

PERBANDINGAN KINERJA ALGORITMA EXTREME GRADIENT BOOSTING DAN RANDOM FOREST UNTUK PREDIKSI HARGA RUMAH DI JABODETABEK

Dhiwa Aqsha ¹⁾

¹⁾ Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Tomang, Grogol Petamburan, RT.6/RW.16, Kota Jakarta Barat, Daerah Khusus
Ibu Kota Jakarta 11440, Indonesia
email : dhiwa.535210087@stu.untar.ac.id

ABSTRACT

The demand for housing continues to increase along with population growth. Predicting house prices is crucial to assist prospective buyers and investors in making more informed decisions. This study aims to predict house prices in the Jabodetabek area by comparing the performance of two machine learning algorithms, namely Extreme Gradient Boosting and Random Forest, to produce accurate price estimates. The prediction process includes data preprocessing, key variable exploration, and model evaluation using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R²). The results show that the Random Forest model performs best, with an MAE of 95,200,513.25, an MSE of 1.47e+19, and an R² of 0.77, outperforming the Extreme Gradient Boosting model with an MAE of 121,836,703.27, an MSE of 3.03e+19, and an R² of 0.52. Thus, this research is expected to serve as an effective tool for stakeholders in mitigating risks in property investment decisions in the Jabodetabek area.

Key words

Analysis, extreme gradient boosting, house price prediction, machine learning, random forest

1. Pendahuluan

Jabodetabek merupakan kawasan megapolitan dengan populasi yang sangat besar, serta kemajuan signifikan dalam bidang pendidikan, perbankan, dan ekonomi, yang membuatnya lebih maju dibandingkan daerah lain di Indonesia [1][2][3]. Kawasan ini terdiri dari gabungan beberapa wilayah metropolitan yang masing-masing memiliki jumlah penduduk lebih dari 5 juta jiwa, membentuk pusat urban yang dinamis dan terintegrasi. Seiring dengan pertumbuhan populasi di Jabodetabek, kebutuhan masyarakat akan tempat tinggal terus meningkat. Rumah menjadi salah satu kebutuhan dasar yang penting bagi manusia. Rumah berfungsi sebagai tempat tinggal, ruang untuk bersantai, berkumpul

bersama keluarga, berlindung, serta tempat beristirahat setelah beraktivitas sehari-hari [4].

Kenaikan harga rumah dari tahun ke tahun disebabkan oleh berbagai faktor, termasuk meningkatnya permintaan akan tempat tinggal, daya tarik rumah sebagai aset investasi jangka panjang, serta pembangunan infrastruktur yang pesat di beberapa area. Faktor-faktor ini saling terikat satu sama lain dalam mendorong kenaikan nilai properti di wilayah-wilayah tertentu [5]. Sebagai kawasan megapolitan, Jabodetabek mengalami pertumbuhan ekonomi yang cepat dan tingkat kepadatan penduduk yang tinggi, yang berdampak pada kenaikan harga perumahan. Kondisi ini menciptakan tantangan tersendiri bagi calon pembeli dan investor properti dalam menilai dan memilih harga rumah yang sesuai dengan kebutuhan serta preferensi mereka.

Dalam investasi jangka panjang seperti rumah, diperlukan analisis harga rumah untuk mengurangi risiko kerugian. Analisis harga rumah juga dapat membantu pembeli untuk memprediksi harga berdasarkan preferensi yang dimiliki pembeli. Oleh karena itu, dibutuhkan sistem dengan penerapan pemodelan *machine learning*. *Machine learning* telah digunakan sebagai metode analitis yang dapat membantu memprediksi harga properti dengan mempertimbangkan berbagai faktor yang relevan. Prediksi harga rumah dapat memanfaatkan berbagai algoritma *machine learning*, seperti *extreme gradient boosting* dan *random forest*, untuk menghasilkan estimasi harga berdasarkan data masa lalu dan kondisi saat ini.

Dalam penelitian sebelumnya, Cep Haryanto, Nining Rahaningsih, dan Fadhil Muhammad Basysyar melakukan studi berjudul *Komparasi Algoritma Machine Learning dalam Memprediksi Harga Rumah*. Penelitian ini membandingkan algoritma *Multiple Linear Regression* dan *Random Forest Regression* untuk memprediksi harga rumah, dengan hasil menunjukkan bahwa *Random Forest Regression* memiliki akurasi yang lebih tinggi, yaitu sebesar 81,6%, dibandingkan *Multiple Linear Regression* yang mencapai 78,5% [6].

Dalam penelitian sebelumnya juga, Nuzuliarini Nuris dari Universitas Bina Sarana Informatika melakukan studi berjudul *Analisis Prediksi Harga Rumah Pada Machine*

Learning Menggunakan Metode Regresi Linear. Penelitian ini menggunakan algoritma *Regresi Linear* dengan data yang diperoleh dari website *rumah123.com*. Hasilnya menunjukkan tingkat akurasi sebesar 86,41%, dengan variabel luas bangunan memiliki pengaruh paling kuat terhadap harga rumah (korelasi 0,80), sedangkan jenis sertifikat tidak memberikan dampak signifikan (korelasi 0,00). Model dievaluasi menggunakan metrik seperti *Mean Squared Error* (MSE) sebesar 464,87 dan *Mean Absolute Error* (MAE) sebesar 313,00 [7]. Berdasarkan latarbelakang yang telah dipaparkan, penelitian ini secara garis besar akan membandingkan algoritma *Extreme Gradient Boosting* dan *Random Forest* untuk memprediksi harga rumah di Jabodetabek. Penelitian ini bertujuan untuk memprediksi harga rumah sesuai dengan preferensi mereka dan juga membantu para investor untuk lebih akurat memprediksi harga rumah, sehingga dapat mengurangi risiko kerugian. point.

2. Metode Penelitian

Dalam penelitian ini, penulis menerapkan sejumlah tahapan untuk mencapai tujuan yang diinginkan. Tahapan-tahapan tersebut diperlihatkan pada Gambar 1: Tahapan Penelitian, yang menyajikan gambaran terstruktur mengenai langkah-langkah utama yang diambil penulis dalam proses penelitian ini. Gambar tersebut membantu memperjelas alur dan strategi yang digunakan dalam pencapaian tujuan penelitian.



Gambar 1 Tahapan Penelitian

2.1 Pengumpulan Data

Dalam penelitian ini, data yang digunakan tergolong sebagai data sekunder. Data sekunder adalah jenis

informasi yang diperoleh tidak langsung dari sumber asli, melainkan melalui perantara atau pihak ketiga [8][9]. Data sekunder dalam penelitian ini diambil dari situs *Kaggle.com*, dengan judul dataset “Daftar Harga Rumah Jabodetabek” dan nama file *jabodetabek_house_price.csv*.

Dataset ini terdiri dari 3.554 sampel dan mencakup 27 variabel, dengan 26 variabel independen dan 1 variabel dependen. Keterangan serta atribut dari setiap variabel dijelaskan pada Tabel 1 berikut:

Tabel 1 Variabel dalam dataset

Variabel	Keterangan
url	Tautan unik yang mengarahkan ke halaman detail properti di situs <i>rumah123.com</i> .
price_in_rp	Harga properti dalam Rupiah.
title	Judul iklan properti
address	Alamat lengkap properti, termasuk jalan dan area perumahan.
District	Kecamatan atau area tempat properti berada, berguna untuk segmentasi wilayah.
City	Kota tempat properti berada, seperti Jakarta, Bekasi, atau Depok.
Lat	Latitude (garis lintang) lokasi properti untuk pemetaan geografis.
Long	Longitude (garis bujur) lokasi properti untuk pemetaan geografis.
facilities	Fasilitas yang tersedia di properti atau lingkungan sekitar, seperti taman, kolam renang, atau <i>jogging track</i> .
property_type	Jenis properti, seperti rumah, apartemen, atau ruko.
ads_id	ID unik iklan untuk identifikasi dalam sistem.
bedrooms	Jumlah kamar tidur di properti tersebut.
bathrooms	Jumlah kamar mandi.
land_size_m2	Luas tanah dalam meter persegi (m ²), digunakan untuk menilai nilai properti berdasarkan ukuran tanah.
building_size_m2	Luas bangunan dalam meter persegi (m ²).
carports	Jumlah tempat parkir kendaraan yang tersedia di properti.
certificate	Jenis sertifikat kepemilikan properti, seperti SHM (Sertifikat Hak Milik) atau HGB (Hak Guna Bangunan).
electricity	Kapasitas listrik dalam MAH.
maid_rooms	Jumlah kamar tidur untuk asisten rumah tangga.
maid_bathrooms	Jumlah kamar mandi untuk asisten rumah tangga.
floors	Jumlah lantai yang dimiliki properti.
building_age	Usia bangunan dalam tahun.

year_built	Tahun dibangunnya properti.
property_condition	Kondisi properti, misalnya "bagus" atau "bagus sekali".
building_orientation	Arah orientasi bangunan (misalnya menghadap utara atau selatan).
garages	Jumlah garasi yang tersedia di properti.
furnishing	Status perabotan properti

2.2 Data Preprocessing

Pada tahap ini, data diproses melalui beberapa langkah, salah satunya adalah pengecekan terhadap *missing values*. Proses ini bertujuan untuk membersihkan dataset dari data yang kosong atau hilang, sehingga kualitas data yang digunakan menjadi lebih baik. Langkah selanjutnya adalah normalisasi data, yang berfungsi sebagai proses transformasi data untuk menyeimbangkan nilai pada setiap *record*, mengingat adanya perbedaan rentang nilai pada masing-masing data. Tahap terakhir adalah pembagian data menjadi dua bagian, yaitu data latih dan data uji.

2.3 Split Data

Pembagian data latih dan data uji adalah langkah penting dalam pengembangan model *Machine Learning*. Data latih digunakan untuk melatih model agar dapat mengenali dan memahami pola-pola dalam data, sehingga model dapat melakukan prediksi dengan lebih akurat saat diuji pada data baru. Biasanya Sebagian besar data dialokasikan untuk set pelatihan. Hal ini bertujuan agar model memiliki cukup banyak informasi untuk mempelajari pola data dengan baik sebelum diterapkan pada data pengujian. Dalam proses pelatihan, model akan menyesuaikan parameter internalnya menggunakan data pelatihan, sehingga mampu memahami hubungan antara fitur-fitur dalam data dan output target yang diharapkan [10].

Sebaliknya, set pengujian berfungsi untuk mengevaluasi kemampuan model dalam memprediksi data baru yang belum pernah dilihat sebelumnya, sehingga dapat mengukur seberapa baik model yang telah dilatih bekerja pada data yang tidak dikenal. Data untuk set pengujian biasanya dialokasikan dalam proporsi yang lebih kecil. Untuk memastikan pengujian model dilakukan pada data yang benar-benar independen. Evaluasi kinerja model dilakukan dengan set pengujian ini, memungkinkan penilaian akurasi model dalam memprediksi data baru sekaligus mengidentifikasi potensi *overfitting* [16].

2.4 Modelling

Model algoritma *Extreme Gradient Boosting* dan *Random Forest* yang dirancang dalam penelitian ini akan berfungsi sebagai alat prediktif untuk memperkirakan harga rumah. Model-model ini menggunakan berbagai

fitur dalam dataset sebagai dasar untuk menentukan harga, dengan tujuan menghasilkan prediksi yang akurat berdasarkan pola dan hubungan antarvariabel yang ada di dalam data.

2.4.1 Extreme Gradient Boosting

XGBoost (*Extreme Gradient Boosting*) adalah algoritma *tree boosting* yang sangat dapat diskalakan dan dikenal sebagai salah satu metode *machine learning* yang efektif dan banyak digunakan untuk memprediksi nilai dari suatu dataset. Algoritma ini menggunakan pohon keputusan sebagai model dasar, kemudian menerapkan teknik penguatan untuk meningkatkan performa prediksi. XGBoost dilengkapi dengan fitur-fitur seperti regularisasi, pemrosesan paralel, dan penanganan nilai yang hilang, sehingga memperkuat keandalannya. Selain itu, XGBoost mengoptimalkan pola akses cache, kompresi data, dan sharding untuk membentuk sistem penguatan pohon yang efisien dan skalabel. Kombinasi fitur ini memungkinkan XGBoost memproses data berukuran besar dengan efisiensi sumber daya yang lebih tinggi dibandingkan sistem lainnya [13][14].

Pada model pohon boosting tradisional, informasi yang digunakan untuk meningkatkan model hanya berdasarkan turunan pertama dari fungsi kerugian. Tantangan muncul pada pelatihan pohon ke-*n*, di mana sulit menerapkan pelatihan terdistribusi karena sisa dari pohon sebelumnya (ke-*n-1*) diperlukan untuk kelanjutan pelatihan. XGBoost mengatasi hal ini dengan menggunakan ekspansi Taylor orde kedua pada fungsi kerugian, yang memberikan informasi tambahan untuk meningkatkan akurasi. Selain itu, XGBoost secara otomatis memanfaatkan *multithreading* pada CPU untuk pemrosesan paralel, serta menerapkan berbagai teknik untuk mengurangi risiko *overfitting* [15].

2.4.2 Random Forest

Random Forest adalah metode *ensemble learning* yang sering digunakan untuk tugas-tugas seperti klasifikasi, regresi, dan lainnya. Pendekatan ini bertujuan untuk menyelesaikan berbagai jenis masalah, seperti klasifikasi untuk mengelompokkan data ke dalam kategori tertentu dan regresi untuk memperkirakan nilai berdasar data yang tersedia. Inti dari metode ini adalah membangun banyak pohon keputusan yang dilatih bersamaan, sehingga meningkatkan kekuatan prediksi.

Dalam klasifikasi, Random Forest menghasilkan prediksi berdasarkan suara mayoritas dari seluruh pohon yang dibangun: kelas yang paling banyak dipilih oleh pohon-pohon tersebut akan menjadi hasil akhir. Sementara dalam regresi, hasil akhirnya diperoleh dengan mengambil rata-rata dari prediksi setiap pohon, sehingga menghasilkan estimasi yang lebih stabil.

Keunggulan utama Random Forest adalah kemampuannya untuk mengatasi *overfitting*, suatu masalah yang sering terjadi pada pohon keputusan tunggal. *Overfitting* terjadi saat pohon keputusan terlalu

sesuai dengan data pelatihan, sehingga kurang mampu menggeneralisasi data baru. Dengan menggabungkan hasil dari berbagai pohon keputusan yang berbeda, Random Forest dapat mengurangi risiko *overfitting*, memberikan hasil yang lebih stabil dan tahan terhadap variabilitas. Namun, meskipun Random Forest sering kali unggul dibandingkan pohon keputusan tunggal, metode ini mungkin masih memiliki akurasi yang lebih rendah dibandingkan pendekatan seperti *gradient boosted trees*. Pemilihan metode terbaik seringkali bergantung pada karakteristik data dan tujuan analisis yang spesifik.

Random Forest dikenal luas karena berbagai kelebihanannya, seperti ketahanannya terhadap perbedaan skala dan transformasi nilai fitur, kemampuannya menangani fitur-fitur yang tidak relevan, serta interpretasi model yang relatif mudah. Namun, kelemahannya adalah bahwa tingkat akurasinya mungkin tidak selalu tinggi dalam setiap kasus. Oleh karena itu, dalam memilih metode ini, perlu mempertimbangkan antara keunggulan stabilitas prediksi dan tingkat akurasi yang diharapkan sesuai dengan tugas yang dihadapi. Metode ini juga memiliki beberapa teknik untuk mengukur pentingnya variabel, antara lain:

- Variable Importance (*Pentingnya Variabel*): Teknik ini mengukur seberapa besar pengaruh suatu variabel terhadap performa model dengan menghitung penurunan akurasi jika variabel tersebut dihapus.
- Permutation Importance (*Pentingnya Permutasi*): Metode ini melibatkan permutasi nilai dalam variabel tertentu dan mengukur dampaknya pada akurasi model. Jika akurasi menurun signifikan, variabel tersebut dianggap penting.
- Mean Decrease in Impurity Feature Importance (*Pentingnya Fitur Berdasarkan Penurunan Ketidakmurnian Rata-rata*): Teknik ini menilai kontribusi setiap variabel dalam pengurangan ketidakmurnian atau *classification error* di setiap pohon dalam Random Forest. Variabel yang lebih banyak mengurangi ketidakmurnian dianggap lebih signifikan.

Ketiga teknik ini berperan dalam memahami dan menyeleksi fitur-fitur yang penting, serta memberikan wawasan lebih mendalam tentang faktor-faktor yang memengaruhi prediksi model. Skor pentingnya variabel yang tinggi menunjukkan dampak besar variabel tersebut terhadap kinerja model, sehingga dapat membantu dalam proses seleksi fitur atau optimasi model.

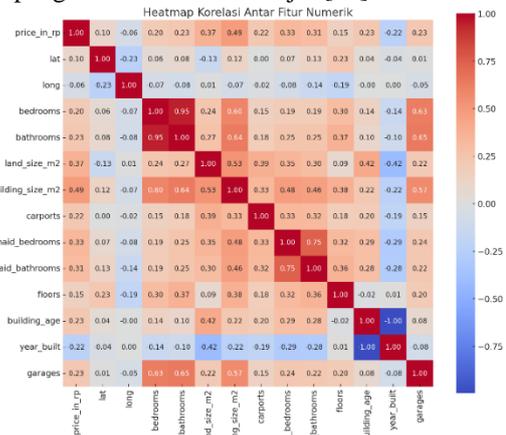
2.5 Evaluasi Model

Tahap evaluasi berfokus pada penilaian kualitas model yang telah dilatih menggunakan algoritma *machine learning* pada tahap sebelumnya. Pada tahap ini, model dievaluasi untuk mengukur performanya dan menentukan seberapa baik model tersebut mampu menghasilkan prediksi yang akurat sesuai dengan tujuan penelitian [11][12]. point.

3. Hasil dan Pembahasan

3.1 Eksplorasi Data

Eksplorasi data bertujuan untuk memperoleh pemahaman yang mendalam mengenai komponen dan isi data yang dianalisis, sehingga dapat mengidentifikasi informasi penting yang relevan dan bermanfaat dalam proses pengolahan data lebih lanjut [16].



Gambar 2 Korelasi Antar Fitur Numerik Dalam Dataset

Diagram heatmap pada gambar 2 menunjukkan hubungan antar fitur numerik dalam dataset. Visualisasi ini membantu mengidentifikasi fitur-fitur yang memiliki hubungan kuat satu sama lain. Heatmap menunjukkan hubungan yang kuat antara beberapa fitur spesifikasi rumah, seperti jumlah kamar tidur dan kamar mandi. Korelasi tinggi ini mengindikasikan bahwa rumah dengan lebih banyak kamar tidur umumnya memiliki lebih banyak kamar mandi, yang mencerminkan standar fasilitas di rumah-rumah besar. Selain itu, fitur seperti garages, maid_bedrooms, dan maid_bathrooms juga berkorelasi, menunjukkan bahwa fasilitas tambahan ini sering ditemukan bersama di properti yang lebih luas atau lebih mewah.

Dari sisi harga, variabel *building_size_m2* (luas bangunan) dan *land_size_m2* (luas tanah) menunjukkan hubungan yang kuat terhadap *price_in_rp*, menegaskan bahwa luas bangunan dan tanah adalah faktor penting dalam menentukan harga rumah. Semakin besar ukuran properti, semakin tinggi pula kecenderungan harga yang diberikan. Insight ini dapat membantu dalam memilih variabel yang relevan untuk membangun model prediksi harga rumah yang lebih akurat.



Gambar 3 Tren Harga Rumah Berdasarkan Jumlah Lantai.

Berdasarkan visualisasi harga rumah berdasarkan jumlah lantai yang dapat dilihat pada Gambar 3, menunjukkan bahwa terdapat tren peningkatan harga seiring bertambahnya jumlah lantai rumah. Rumah bertingkat lebih dari satu lantai cenderung memiliki harga yang lebih tinggi, kemungkinan karena faktor seperti luas bangunan atau fasilitas tambahan. Selain itu, pada rumah dengan satu lantai, harga cenderung lebih stabil dengan rentang yang tidak terlalu luas, mengindikasikan bahwa spesifikasi dan fitur di kelas ini relatif homogen.

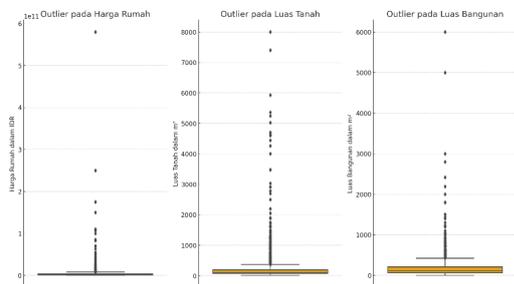
Namun, pada rumah dengan dua lantai atau lebih, rentang harga menunjukkan variasi yang cukup besar. Hal ini menunjukkan bahwa selain jumlah lantai, faktor-faktor lain seperti lokasi, kualitas material, dan desain rumah turut memengaruhi harga. Rumah bertingkat dua atau lebih juga tampaknya memiliki daya tarik tersendiri, terutama bagi pasar yang mengutamakan ruang yang lebih besar atau properti dengan kesan premium.

3.2 Pengecekan Kualitas Data

a. Missing Values

Terdapat beberapa fitur yang memiliki nilai hilang (missing values) dengan proporsi yang bervariasi. Kolom `building_orientation` memiliki nilai hilang tertinggi, yaitu sekitar 46.36%, diikuti oleh `year_built` dan `building_age` dengan masing-masing sekitar 40.67% nilai hilang. Kehadiran missing values yang signifikan ini dapat mempengaruhi kualitas analisis, terutama jika kolom-kolom tersebut memiliki keterkaitan penting dalam konteks prediksi harga rumah.

b. Outliers



Gambar 4 Outliers pada 3 fitur utama

Gambar 4 menunjukkan bahwa adanya outliers pada 3 fitur utama. Pada variabel harga rumah, terlihat adanya sejumlah outlier dengan nilai yang sangat tinggi dibandingkan dengan mayoritas data lainnya. Ini

mengindikasikan bahwa terdapat beberapa properti mewah atau premium dengan harga yang jauh di atas rata-rata. Kehadiran outlier semacam ini perlu diperhatikan, karena dapat mempengaruhi hasil analisis atau prediksi jika tidak ditangani dengan tepat.

Pada variabel luas tanah dan luas bangunan, outlier juga terlihat menonjol. Beberapa properti memiliki luas tanah dan bangunan yang jauh lebih besar dibandingkan dengan rata-rata dalam dataset. Outlier pada variabel ini mungkin mencerminkan rumah-rumah di kawasan elit atau properti dengan spesifikasi khusus seperti villa atau mansion yang umumnya memiliki lahan luas. Hal ini juga mengindikasikan adanya variasi yang besar dalam ukuran properti, yang perlu dipertimbangkan dalam analisis lebih lanjut.

3.3 Data Preparation

Proses data preparation dilakukan untuk memastikan bahwa data yang digunakan dalam analisis memiliki kualitas yang optimal dan siap untuk digunakan dalam pemodelan prediktif. Tahapan data preparation yang dilakukan meliputi:

1. Penanganan Missing Values

Beberapa kolom dalam dataset memiliki persentase nilai hilang yang cukup tinggi, terutama pada variabel `building_orientation`, `year_built`, dan `building_age`. Untuk memastikan kualitas data tetap optimal, dilakukan penghapusan baris atau kolom yang memiliki proporsi nilai hilang yang signifikan. Kolom dengan jumlah missing values yang melebihi ambang batas tertentu akan dihapus sepenuhnya, sementara baris dengan nilai hilang pada kolom-kolom penting juga akan dieliminasi. Langkah ini diambil untuk menjaga integritas data dan mencegah bias yang dapat muncul dari proses imputasi yang mungkin tidak merefleksikan nilai sebenarnya.

2. Penanganan Outlier

Berdasarkan hasil eksplorasi data, ditemukan sejumlah outlier pada variabel `price_in_rp`, `land_size_m2`, dan `building_size_m2`, yang berpotensi mempengaruhi hasil analisis atau prediksi secara signifikan. Untuk mengatasi hal ini, dilakukan penghapusan outlier pada data dengan metode *interquartile range (IQR)*, di mana data yang berada di luar rentang $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$ dianggap sebagai outlier dan dihapus dari dataset. Langkah ini diambil untuk menjaga konsistensi hasil model serta menghindari bias yang disebabkan oleh nilai ekstrem yang tidak representatif terhadap populasi data.

3. Normalisasi Data

Dilakukan Normalisasi atau standardisasi dikarenakan terdapat perbedaan antar variable, terutama antara variabel-variabel yang mengukur ukuran properti (misalnya, luas tanah dan luas bangunan) dengan variabel harga. Normalisasi data bertujuan untuk menyeimbangkan skala antar variabel sehingga dapat mencegah bias dalam algoritma pemodelan, khususnya model berbasis jarak atau gradien.

4. Konversi Variabel Kategorikal

Beberapa variabel dalam dataset berbentuk kategorikal, seperti *property_type* dan *furnishing*. Variabel-variabel ini perlu dikonversi menjadi format numerik agar dapat digunakan dalam pemodelan. Pengkodean dilakukan dengan metode *one-hot encoding* untuk variabel yang tidak memiliki urutan (nominal) atau *label encoding* untuk variabel yang memiliki urutan (ordinal).

3.4 Evaluasi model

Evaluasi model dilakukan untuk membandingkan performa dua algoritma, yaitu *Extreme Gradient Boosting* dan *Random Forest*, dalam memprediksi harga rumah. Tabel di bawah ini menyajikan tiga metrik evaluasi utama yang digunakan: *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *R-squared* (R^2).

Tabel 2 Evaluasi 3 Metrik Utama

Model	Mean Absolute Error	Mean Squared Error	R-squared
Extreme Gradient Boosting	1218367203.27	3.0251714430689	0.52
Random Forest	952005136.25	1.4686021378655	0.77

Mean Absolute Error (MAE) menunjukkan bahwa model *Random Forest* memiliki nilai kesalahan rata-rata yang lebih rendah dibandingkan *Extreme Gradient Boosting*, yaitu sekitar 95,200,513.25 dibandingkan 121,836,703.27. Hal ini menunjukkan bahwa *Random Forest* lebih akurat dalam memprediksi harga rumah dengan selisih kesalahan prediksi yang lebih kecil.

Mean Squared Error (MSE) juga menunjukkan hasil yang serupa, di mana nilai MSE *Random Forest* lebih rendah dibandingkan dengan *Extreme Gradient Boosting*. Nilai MSE yang lebih rendah pada *Random Forest* menunjukkan bahwa model ini lebih baik dalam menangani prediksi dengan kesalahan yang lebih kecil, terutama dalam menghadapi outlier.

R-squared, yang mengukur seberapa besar variasi dalam data yang dapat dijelaskan oleh model, memiliki nilai yang lebih tinggi pada *Random Forest* (0.77) dibandingkan *Extreme Gradient Boosting* (0.52). Ini menunjukkan bahwa *Random Forest* lebih mampu menangkap hubungan antara fitur dan harga rumah, sehingga model ini dianggap lebih andal untuk tugas prediksi dalam penelitian ini.

4. Kesimpulan

Berdasarkan hasil evaluasi menggunakan metrik *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *R-squared* (R^2), model *Random Forest* terbukti memiliki performa lebih baik dibandingkan *Extreme Gradient Boosting*. Model *Random Forest*

menunjukkan nilai MAE dan MSE yang lebih rendah, yang menunjukkan akurasi prediksi yang lebih tinggi dengan kesalahan yang lebih kecil. Selain itu, nilai R^2 model *Random Forest* yang lebih tinggi (0.77) mengindikasikan bahwa model ini mampu menjelaskan variasi harga rumah dengan lebih baik.

REFERENSI

- [1] E. D. Viana, F. Febrianti, and F. R. Dewi, "Literasi Keuangan, Inklusi Keuangan dan Minat Investasi Generasi Z di Jabodetabek," *Jurnal Manajemen dan Organisasi*, vol. 12, no. 3, pp. 252–264, Jan. 2022.
- [2] Permata, Nike Dyah, "Evaluasi Fungsi Pemanfaatan Hutan Kota Sebagai Kawasan Rekreasi Masyarakat Jabodetabek," *Ipb.ac.id*, 2019.
- [3] F. Febyanti, "Pemodelan Faktor-Faktor yang Mempengaruhi Harga Rumah di Jabodetabek Menggunakan Metode Regresi Probit," *Jurnal Riset Statistika*, pp. 50–56, Jul. 2022.
- [4] R. M. Sanusi, A. Siswo, and R. Wijaya, "Prediksi Harga Rumah Di Kota Bandung Bagian Timur Dengan Menggunakan Metode Regresi," *eProceedings of Engineering*, vol. 7, no. 3, 2020, Accessed: Oct. 31, 2024. [Online].
- [5] Umami Athiyah, Arnelka Hananta, Taufik Maulidi, Vico, T. Felix, and E. Angeline, "Sistem Pendukung Keputusan Prediksi Harga Rumah Kost untuk Mahasiswa IT Telkom Purwokerto Menggunakan Metode Fuzzy Tsukamoto Berbasis Web," *Journal of Dinda Data Science Information Technology and Data Analytics*, vol. 1, no. 2, pp. 77–81, Aug. 2021.
- [6] Cep Haryanto, Nining Rahaningsih, and Fadhil Muhammad Basysyar, "KOMPARASI ALGORITMA MACHINE LEARNING DALAM MEMPREDIKSI HARGA RUMAH," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 533–539, Mar. 2023.
- [7] N. Nuris, "Analisis Prediksi Harga Rumah Pada Machine Learning Metode Regresi Linear," *Explore*, vol. 14, no. 2, pp. 108–112, Jul. 2024.
- [8] Sutisna Sutisna and Mirsandi Nazar Yuniar, "Klasifikasi Kualitas Air Bersih Menggunakan Metode Naïve baiyes," *Jurnal Sains dan Teknologi*, vol. 5, no. 1, pp. 243–246, 2023.
- [9] Rasiban Rasiban, S. Bila, and Triwahyudi Triwahyudi, "Pemilihan Rute Terbaik Dari Lokasi Calon Pelanggan Ke Terminal Akses Menggunakan Metode A*Star PT. Bahtera Anugrah Electical," *Jurnal Sains dan Teknologi*, vol. 5, no. 1, pp. 331–335, 2023.
- [10] H. Tao et al., "Training and Testing Data Division Influence on Hybrid Machine Learning Model Process: Application of River Flow Forecasting," *Complexity*, vol. 2020, pp. 1–22, Oct. 2020.
- [11] Hadirat Halawa, Okta Jaya Harmaja, Wicarda sandi Hulu, and seriani Loi, "Implementasi Algoritma K-Means Clustering Untuk Pengelompokan Penyakit Pasien Pada Puskesmas Pulo Brayon," *Jurnal Sains dan Teknologi*, vol. 5, no. 1, pp. 150–157, 2023, Accessed: Oct. 31, 2024.
- [12] "IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MEMBER BARU MENGGUNAKAN LINEAR REGRESSION PADA PT. GSI," *Jurnal*

- Tekinkom (Teknik Informasi dan Komputer), vol. 6, no. 1, pp. 118–126, 2023.
- [13] S. Usman, “Predictive Sparepart Maintenance Menggunakan Algoritma Machine Learning Extreme Gradient Boosting Regressor,” *Journal of System and Computer Engineering (JSCE)*, vol. 5, no. 2, pp. 249–258, Jul. 2024.
- [14] T. Chen and C. Guestrin, “XGBoost: a Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, vol. 13–17, pp. 785–794, 2016.
- [15] W. Li, Y. Yin, X. Quan, and H. Zhang, “Gene Expression Value Prediction Based on XGBoost Algorithm,” *Frontiers in Genetics*, vol. 10, Nov. 2019,
- [16] T. M. Khalil and Sahid Sahid, “Klasifikasi informasi hoaks pada media sosial twitter menggunakan algoritma random forest berbasis particle swarm optimization,” *Jurnal Kajian dan Terapan Matematika*, vol. 8, no. 3, pp. 199–209, 2022, Accessed: Oct. 31, 2024. [Online].

Dhiwa Aqsha, saat ini sebagai mahasiswa program studi Teknik Informatika Universitas Tarumanagara.