikan usia bangunan. Variabel target *House_Price* dinormalisasi ke dalam satuan juta rupiah agar memudahkan interpretasi hasil. Setelah proses ini, dataset menjadi lebih terstruktur dan siap digunakan pada tahap pelatihan algoritma.

3.1.3 Tahapan Penerapan Algoritma Prediksi Harga Rumah

Penerapan algoritma prediksi dilakukan melalui beberapa tahapan utama. Pertama, dataset hasil pra-pemrosesan dibagi menjadi data latih sebesar 80% dan data uji sebesar 20% untuk menguji kemampuan generalisasi model. Selanjutnya, masing-masing algoritma diterapkan secara terpisah pada data latih.
Regresi Linier digunakan sebagai pendekatan dasar untuk memodelkan hubungan linear antara variabel prediktor dan harga rumah. Model ini mempelajari koefisien setiap fitur untuk menghasilkan estimasi harga berdasarkan kombinasi linear variabel input.

Algoritma *Random Forest* diterapkan dengan membangun sejumlah pohon keputusan dari subset data dan fitur yang dipilih secara acak. Setiap pohon menghasilkan prediksi harga, kemudian nilai akhir diperoleh dari rata-rata seluruh prediksi pohon, sehingga model menjadi lebih stabil dan tahan terhadap *overfitting*.

Selanjutnya, algoritma *Gradient Boosting* diterapkan melalui proses pembelajaran bertahap, di mana setiap model baru berfungsi untuk memperbaiki kesalahan prediksi dari model sebelumnya. Pendekatan ini memungkinkan model menangkap pola data yang lebih kompleks secara iteratif. Untuk memastikan hasil yang diperoleh tidak bergantung pada satu pembagian data tertentu, seluruh algoritma dievaluasi menggunakan metode *10-Fold Cross Validation*.

3.1.4 Hasil Evaluasi Model pada Data Uji

Setelah proses pelatihan model selesai dilakukan, tahap selanjutnya adalah mengevaluasi kinerja masing-masing algoritma menggunakan data uji untuk menilai kemampuan prediksi model terhadap data yang belum pernah dilihat sebelumnya. Hasil evaluasi kinerja model pada data uji disajikan secara ringkas pada Tabel 2. Evaluasi dilakukan menggunakan metrik *Root Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, dan koefisien determinasi (R^2) guna memberikan gambaran yang komprehensif mengenai tingkat kesalahan prediksi dan kemampuan model dalam menjelaskan variasi data harga rumah. Berdasarkan hasil yang ditunjukkan pada Tabel 2, ketiga algoritma menunjukkan performa yang kompetitif dengan nilai R^2 di atas 0,84. Regresi Linier menghasilkan nilai *RMSE* dan *MAE* paling rendah serta nilai R^2 tertinggi sebesar 0,8539, yang mengindikasikan bahwa model ini memiliki kemampuan prediksi yang lebih baik dibandingkan dengan algoritma *Random Forest* dan *Gradient Boosting* pada dataset yang digunakan.

Tabel 2. Hasil Evaluasi Model pada Data Uji

Algortima	RMSE (Juta)	MAE (Juta)	R^2
Regresi Linier	0.1024	0.0749	0.8539
Random Forest	0.1039	0.0782	0.8495
Gradient Boosting	0.1035	0.0770	0.8508

3.1.5 Hasil Validasi Silang (10-Fold Cross Validation)

Validasi silang dilakukan untuk menilai kestabilan performa model pada berbagai pembagian data menggunakan metode *10-Fold Cross Validation*. Hasil pengujian validasi silang disajikan pada Tabel 3 yang menunjukkan nilai rata-rata *Root Mean Squared Error (RMSE)* dan *koefisien determinasi (R^2)* dari masing-masing algoritma. Berdasarkan Tabel 3, seluruh model memiliki nilai R^2 rata-rata di atas 0,85, yang mengindikasikan bahwa model mampu menjelaskan variabilitas data harga rumah dengan baik dan memiliki performa yang relatif stabil pada setiap lipatan data. Model regresi linier memperoleh nilai R^2 rata-rata tertinggi sebesar 0,8630 dengan nilai *RMSE* rata-rata terendah, diikuti oleh *Gradient Boosting* dengan nilai R^2 sebesar 0,8618, sementara *Random Forest* menunjukkan performa yang sedikit lebih rendah dibandingkan dua model lainnya. Hasil ini menegaskan bahwa ketiga algoritma memiliki tingkat konsistensi yang baik dan tidak bergantung pada satu subset data tertentu.

Tabel 3. Hasil Validasi Silang (10-Fold CV)

Algoritma	RMSE Rata-rata (CV)	R^2 Rata-rata (CV)
Regresi linier	0.1003	0.8630
Random Forest	0.1033	0.8545
Gradient Boosting	0.1007	0.8618

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

3.1.6 Hasil Optimasi Parameter

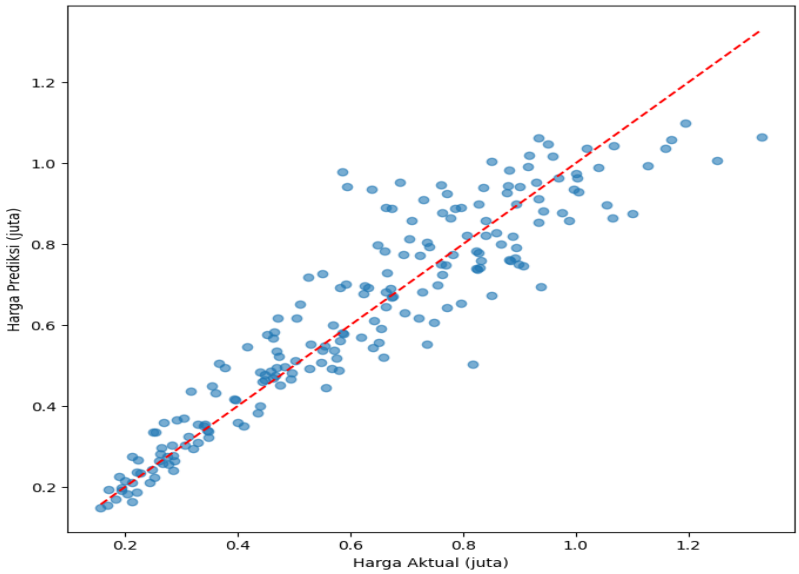
Optimasi parameter dilakukan menggunakan metode *GridSearchCV* pada algoritma Random Forest dan Gradient Boosting untuk memperoleh konfigurasi parameter yang paling optimal. Proses optimasi ini bertujuan untuk meningkatkan performa model dengan menyesuaikan kombinasi parameter utama pada masing-masing algoritma ensemble. Hasil optimasi yang disajikan pada Tabel 4 menunjukkan bahwa kedua algoritma mengalami peningkatan nilai *koefisien determinasi (R^2)* dibandingkan sebelum dilakukan optimasi. Model *Gradient Boosting* menghasilkan nilai R^2 tertinggi sebesar 0,8679, sedangkan *Random Forest* memperoleh nilai R^2 sebesar 0,8577. Meskipun demikian, peningkatan performa yang diperoleh dari proses optimasi parameter ini relatif tidak signifikan apabila dibandingkan dengan model *regresi linier*, yang sebelumnya telah menunjukkan performa yang kompetitif tanpa proses optimasi yang kompleks.

Tabel 4. Parameter Optimal Model Ensemble

Algoritma	Parameter Optimal	R^2 (CV)
Random Forest	n estimators=100, max depth=10, min samples split=5	0.8577
Gradient Boosting	n estimators=100, learning rate=0.05, max depth=3	0.8679

3.1.7 Feature Importance dan Visualisasi Prediksi

Analisis kontribusi fitur pada model ensemble menunjukkan bahwa *Square_Footage* merupakan variabel yang paling dominan dalam memengaruhi harga rumah, kemudian diikuti oleh *Lot_Size* dan *Year_Built*. Temuan ini mengindikasikan bahwa luas bangunan dan ukuran lahan memiliki peran yang lebih signifikan dibandingkan variabel lainnya dalam menentukan nilai properti. Visualisasi perbandingan antara nilai aktual dan hasil prediksi pada model Regresi Linier ditunjukkan pada Gambar 2. Sebagian besar titik data terlihat berada di sekitar garis diagonal, yang menandakan bahwa model mampu menghasilkan prediksi yang mendekati nilai sebenarnya serta memiliki tingkat akurasi yang baik.



Gambar 2. Grafik Perbandingan Nilai Aktual dan Prediksi (Linear Regression)

3.2 *Analisis Dan Pembahasan*

3.2.1 Perbandingan Kinerja Antar Model

Regresi Linier menunjukkan performa terbaik pada data uji dengan nilai R^2 sebesar 0.8539. Hal ini mengindikasikan bahwa hubungan antarvariabel pada dataset relatif linear, sehingga model linear mampu menangkap pola dengan baik. Random Forest dan Gradient Boosting memberikan performa yang kompetitif, tetapi peningkatan kompleksitas keduanya tidak menghasilkan kenaikan akurasi yang signifikan.

3.2.2 Analisis Pola Hubungan Variabel

Fitur Square_Footage memiliki korelasi tertinggi terhadap harga rumah (0.9238), sehingga tidak mengherankan jika fitur ini muncul sebagai faktor paling dominan. Lot_Size dan Year_Built juga memberikan kontribusi positif, menandakan bahwa ukuran lahan dan usia bangunan berpengaruh nyata terhadap nilai properti. Sebaliknya, fitur Num_Bathrooms dan Age memiliki kontribusi rendah dan cenderung tidak berpengaruh signifikan.

3.2.3 Evaluasi Kesesuaian Model

Hasil validasi silang menunjukkan bahwa seluruh model memperoleh nilai R^2 rata-rata di atas 0,85, yang menandakan bahwa performanya tetap konsisten pada berbagai skenario pembagian data. Temuan ini juga mengindikasikan jika model

JURIKOM (Jurnal Riset Komputer), Vol. 12 No. 6, Desember 2025
e-ISSN 2715-7393 (Media Online), p-ISSN 2407-389X (Media Cetak)
DOI 10.30865/jurikom.v12i6.9369
Hal 1018-1025
https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom

dapat melaksanakan generalisasi secara baik ketika diterapkan pada data baru yang tidak disertakan dalam proses pelatihan.

3.2.4 Perbandingan dengan Penelitian Sebelumnya

Temuan penelitian ini konsisten dengan sejumlah studi sebelumnya [1], [2], dan [4] yang menegaskan bahwa luas bangunan merupakan variabel paling berpengaruh untuk memperkirakan harga rumah. Di samping itu, hasil bahwa model linear dapat mengungguli model ensemble pada dataset ini juga sejalan dengan penelitian lain yang menjelaskan bahwa hubungan harga properti dapat berbentuk linear apabila dataset tidak memuat informasi geografis.

3.2.5 Keterbatasan Penelitian

Studi yang dilaksanakan mempunyai sejumlah keterbatasan yang penting untuk diberikan perhatian. Pertama, jumlah dataset yang digunakan hanya sebanyak 1.000 baris, sehingga masih tergolong relatif kecil untuk mengeksplorasi performa model ensemble yang biasanya membutuhkan data dalam jumlah besar agar pola non-linear dapat terdeteksi secara optimal. Kedua, dataset tidak memuat fitur lokasi seperti koordinat geografis, kode pos, maupun jarak ke pusat kota, padahal variabel spasial umumnya memiliki pengaruh signifikan terhadap variasi harga properti. Ketiadaan fitur tersebut menyebabkan model hanya mampu menangkap pengaruh faktor fisik bangunan tanpa mempertimbangkan aspek lingkungan secara menyeluruh. Selain itu, penelitian ini belum mencakup penggunaan algoritma lain seperti "Support Vector Regression (SVR), XGBoost, atau Artificial Neural Network (ANN)" yang dapat memberikan sudut pandang lebih luas mengenai perbandingan performa model pada data dengan karakteristik serupa. Terakhir, evaluasi model hanya dilakukan pada satu dataset tanpa uji eksternal, sehingga generalisasi model terhadap data dari wilayah atau kondisi berbeda belum dapat dipastikan sepenuhnya.

3.2.6 Implikasi Penelitian

Penelitian ini dapat dijadikan acuan dalam pengembangan sistem prediksi harga properti yang ringan, cepat, dan mudah diimplementasikan. Regresi Linier terbukti cukup akurat sehingga dapat dimanfaatkan dalam aplikasi berbasis web atau sistem pendukung keputusan dengan kebutuhan komputasi minimal.

3.3 Model Terbaik dan Alasan Pemilihannya

Berdasarkan hasil evaluasi pada data uji dan validasi silang, Regresi Linier ditetapkan sebagai model terbaik dalam penelitian ini. Model ini menunjukkan kinerja paling unggul dengan nilai R^2 tertinggi pada data uji sebesar 0.8539 serta performa yang stabil pada validasi silang 10-fold dengan R^2 rata-rata 0.8630. Selain menghasilkan akurasi yang kompetitif, Regresi Linier juga memiliki struktur yang lebih sederhana, efisien dari sisi komputasi, dan mudah diinterpretasikan dibandingkan model ensemble seperti Random Forest dan Gradient Boosting. Keunggulan ini membuat Regresi Linier sangat sesuai dengan karakteristik dataset yang cenderung linear dan tidak terlalu kompleks. Dengan demikian, model ini dinilai paling layak digunakan dalam skenario prediksi harga rumah dengan tingkat kompleksitas rendah hingga menengah, serta cocok diterapkan pada sistem prediksi berbasis data yang membutuhkan kecepatan pemrosesan dan transparansi hasil.

4. KESIMPULAN

Berdasarkan hasil eksperimen menggunakan *Home Value Insights Dataset*, ketiga algoritma yang diuji—Regresi Linier, *Random Forest*, dan *Gradient Boosting*—menunjukkan performa yang baik dengan nilai R^2 di atas 0,85. Di antara ketiganya, Regresi Linier memberikan hasil paling optimal dan stabil dengan nilai R^2 sebesar 0,8539, *RMSE* sebesar 0,1024 juta, dan R^2 (*CV*) sebesar 0,8630. Hasil tersebut mengindikasikan jika model sederhana ini dapat menguraikan sebagian besar variasi harga rumah secara akurat dan efisien. Fitur *Square_Footage* dan *Lot_Size* diketahui memiliki pengaruh paling signifikan terhadap harga rumah, sedangkan jumlah kamar mandi dan umur bangunan hanya memberikan kontribusi kecil terhadap nilai properti. Optimasi parameter menggunakan *GridSearchCV* pada model *Random Forest* dan *Gradient Boosting* menghasilkan peningkatan akurasi yang tidak terlalu signifikan, menandakan bahwa peningkatan kompleksitas model tidak selalu sebanding dengan peningkatan performa, khususnya pada data dengan hubungan yang bersifat linier. Visualisasi perbandingan antara nilai aktual serta prediksi mengindikassikan sebaran titik data yang berdekatan dengan garis diagonal, menegaskan kemampuan generalisasi yang baik dari model Regresi Linier. Untuk penelitian mendatang, disarankan agar dataset diperluas dengan menambahkan variabel seperti lokasi, jarak ke pusat kota, aksesibilitas, dan faktor ekonomi wilayah agar hasil yang diperoleh lebih representatif, serta menguji algoritma lain seperti *Support Vector Regression (SVR)*, *Extreme Gradient Boosting (XGBoost)*, dan *Artificial Neural Network (ANN)* untuk memperoleh perbandingan yang lebih komprehensif terhadap tingkat akurasi dan efisiensi model. lanjutan seperti *stacking* atau *hybrid model* juga dapat menjadi alternatif untuk meningkatkan performa prediksi. Model hasil pelatihan sebaiknya disimpan dalam format yang dapat digunakan ulang serta diuji di lingkungan produksi agar hasilnya dapat diterapkan langsung dalam sistem prediksi harga rumah berbasis digital.

REFERENCES

[1] K. D. Sanjaya, "Prediksi Harga Rumah Dengan Metode Regresi Linear Dan Support Vector Regression Di Daerah Tebat Jakarta

Selatan," *J. Komput. dan Inform.*, vol. 19, no. 2, pp. 95–102, 2024.
[2] R. Tanamal, N. Minoque, T. Wiradinata, Y. Soekamto, and T. Ratih, "House Price Prediction Model Using Random Forest in Surabaya City," *TEM J.*, vol. 12, no. 1, pp. 126–132, 2023, doi: 10.18421/TEM121-17.
[3] U. M. Semarang, M. Of, F. Influencing, and G. Development, "Jurnal statistika," vol. 12, no. 2, pp. 29–41, 2024, doi: 10.14710/JSUNIMUS.12.2.2024.29-41.
[4] M. B. S. Qolbi, T. N. Puteh, R. Rivandi, and C. Rozikin, "Prediksi Harga Rumah Di Jakarta Pusat Menggunakan Algoritma Machine Learning," *J. Ilmu Komput. dan Bisnis*, vol. 16, no. 1, pp. 16–24, 2025, doi: 10.47927/jikb.v16i1.840.
[5] R. Annamoradnejad and I. Annamoradnejad, "Machine Learning for Housing Price Prediction," *Encycl. Data Sci. Mach. Learn.*, no. September 2022, pp. 2728–2739, 2022, doi: 10.4018/978-1-7998-9220-5.ch163.
[6] Dhiwa Aqsha, "Perbandingan Kinerja Algoritma Extreme Gradient Boosting Dan Random Forest Untuk Prediksi Harga Rumah Di Jabodetabek," *J. Ilmu Komput. dan Sist. Inf.*, vol. 13, no. 1, pp. 1–7, 2025, doi: 10.24912/jiksi.v13i1.32863.
[7] S. Suakanto, A. Christy, V. J. L. Engel, and D. Angela, "Pengembangan Sistem Prediksi Harga Pasar Properti Menggunakan Big Data Platform," *J. Telemat.*, vol. 13, no. 1, pp. 19–26, 2019, doi: 10.61769/telematika.v13i1.257.
[8] A. Fuadah, A. M. Siregar, and Y. Cahyana, "Model Prediksi Harga Rumah Di Kabupaten Bandung Menggunakan Multiple Linear Regression Dan Support Vector Regression," *Sci. Student J. Information, Technol. Sci.*, vol. 5, no. 2, pp. 10–16, 2024.
[9] R. R. Hallan and I. N. Fajri, "Prediksi Harga Rumah menggunakan Machine Learning Algoritma Regresi Linier," *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 7, no. 1, pp. 57–62, 2025, doi: 10.47233/jteksis.v7i1.1732.
[10] A. Jl *et al.*, "Analisis Prediksi Harga Rumah di Bandung Menggunakan Regresi Linear Berganda Rafif Nauval Tuah Siregar Vijay Sitorus Universitas Negeri Medan Willy Pramudia Ananta perbandingan melalui penalaran berbasis kasus ," yang dilakukan oleh I-Cheng Yeh , Tzu- yan," vol. 1, no. 6, 2023.
[11] A. Widyastuti, "Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Metode Multiple Linear Regression," *SUBMIT J. Ilm. Teknol. Infomasi dan Sains*, vol. 4, no. 1, pp. 30–35, 2024, doi: 10.36815/submit.v4i1.3343.
[12] H. Hakim, D. Kamil, and B. Alatas, "Pendekatan Machine Learning untuk Estimasi Harga Rumah dengan Regresi Linier," *ALPHA J. Sci. Technol.*, vol. 1, no. 1, pp. 18–22, 2025, doi: 10.70716/alpha.v1i1.99.
[13] R. Fauzan Almahdy and W. D. Mega Pradnya, "Prediksi Harga Rumah Di Kabupaten Bantul Menggunakan Algoritma Support Vector Regression," *J. Tek. Inform. dan Sist. Inf.*, vol. 11, no. 2, pp. 152–165, 2024.
[14] R. Hidayat *et al.*, "Implementasi Algoritma Random Forest Regression Untuk Memprediksi Penjualan Produksi di Supermarket," *Simkom*, vol. 10, no. 1, pp. 101–109, 2025, doi: 10.51717/simkom.v10i1.703.
[15] A. P. Wardani, H. A. Irawan, M. P. Syah, M. A. Akmal, N. U. Nariswari, and K. M. Hindrayani, "Analisis dan Prediksi Harga Properti Rumah di Kota Surabaya dengan Algoritma Random Forest," *Pros. Semin. Nas. Sains Data*, vol. 4, no. 1, pp. 885–894, 2024, doi: 10.33005/senada.v4i1.375.
[16] Mohit Jain and Arjun Srihari, "House price prediction with Convolutional Neural Network (CNN)," *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 405–415, 2023, doi: 10.30574/wjaets.2023.8.1.0048.
[17] F. A. Rangkuti, Khairunnisa, and S. Sundari, "Implementasi Gradient Boosting Machines Untuk Prediksi Harga Rumah Pada Jakarta Selatan," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 2, pp. 164–172, 2025, doi: 10.69916/jkbti.v4i2.318.
[18] G. Khalda Rifdan, N. Rahaningsih, A. Bahtiar, I. Ali, and N. Dienwati Nuris, "Ramalan Penjualan Rumah Menggunakan Algoritma Linear Regresi Di Tebet Jakarta Selatan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1847–1851, 2024, doi: 10.36040/jati.v8i2.9022.
[19] A. Fauzi, N. Maulidah, R. Supriyadi, H. Nalatissifa, and S. Diantika, "Prediksi Harga Properti Di Indonesia Menggunakan Algoritma Random Forest," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 1, pp. 43–49, 2025, doi: 10.31004/riggs.v4i1.367.
[20] M. N. Hibatulloh, G. D. Prakoso, A. D. Putri Yunus, and T. D. Putra, "Prediksi Harga Rumah di Bandung 2024 Menggunakan Ensemble Learning: Analisis Komparatif dan Interpretabilitas," *J. Inform. J. Pengemb. IT*, vol. 10, no. 2, pp. 484–493, 2025, doi: 10.30591/jpit.v10i2.8200.
[21] R. Riyandi, R. Roy Hakiki, Dealmus, Y. O. Reins Dima, and N. P, "Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Metode Regresi Linear Berganda Berbasis Shiny R," *Inform. J. Ilmu Komput.*, vol. 21, no. 1, pp. 1–13, 2025, doi: 10.52958/iftk.v21i1.10563.
[22] S. Anggellica, B. N. Sari, and I. Maulana, "Prediksi Harga Rumah Menggunakan Multiple Linear Regression (Studi Kasus : Kabupaten Karawang Pada Website Lamudi.Co.Id)," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 2, 2025, doi: 10.23960/jitet.v13i2.6370.
[23] Prokshitha, "Home Value Insights Dataset," 2024, *Kaggle Datasets, San Francisco (opsional, markas Kaggle)*. [Online]. Available: https://www.kaggle.com/datasets/prokshitha/home-value-insights