



Analisis Perbandingan Prediksi Harga Rumah Dengan Random Forest, Gradient Boosting, dan XGBoost

Bety Wulan Sari^{1,*}, Donni Prabowo²

^{1,2} Universitas AMIKOM Yogyakarta, Yogyakarta, Indonesia

Informasi Artikel

Sejarah Artikel:
 Submit : 24 Juni 2025
 Revisi : 26 Juni 2025
 Diterima : 29 Juni 2025
 Diterbitkan: 30 Juni 2025

Kata Kunci

harga rumah, regresi, XGBoost, Random Forest, Gradient Boosting, Yogyakarta

Correspondence

E-mail: bety@amikom.ac.id*

A B S T R A K

Prediksi harga rumah menjadi tantangan penting dalam bidang properti, khususnya di wilayah Yogyakarta yang memiliki variasi harga cukup ekstrem. Penelitian ini bertujuan untuk membandingkan performa tiga algoritma regresi yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost* digunakan untuk membangun model prediksi harga rumah berdasarkan fitur seperti luas tanah, luas bangunan, jumlah kamar tidur, kamar mandi, dan garasi. Data yang dianalisis mencakup 1.642 entri dengan harga rumah berkisar antara Rp 7 juta hingga Rp 4,37 miliar, harga rata-rata sebesar Rp 1,14 miliar, dan modus Rp 775 juta. Evaluasi model dilakukan menggunakan metrik *Mean Squared Error* (MSE) dan koefisien determinasi (R^2), di mana *XGBoost* menghasilkan performa terbaik dengan MSE sebesar $1,56 \times 10^{14}$ rupiah², R^2 sebesar 0,7746, dan *Root Mean Squared Error* (RMSE) sekitar 12,5 juta rupiah. Hasil ini menunjukkan bahwa *XGBoost* lebih unggul dalam menangani data tabular kompleks dan memiliki akurasi prediksi yang lebih baik dibanding dua model lainnya. Model prediktif ini berpotensi digunakan oleh pengembang properti, *agen real estate*, maupun pemerintah daerah sebagai alat bantu dalam penetapan harga, evaluasi pasar, dan perencanaan tata ruang yang berbasis data. Temuan ini memberikan gambaran bahwa pemilihan algoritma yang tepat dapat meningkatkan kualitas prediksi harga properti.

Abstract

House price prediction poses a significant challenge in the property sector, especially in the Yogyakarta region, which exhibits a wide range of price variations. This study aims to compare the performance of three regression algorithms such as *Random Forest*, *Gradient Boosting*, and *XGBoost*, in building predictive models based on features such as land area, building area, number of bedrooms, bathrooms, and garage availability. The dataset analyzed consists of 1,642 entries, with house prices ranging from IDR 7 million to IDR 4.37 billion, an average price of IDR 1.14 billion, and a mode of IDR 775 million. Model evaluation was conducted using *Mean Squared Error* (MSE) and the coefficient of determination (R^2), where *XGBoost* achieved the best performance with an MSE of 1.56×10^{14} IDR², an R^2 of 0.7746, and a *Root Mean Squared Error* (RMSE) of approximately IDR 12.5 million. These results indicate that *XGBoost* outperforms the other two models in handling complex tabular data and provides more accurate predictions. The predictive model has practical potential to be utilized by property developers, real estate agents, and local governments as a decision-support tool for price estimation, market evaluation, and data-driven urban planning. These findings highlight that selecting the appropriate algorithm can significantly enhance the quality of house price prediction.

This is an open access article under the CC-BY-SA license



1. Pendahuluan

Dalam beberapa tahun terakhir, pertumbuhan sektor properti di Yogyakarta menunjukkan *tren* yang signifikan. Kota yang dikenal sebagai pusat pendidikan dan budaya ini semakin berkembang menjadi destinasi investasi, wisata, dan hunian, sehingga permintaan terhadap lahan dan rumah

tinggal meningkat secara drastis. Kenaikan permintaan tersebut berdampak langsung terhadap harga rumah di berbagai wilayah, bahkan menjadikan Yogyakarta sebagai salah satu kota dengan pertumbuhan harga properti tertinggi di Indonesia. Hal ini akan membuat harga rumah semakin hari semakin naik dengan daya beli masyarakat yang melonjak tinggi [1]. Fenomena ini tidak hanya dialami di kawasan pusat kota seperti Gondokusuman dan Umbulharjo, tetapi juga merambah ke wilayah pinggiran seperti Sleman, Bantul, hingga Gunungkidul.

Tingginya variasi harga rumah disebabkan oleh beragam faktor, mulai dari lokasi, akses jalan, luas tanah dan bangunan, hingga kedekatan dengan fasilitas umum seperti kampus, pusat perbelanjaan, dan tempat ibadah [2], [3]. Bagi calon pembeli, *developer*, maupun investor, fluktuasi harga ini menimbulkan tantangan tersendiri dalam membuat keputusan yang tepat dan tepat waktu. Salah satu pendekatan yang umum digunakan dalam estimasi harga properti adalah metode *regresi*, yakni suatu teknik statistika dan *machine learning* yang digunakan untuk memodelkan hubungan antara variabel *independen* (fitur properti) dengan variabel *dependen* (harga rumah). Regresi memiliki keunggulan karena mampu memberikan estimasi numerik, berbeda dengan klasifikasi yang hanya memberikan label [4], [5], [6].

Dalam beberapa tahun terakhir, perkembangan teknologi *machine learning* membuka peluang baru dalam implementasi algoritma regresi berbasis data [7], [8]. Berbeda dengan metode konvensional seperti regresi linear yang mengasumsikan hubungan linier antar variabel, algoritma *machine learning* mampu menangkap pola kompleks dan interaksi *non-linier* yang tersembunyi dalam data. Kemampuan ini menjadi sangat penting ketika menghadapi data harga rumah yang tidak hanya bersifat *non-linier*, tetapi juga dipengaruhi oleh banyak variabel saling terkait seperti lokasi, luas tanah, dan fasilitas. Oleh karena itu, kebutuhan akan model *machine learning* menjadi semakin mendesak untuk menghasilkan prediksi yang lebih akurat, adaptif, dan relevan dalam konteks dunia nyata [9], [10], [11]. Tidak lagi terbatas pada *regresi linear klasik*, kini berbagai algoritma *modern* seperti *Random Forest*, *Gradient Boosting*, dan *XGBoost* banyak digunakan dalam berbagai studi prediksi harga rumah di dunia. Ketiga algoritma ini dikenal sebagai model yang menggabungkan kekuatan banyak model sederhana untuk membentuk prediksi yang lebih akurat dan stabil. Keunggulan dari model-model ini antara lain adalah kemampuannya dalam menangani data *non-linear*, fitur kategorikal, serta mengurangi risiko *overfitting* [12], [13], [14].

Penelitian sebelumnya membandingkan tiga metode *regresi* yaitu *Linear Regression*, *Random Forest Regression*, dan *Gradient Boosted Trees Regression*, untuk memprediksi harga rumah di wilayah Jakarta, menggunakan data dari *Kaggle* yang terdiri dari 2.011 *record* dengan enam fitur utama, seperti harga, luas tanah, dan jumlah kamar. Proses analisis dilakukan dengan pendekatan CRISP-DM dan evaluasi menggunakan 10-fold *cross-validation* melalui platform *KNIME*. Hasil menunjukkan bahwa metode *Random Forest Regression* memberikan performa terbaik dengan nilai RMSE terendah sebesar 0,440 dan akurasi tertinggi sebesar 81,5%, mengungguli *Gradient Boosted Trees* (RMSE 0,508) dan *Linear Regression* (RMSE 0,515). Dengan demikian, *Random Forest* terbukti lebih efektif dalam menghasilkan prediksi harga rumah yang akurat pada *dataset* ini [15].

Penelitian berikutnya menggunakan metode *Gradient Boosting Machine* (GBM) untuk memprediksi harga rumah di Jakarta Selatan berdasarkan fitur seperti luas tanah, luas bangunan, jumlah kamar tidur, kamar mandi, dan garasi. Hasil analisis menunjukkan bahwa luas tanah dan luas bangunan merupakan faktor paling dominan dalam memengaruhi harga. Model dievaluasi menggunakan metrik MSE, MAE, dan R^2 , yang menghasilkan nilai MSE sebesar $1,6e+19$, MAE sekitar 30 miliar rupiah, dan R^2 sebesar 0,657, menandakan model memiliki akurasi sedang namun masih menghasilkan kesalahan prediksi yang signifikan, terutama pada rumah dengan harga sangat tinggi. Visualisasi residual dan distribusi error menunjukkan adanya *outlier* serta prediksi yang meleset jauh pada beberapa kasus, sehingga model masih memerlukan optimasi lanjutan agar lebih presisi dan mampu menangkap pola data yang kompleks secara lebih akurat. Namun, studi ini hanya

menggunakan satu algoritma dan belum membandingkan beberapa model *regresi* secara sistematis pada data lokal. [16].

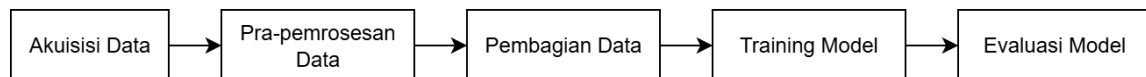
Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja tiga algoritma *regresi* yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost*, dalam memprediksi harga rumah menggunakan data tabular di wilayah Yogyakarta [16], [17], [18]. Pendekatan ini dilakukan dengan serangkaian tahapan preprocessing data, seperti transformasi pada variabel target (harga rumah), *encoding* fitur kategorikal menggunakan *Ordinal Encoder* [19], serta normalisasi fitur numerik menggunakan *StandardScaler* [20]. Setiap model kemudian dilatih dan dievaluasi menggunakan metrik *Mean Squared Error* (MSE) dan koefisien determinasi (R^2 Score) untuk menilai tingkat akurasi prediksi [21], [22].

Kontribusi ilmiah dari penelitian ini mencakup penggunaan *dataset* lokal dari wilayah Yogyakarta yang jarang digunakan dalam studi serupa, dan analisis komparatif terhadap tiga model yang populer dalam regresi untuk mengidentifikasi algoritma paling optimal dalam konteks prediksi harga rumah di Indonesia.

Dengan fokus pada data rumah-rumah yang berlokasi di Provinsi Daerah Istimewa Yogyakarta, penelitian ini tidak hanya memberikan pendekatan teknis dalam membangun sistem prediktif, namun juga memberikan gambaran realistis tentang faktor-faktor yang berkontribusi terhadap harga rumah di wilayah tersebut. Data yang digunakan dalam penelitian ini diperoleh dari *platform publik Kaggle*, yang merupakan *dataset* terbuka berisi informasi harga rumah, lokasi (kabupaten dan kecamatan), serta fitur properti lainnya seperti luas tanah, luas bangunan, jumlah kamar tidur, kamar mandi, dan garasi [23]. *Dataset* ini merepresentasikan kondisi pasar properti di wilayah Yogyakarta dan sekitarnya. Diharapkan, hasil dari penelitian ini dapat memberikan manfaat bagi pihak-pihak yang berkepentingan, seperti pembeli rumah, agen properti, pengembang perumahan, hingga perencanaan tata kota.

2. Metodologi Penelitian

Untuk memberikan gambaran umum terhadap keseluruhan proses, penelitian ini digambarkan dalam bentuk diagram alur metodologi yang mencakup proses *preprocessing* data, pembuatan *pipeline* model, pelatihan algoritma *regresi*, hingga evaluasi performa model yang disajikan dalam gambar 1.



Gambar 1. Alur Penelitian

2.1. Akuisisi Data

Penelitian ini menggunakan data harga rumah yang terdiri dari berbagai fitur properti seperti lokasi (kabupaten dan kecamatan), ukuran bangunan, luas tanah, dan atribut lainnya yang berpengaruh terhadap nilai properti. Data bersifat tabular dan berasal dari *Kaggle* yang mencerminkan kondisi harga tanah di wilayah Yogyakarta dan sekitarnya [23]. Gambar 2 menyajikan fitur pada *dataset* harga tanah di Yogyakarta.

```

RangeIndex: 2020 entries, 0 to 2019
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   price                  2020 non-null  int64  
1   listing-location       2020 non-null  object  
2   bed                    2020 non-null  int64  
3   bath                   2020 non-null  int64  
4   carport                 2020 non-null  int64  
5   surface_area           2020 non-null  int64  
6   building_area          2020 non-null  int64  
dtypes: int64(6), object(1)
memory usage: 110.6+ KB
  
```

Gambar 2. Fitur *dataset* harga tanah di Yogyakarta

Target variabel dalam penelitian ini adalah *price* (harga rumah), sedangkan fitur-fitur *independen* mencakup gabungan antara data *kategorikal* (seperti kabupaten, kecamatan) dan *numerik* (misalnya luas_bangunan, jumlah_kamar, luas_tanah). Gambar 3 menyajikan sample data.

	price	listing-location	bed	bath	carport	surface_area	building_area
0	Rp 1,79 Miliar	Ngaglik, Sleman	3.0	3.0	2.0	120 m ²	110 m ²
1	Rp 170 Juta	Jombor, Sleman	3.0	2.0	1.0	102 m ²	126 m ²
2	Rp 695 Juta	Berbah, Sleman	2.0	2.0	1.0	100 m ²	100 m ²
3	Rp 560 Juta	Prambanan, Sleman	3.0	1.0	1.0	109 m ²	67 m ²
4	Rp 200 Juta	Moyudan, Sleman	2.0	1.0	1.0	60 m ²	30 m ²

Gambar 3. Sample data

Sebelum dilakukan pelatihan model, target variabel ditransformasikan untuk mengurangi *skewness* pada distribusi harga, sehingga dapat meningkatkan stabilitas dan akurasi model *regresi*.

2.2. Pra-pemrosesan Data

Data preprocessing dilakukan secara sistematis untuk menyiapkan data mentah menjadi data yang layak untuk dimasukkan ke dalam model machine learning. Tahapan preprocessing meliputi:

1. Penghapusan Fitur *Redundan*: Kolom *listing-location* dihapus karena *redundan* atau tidak informatif.
2. Transformasi *Logaritmik*: Harga (*price*) ditransformasikan dengan fungsi $\log_{10}()$ untuk membuat distribusinya lebih normal.
3. *Encoding* Fitur *Kategorikal*: Dua kolom *kategorikal* (kabupaten, kecamatan) diolah menggunakan *Ordinal Encoding* agar dapat digunakan dalam algoritma berbasis pohon.
4. Standarisasi Fitur *Numerik*: Fitur *numerik* dinormalisasi menggunakan *StandardScaler* untuk menyamakan skala antar fitur.

Preprocessing ini dibungkus ke dalam sebuah *ColumnTransformer* yang kemudian dimasukkan ke dalam *pipeline* model agar semua tahapan *preprocessing* dilakukan secara otomatis saat pelatihan dan prediksi.

2.3. Pembagian Data

Data dibagi menjadi dua subset menggunakan fungsi *train_test_split* dari *scikit-learn* dengan rasio 80:20, di mana:

1. 80% data digunakan sebagai data latih (X_{train}, y_{train_log})
2. 20% data digunakan sebagai data uji (X_{test}, y_{test_log}).

Pendekatan pembagian sederhana ini dipilih untuk menjaga efisiensi proses pelatihan dan evaluasi model, mengingat ukuran *dataset* yang relatif sedang. Namun, untuk meningkatkan keandalan evaluasi dan mengurangi potensi bias dari pembagian data acak, validasi silang seperti *K-Fold* disarankan pada penelitian selanjutnya.

2.4. Training Model

Penelitian ini menggunakan tiga algoritma regresi, yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost*. Ketiga algoritma tersebut dipilih karena kemampuannya yang terbukti efektif dalam menangani data *tabular* [24], [25], [26], mengelola interaksi *non-linear* antar fitur [27], [28], [29], dan menghasilkan prediksi yang stabil [30], [31]. *Random Forest Regressor* adalah metode berbasis bagging, yang membangun banyak pohon keputusan secara paralel dan menggabungkan hasil prediksinya melalui rata-rata. Algoritma ini relatif tahan terhadap *overfitting* dan mampu menangani variabel kategorikal maupun numerik.

Gradient Boosting Regressor bekerja secara bertahap dengan membentuk pohon keputusan baru untuk memperbaiki kesalahan model sebelumnya. Teknik *boosting* ini cenderung memberikan hasil prediksi yang lebih presisi dibanding metode *bagging*, tetapi lebih sensitif terhadap *overfitting*.

XGBoost (*Extreme Gradient Boosting*) merupakan varian lanjutan dari *gradient boosting* yang dilengkapi dengan teknik regularisasi, pengaturan *subsampling*, dan optimasi komputasi yang efisien. XGBoost telah menjadi salah satu algoritma paling populer dalam kompetisi *machine learning* karena kinerjanya yang unggul dalam berbagai kasus prediksi.

2.4. Evaluasi Model

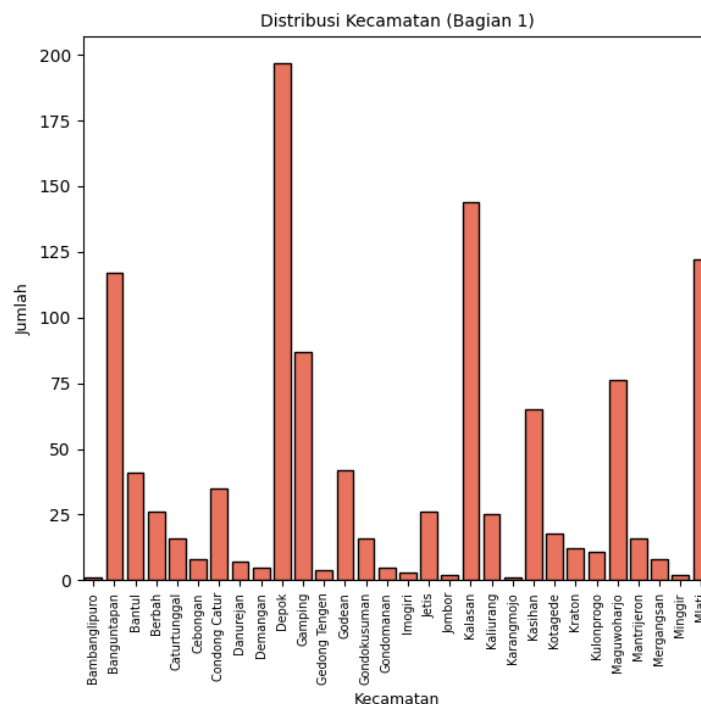
Untuk mengukur performa masing-masing model, digunakan dua metrik evaluasi:

1. *Mean Squared Error* (MSE): Mengukur selisih kuadrat antara nilai prediksi dan aktual. Semakin kecil nilai MSE, semakin baik performa model.
2. R^2 Score (Koefisien Determinasi): Menunjukkan seberapa baik variasi harga rumah dapat dijelaskan oleh model. Nilai R^2 mendekati 1 menunjukkan prediksi yang baik.

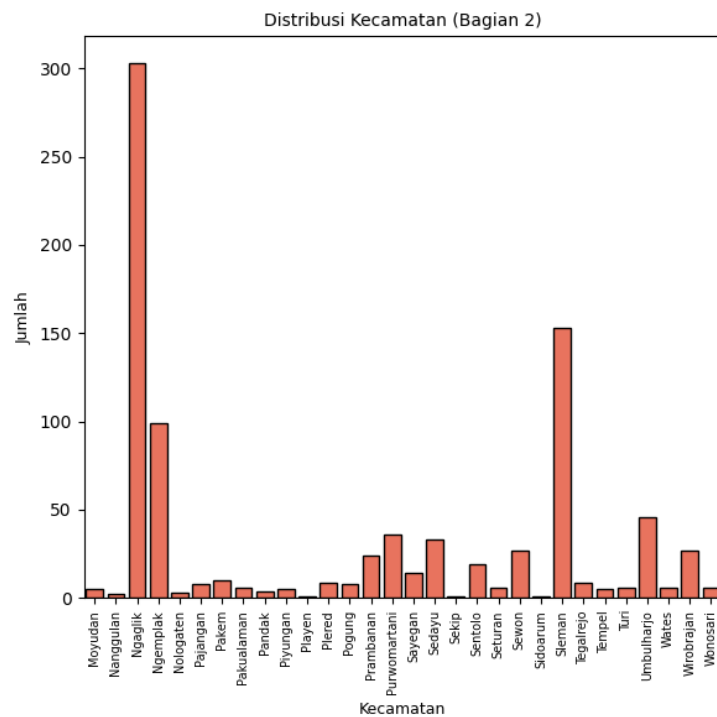
3. Hasil dan Pembahasan

Berdasarkan data yang dianalisis, harga rumah di Yogyakarta bervariasi mulai dari yang terendah sebesar Rp 7 juta hingga yang tertinggi mencapai Rp 4,37 miliar, dengan harga rata-rata sekitar Rp 1,14 miliar dan harga yang paling sering muncul (*modus*) sebesar Rp 775 juta. Kecamatan dengan jumlah rumah terbanyak adalah Ngaglik (226 rumah), diikuti oleh Depok (147), Sleman (134), Kalasan (134), dan Banguntapan (106). Sementara itu, distribusi rumah berdasarkan kabupaten/kota menunjukkan dominasi wilayah Sleman dengan 1.065 rumah, disusul Bantul (292), Yogyakarta (243), Kulon Progo (35), dan Gunung Kidul (7).

Untuk memahami karakteristik awal data yang digunakan dalam penelitian ini, dilakukan analisis distribusi terhadap lokasi properti berdasarkan kecamatan. Analisis ini penting untuk melihat sebaran geografis data serta potensi ketimpangan jumlah entri antar wilayah, yang dapat memengaruhi performa model prediktif. Gambar 4 dan 5 menyajikan distribusi lokasi properti terhadap kecamatan.



Gambar 4. Distribusi lokasi bagian pertama



Gambar 5. Distribusi lokasi bagian kedua

Gambar 4 dan 5 menunjukkan distribusi jumlah data harga rumah berdasarkan kecamatan di wilayah Yogyakarta dan sekitarnya. Terlihat bahwa terdapat konsentrasi data yang tinggi pada beberapa kecamatan seperti Ngaglik, Depok, dan Sleman, sementara kecamatan lain memiliki jumlah *entri* yang sangat sedikit. Distribusi yang tidak merata ini menjadi pertimbangan penting dalam proses pelatihan dan evaluasi model machine learning, terutama dalam hal generalisasi terhadap wilayah dengan data terbatas. Gambar 6 memaparkan banyaknya rumah dan rata-rata harganya.

	kabupaten	house_count	average_price
0	Bantul	321	9.815950e+08
1	Gunung Kidul	8	9.181250e+08
2	Kulon Progo	37	4.755405e+08
3	Sleman	1312	1.778995e+09
4	Yogyakarta	342	3.050509e+09

Gambar 6. Banyaknya rumah dan rata-rata harga

Setelah dilakukan proses pelatihan dan pengujian terhadap ketiga model regresi yang digunakan, diperoleh hasil evaluasi sebagaimana disajikan pada Tabel 1. Evaluasi dilakukan terhadap data uji menggunakan dua metrik utama, yaitu *Mean Squared Error* (MSE) dan koefisien determinasi (R^2 Score).

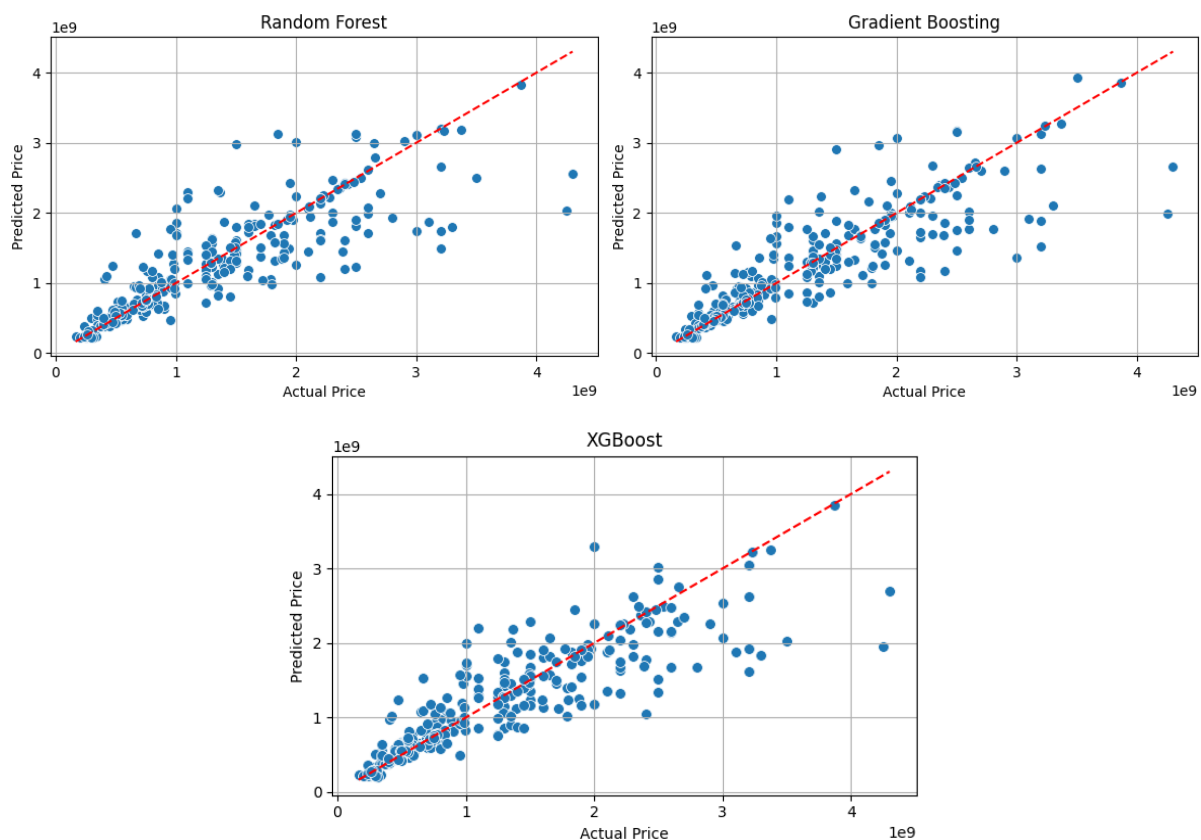
Tabel 1. Hasil evaluasi ketiga model

Model	MSE (dalam satuan rupiah ²)	R^2 Score
<i>Random Forest</i>	174.650.899.586.738.144	0.7471
<i>Gradient Boosting</i>	163.016.231.938.791.712	0.7639
<i>XGBoost</i>	155.621.063.915.530.144	0.7746

Berdasarkan hasil evaluasi di atas, model *XGBoost* menunjukkan performa terbaik dibandingkan dua model lainnya. Hal ini terlihat dari nilai MSE yang paling rendah dan R^2 tertinggi sebesar 0.7746, yang menunjukkan bahwa sekitar 77,46% variasi harga rumah dapat dijelaskan oleh model ini. Sementara itu, *Random Forest* mencatat nilai R^2 sebesar 0.7471 dan *Gradient Boosting* sebesar 0.7639, yang juga menunjukkan performa yang cukup baik namun masih di bawah *XGBoost*. *XGBoost* memiliki keunggulan dalam hal pengendalian *overfitting* melalui regulasi *reg_alpha* dan *reg_lambda*, serta fleksibilitas dalam pengaturan parameter *boosting* dan *sampling*. Selain itu, kemampuan algoritma ini dalam menangani interaksi fitur yang kompleks menjadikannya cocok untuk data tabular seperti harga properti.

Untuk memperoleh gambaran visual mengenai performa model regresi dalam memprediksi harga rumah, dilakukan visualisasi perbandingan antara harga aktual dan harga hasil prediksi dari tiga algoritma: *Random Forest*, *Gradient Boosting*, dan *XGBoost*. Visualisasi ditampilkan dalam bentuk scatter plot dengan garis merah putus-putus sebagai garis ideal ($y = x$), yang menunjukkan posisi prediksi sempurna.

Berdasarkan *scatter plot* pada Gambar 7, ketiga model cenderung menghasilkan prediksi yang cukup akurat di rentang harga menengah (sekitar Rp 500 juta – Rp 2 miliar), ditunjukkan oleh kepadatan titik di sekitar garis ideal. Namun, terdapat pola *underprediction* yang cukup jelas pada harga tinggi ($> \text{Rp } 2,5 \text{ miliar}$), terutama pada model *Gradient Boosting*, di mana banyak titik jatuh di bawah garis ideal. Sebaliknya, model *Random Forest* dan *XGBoost* menunjukkan sebaran yang lebih merata dan mendekati garis ideal, meskipun masih terdapat beberapa *overprediction* pada harga rendah dan *outlier* di harga ekstrem. Hal ini mengindikasikan bahwa *XGBoost* dan *Random Forest* lebih stabil dalam menangani variasi harga dibandingkan *Gradient Boosting*. Perbedaan pola *error* ini penting untuk dipertimbangkan dalam pemilihan model yang paling sesuai dengan distribusi harga properti di wilayah Yogyakarta.



Gambar 7. Visualisasi masing-masing model

Berdasarkan gambar *scatter plot* yang membandingkan harga rumah aktual dengan hasil prediksi dari tiga model regresi yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost*, dapat dilihat bahwa ketiganya mampu mengikuti pola umum dari data, meskipun terdapat perbedaan tingkat akurasi. Model *Random Forest* menunjukkan sebaran titik yang cukup rapat di sekitar garis ideal ($y = x$), menandakan bahwa model ini mampu memprediksi harga dengan akurasi yang cukup baik, meskipun masih terdapat beberapa *outlier* terutama pada harga tinggi.

Temuan ini sejalan dengan hasil studi sebelumnya oleh Fitri [15], yang menunjukkan bahwa *Random Forest* memberikan performa terbaik dalam memprediksi harga rumah di Jakarta, dibandingkan *Gradient Boosted Trees* dan *Linear Regression*. Namun, pada konteks data lokal Yogyakarta, penelitian ini menunjukkan bahwa *XGBoost* mampu melampaui performa *Random Forest*, menandakan pentingnya pemilihan model yang sesuai dengan karakteristik data wilayah tertentu. Dari sisi praktik, hasil model ini dapat dimanfaatkan oleh berbagai pihak. Bagi pengembang properti dan agen *real estate*, model prediktif dapat digunakan untuk menentukan harga jual atau beli yang lebih akurat. Pemerintah daerah juga dapat menggunakan model ini sebagai dasar dalam menyusun kebijakan tata ruang, perencanaan perumahan, atau evaluasi nilai properti secara makro.

Adapun keterbatasan dalam penelitian ini mencakup tidak digunakannya validasi silang (*K-Fold*) dalam proses evaluasi, serta tidak dimasukkannya faktor-faktor eksternal seperti aksesibilitas transportasi, kondisi pasar, dan variabel ekonomi makro lainnya. Selain itu, beberapa *outlier* dengan harga ekstrem belum ditangani secara khusus, yang dapat memengaruhi performa model dalam skenario *real-world*. Penelitian lanjutan disarankan untuk mengatasi keterbatasan tersebut dan mengeksplorasi integrasi variabel tambahan serta teknik *feature selection* dan *hyperparameter tuning*.

4. Kesimpulan

Penelitian ini membandingkan tiga algoritma regresi—*Random Forest*, *Gradient Boosting*, dan *XGBoost*—untuk memprediksi harga rumah di Yogyakarta. Hasil menunjukkan bahwa *XGBoost* memberikan performa terbaik dengan MSE sebesar $1,56 \times 10^{14}$ rupiah², R^2 sebesar 0,7746, dan RMSE sekitar 12,5 juta rupiah. Keunggulan *XGBoost* terletak pada kemampuannya menangani interaksi fitur kompleks serta fitur regularisasi yang mengurangi *overfitting*. Visualisasi prediksi menunjukkan bahwa *XGBoost* dan *Random Forest* lebih konsisten dibanding *Gradient Boosting*, khususnya pada harga rumah tinggi. Kebaruan penelitian ini terletak pada penggunaan *pipeline preprocessing* yang terstruktur dan perbandingan model secara sistematis pada data lokal Yogyakarta, yang belum banyak dieksplorasi sebelumnya. Secara praktis, hasil model ini bermanfaat bagi pengembang properti, agen *real estate*, maupun pemerintah daerah dalam perencanaan, penilaian, dan pengambilan keputusan harga rumah. Evaluasi di masa mendatang dapat dilengkapi dengan metrik lain seperti MAE dan MAPE untuk hasil yang lebih komprehensif.

Daftar Pustaka

- [1] E. Febrion Rahayuningtyas, F. Novia Rahayu, Y. Azhar, and I. Artikel, "Prediksi Harga Rumah Menggunakan General Regression Neural Network," J. Inform., vol. 8, no. 1, 2021, [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Real>
- [2] S. N. Asri, A. W. Hasyim, W. Dwi, P. Jurusan, P. Wilayah, and D. Kota, "FAKTOR-FAKTOR YANG MEMPENGARUHI HARGA LAHAN PERMUKIMAN DI KOTA MALANG."
- [3] D. Permata Sari, "FAKTOR-FAKTOR YANG MEMPENGARUHI KEPUTUSAN PEMBELIAN, KUALITAS PRODUK, HARGA KOMPETITIF, LOKASI (LITERATURE REVIEW MANAJEMEN PEMASARAN)," vol. 2, no. 4, 2021, doi: 10.31933/jimt.v2i4.
- [4] Hendra Di Kesuma, D. Apriadi, H. Juliansa, and E. Etriyanti, "Implementasi Data Mining Prediksi Mahasiswa Baru Menggunakan Algoritma Regresi Linear Berganda," J. Ilm. Bin. STMIK Bina Nusantara. Jaya Lubuklinggau, vol. 4, no. 2, pp. 62–66, Oct. 2022, doi: 10.52303/jb.v4i2.74.
- [5] E. Lette, M. Zunaidi, and W. R. Maya, "Prediksi Penjualan Crude Palm Oil (CPO) Menggunakan Metode Regresi Linear Berganda," J. Sist. Inf. Triguna Dharma (JURSI TGD), vol. 1, no. 3, p. 128, May 2022, doi:

10.53513/jursi.v1i3.5106.

- [6] I. Muthahharah and Inayanti Fatwa, "Analisis Regresi Linear Berganda Untuk Media Pembelajaran Daring Terhadap Prestasi Belajar Mahasiswa di STKIP Pembangunan," *J. MSA (Mat. dan Stat. serta Apl.)*, vol. 10, no. 1, pp. 53–60, Jun. 2022, doi: 10.24252/msa.v10i1.25145.
- [7] A. N. Maharadja, I. Maulana, and B. A. Dermawan, "Penerapan Metode Regresi Linear Berganda untuk Prediksi Kerugian Negara Berdasarkan Kasus Tindak Pidana Korupsi," *J. Appl. Informatics Comput.*, vol. 5, no. 1, pp. 95–102, Jul. 2021, doi: 10.30871/jaic.v5i1.3184.
- [8] R. Yohanes and D. Lasut, "Web-Based used Car Price Prediction Application with Linear Regression Method," *bit-Tech*, vol. 7, no. 3, pp. 687–695, Apr. 2025, doi: 10.32877/bt.v7i3.1722.
- [9] F. M. Talaat, A. Aljadani, M. Badawy, and M. Elhosseini, "Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4847–4865, Mar. 2024, doi: 10.1007/s00521-023-09232-2.
- [10] J. Gao, Y. Lu, N. Ashrafi, I. Domingo, K. Alaei, and M. Pishgar, "Prediction of sepsis mortality in ICU patients using machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 228, Aug. 2024, doi: 10.1186/s12911-024-02630-z.
- [11] A. A. Rasheed, "Improving prediction efficiency by revolutionary machine learning models," *Mater. Today Proc.*, vol. 81, pp. 577–583, 2023, doi: 10.1016/j.matpr.2021.04.014.
- [12] E. B. El Hakim and J. Aryanto, "Automated Maintenance System For Freshwater Aquascape Based On The Internet Of Things (Iot)," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 1, Nov. 2024, doi: 10.26877/asset.v6i1.17951.
- [13] Y. S. Zakaria, N. A. Ariffin, A. Ahmad, R. Rainis, A. M. Muslim, and W. M. M. Wan Ibrahim, "Optimizing Tuberculosis Treatment Predictions: A Comparative Study of XGBoost with Hyperparameter in Penang, Malaysia," *Sains Malaysiana*, vol. 54, no. 1, pp. 3741–3752, Jan. 2025, doi: 10.17576/jsm-2025-5401-22.
- [14] L. Alfaris et al., "Predicting Ocean Current Temperature Off the East Coast of America with XGBoost and Random Forest Algorithms Using Rstudio," *ILMU Kelaut. Indones. J. Mar. Sci.*, vol. 29, no. 2, pp. 273–284, Jun. 2024, doi: 10.14710/ik.ijms.29.2.273-284.
- [15] 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.
- [16] 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, May 2025, doi: 10.69916/jkbt.v4i2.318.
- [17] A. N. M. Pudjianto and E. Y. Hidayat, "Perbandingan Prediksi Depresi Mahasiswa dengan Linear Regression, Random Forest, dan Gradient Boosting," *SINTECH (Science Inf. Technol. J.)*, vol. 7, no. 3, pp. 180–189, Dec. 2024, doi: 10.31598/sintechjournal.v7i3.1729.
- [18] M. Bagus Prayogi and F. Apriani, "PREDIKSI ANGKA HARAPAN HIDUP MENGGUNAKAN RANDOM FOREST DAN XGBOOST REGRESSION," 2025.
- [19] N. Hassan, S. T. Sheikh Abdul Kadir, M. L. Husain, B. Satyanarayana, M. A. Ambak, and A. M. Ghaffar, "Weight Prediction for Fishes in Setiu Wetland, Terengganu, using Machine Learning Regression Model," *BIO Web Conf.*, vol. 73, p. 01007, Nov. 2023, doi: 10.1051/bioconf/20237301007.
- [20] D. Ignasius, M. Akrom, and S. Budi, "Comparative Analysis of Linear Regression, Decision Tree, and Gradient Boosting Models for Predicting Drug Corrosion Inhibition Efficiency Using QSAR Descriptors," *Fakt. Exacta*, vol. 17, no. 3, p. 251, Sep. 2024, doi: 10.30998/faktorexacta.v17i3.24679.
- [21] Nurdin, N. Suarna, and W. Prihartono, "ALGORITMA REGRESI LINIER SEDERHANA UNTUK PREDIKSI PENGGUNAAN VOLUME AIR BERDASARKAN JENIS PELANGGAN PDAM," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 4, no. 1, pp. 43–52, Jan. 2025, doi: 10.69916/jkbt.v4i1.187.
- [22] I. D. Hartarti, I. A. Septiyani, D. A. Gultom, Y. Hendrian, and S. L. Kinanti, "Prediksi Harga Rumah di Boston Dengan Model Regresi Linear Menggunakan Python," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 2, pp. 4250–4256, Jun. 2025, doi: 10.31004/riggs.v4i2.1210.
- [23] Pramudya Dika, "Dataset Harga Tanah Yogyakarta," <https://www.kaggle.com/datasets/pramudyadika/yogyakarta-housing-price-ndonesia/data>.
- [24] A. Aliberti, Y. Xin, A. Viticchié, E. Macii, and E. Patti, "Comparative analysis of neural networks techniques to forecast Airfare Prices," in *Proceedings - International Computer Software and*

- Applications Conference, 2023. doi: 10.1109/COMPSAC57700.2023.00157.
- [25] M. Raparathi, D. Dhabliya, T. Kumari, R. Upadhyaya, and A. Sharma, "Implementation and Performance Comparison of Gradient Boosting Algorithms for Tabular Data Classification," 2024, pp. 461–479. doi: 10.1007/978-981-97-4533-3_36.
 - [26] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, p. 88, Feb. 2025, doi: 10.3390/technologies13030088.
 - [27] N. S. Ahmed, "Machine Learning Models for Pavement Structural Condition Prediction: A Comparative Study of Random Forest (RF) and eXtreme Gradient Boosting (XGBoost)," *Open J. Civ. Eng.*, vol. 14, no. 04, pp. 570–586, 2024, doi: 10.4236/ojce.2024.144031.
 - [28] Y. Wu, Z. Zhang, X. Qi, W. Hu, and S. Si, "Prediction of flood sensitivity based on Logistic Regression, eXtreme Gradient Boosting, and Random Forest modeling methods," *Water Sci. Technol.*, vol. 89, no. 10, pp. 2605–2624, May 2024, doi: 10.2166/wst.2024.146.
 - [29] K. Ita and J. Prinze, "Machine learning for skin permeability prediction: random forest and XG boost regression," *J. Drug Target.*, vol. 32, no. 1, pp. 57–65, Jan. 2024, doi: 10.1080/1061186X.2023.2284096.
 - [30] A. K. Jana and S. Saha, "Comparative Performance analysis of Machine Learning Algorithms for stability forecasting in Decentralized Smart Grids with Renewable Energy Sources," in 2024 International Conference on Electrical, Computer and Energy Technologies (ICECET, IEEE, Jul. 2024, pp. 1–7. doi: 10.1109/ICECET61485.2024.10698410.
 - [31] Y. Qu et al., "Migratable Power System Transient Stability Assessment Method Based on Improved XGBoost," *Energy Eng.*, vol. 121, no. 7, pp. 1847–1863, 2024, doi: 10.32604/ee.2024.048300.