Seminar Nasional Sains Data 2024 (SENADA 2024)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Analisis dan Prediksi Harga Properti Rumah di Kota Surabaya dengan Algoritma Random Forest

Ajeng Puspa Wardani[1], Herlambang Awan Irawan[2], Maulidya Prastita Syah[3], Muhammad Azkiya' Akmal[4] , Naura Ulayya Nariswari[5]

[1, 2, 3, 4, 5]*Program Studi Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur*

[1]22083010040@student.upnjatim.ac.id
[2]22083010101@student.upnjatim.ac.id
[3]22083010039@student.upnjatim.ac.id
[4]22083010084@student.upnjatim.ac.id
[5]22083010034@student.upnjatim.ac.id
Corresponding author email: 22083010040@student.upnjatim.ac.id

Abstract: *Generation Z's attention to house prices in Indonesia has increased along with the surge in property prices. In September 2023, house prices rose by 2% compared to the previous year, with the House Price Index (HPI) reaching 211.9 in Q3-2023, the highest annual growth since the pandemic. However, the average income of Gen Z remains below 2.5 million rupiah per month. The Surabaya property market, the second-largest metropolitan area in Indonesia, is influenced by economic growth and urbanization. This study aims to predict house prices in Surabaya using the Random Forest algorithm optimized with GridSearchCV. The analysis shows that the main factors affecting house prices are land dimensions, followed by building dimensions, the number of bedrooms, and the number of bathrooms. The region does not significantly affect prices. Model evaluation in three scenarios shows that complete data without outlier removal provides the best performance based on Mean Absolute Error (MAE). Although Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) remain high due to outliers, maintaining complete data while finding methods to handle outliers improves prediction accuracy. This model helps prospective buyers, sellers, and property agents make smarter decisions in the Surabaya property market.*
Keywords: *House prices, Surabaya property, Random Forest, GridSearchCV, Price Prediction.*

Abstrak: Perhatian Generasi Z terhadap biaya perumahan telah melonjak secara signifikan akibat lonjakan harga yang cukup tajam. Pada September 2023, harga rumah naik 2% dibanding tahun sebelumnya, dengan Indeks harga rumah mencapai 211,9 untuk kuartal III-2023, mengalami pertumbuhan tahunan tertinggi sebesar 8,7 persen sejak pandemi. Namun, pendapatan rata-rata Gen Z masih di bawah 2,5 juta rupiah per bulan. Pasar properti Surabaya, kota terbesar kedua di Indonesia, dipengaruhi oleh pertumbuhan ekonomi dan urbanisasi. Penelitian ini bertujuan memprediksi harga rumah di Surabaya menggunakan algoritma Random Forest yang dioptimalkan dengan GridSearchCV. Hasil korelasi menunjukkan bahwa faktor utama yang mempengaruhi harga rumah adalah dimensi tanah, diikuti oleh dimensi bangunan, jumlah kamar tidur, dan jumlah kamar mandi. Wilayah tidak signifikan mempengaruhi harga. Evaluasi model dalam tiga skenario menunjukkan bahwa data utuh tanpa penghapusan outlier memberikan performa terbaik berdasarkan Mean Absolute Error (MAE). Meskipun Mean Squared Error (MSE) dan Root Mean Squared Error (RMSE) tetap tinggi karena outlier, mempertahankan data utuh sambil mencari metode untuk menangani outlier meningkatkan keakuratan prediksi. Model ini membantu calon pembeli, penjual, dan agen properti dalam membuat keputusan lebih cerdas di pasar properti Surabaya.
Kata kunci: Harga rumah, Properti Surabaya, Random Forest, GridSearchCV, Prediksi Harga.

I. PENDAHULUAN

Perhatian Generasi Z terhadap biaya perumahan telah melonjak secara signifikan dalam beberapa tahun terakhir akibat lonjakan harga yang cukup tajam. Generasi ini kini mulai memasuki dunia kerja dan menghadapi tantangan ekonomi seperti kebutuhan hunian. Harga rumah di Indonesia meningkat sebesar 2% pada bulan September 2023 dibandingkan tahun sebelumnya [1]. Indeks harga properti residensial mencapai 211,9 pada kuartal ketiga 2023, menunjukkan peningkatan tahunan tertinggi sebesar 8,7 persen sejak pandemi, menurut studi yang dilakukan oleh Housing Finance Center (HFC). Namun, pendapatan rata-rata Gen Z masih di bawah 2,5 juta rupiah per bulan pada tahun 2023

[2]. Harga rumah tipe 21 terendah berkisar antara 250 dan 450 dolar [3]. Pada kuartal III-2023, rumah ukuran besar atau tipe 70 dengan harga mulai dari Rp500 juta hingga Rp1 miliar mengalami peningkatan sebesar 12% setiap tahunnya.

Di Indonesia, khususnya di kota besar seperti Surabaya, terjadi perkembangan besar dalam beberapa tahun terakhir. Pasar properti di Surabaya, kota metropolitan terbesar kedua di Indonesia, menunjukkan dinamika yang tinggi karena berbagai faktor seperti pertumbuhan ekonomi, urbanisasi, pembangunan infrastruktur, dan peningkatan daya beli masyarakat. Mengingat banyaknya variabel yang mempengaruhi harga properti, diperlukan metode analisis yang efisien untuk memahami pola harga dan membuat proyeksi yang akurat. Untuk membuat keputusan yang tepat, calon pembeli rumah dari generasi Z membutuhkan informasi yang komprehensif dan mutakhir. Selain itu, terdapat beberapa transaksi penjualan rumah yang tidak tercatat atau tidak dipublikasikan di platform daring, yang menyulitkan calon pembeli dalam mengetahui harga yang sebenarnya.

Sebuah solusi yang dapat digunakan untuk menganalisa dan memprediksi harga properti rumah di Surabaya diperlukan untuk mengatasi masalah Generasi Z tentang harga rumah yang tidak sebanding dengan ketersediaan informasi terbaru dan akurat. Salah satu metode analisis data yaitu Random Forest, merupakan metode yang dapat memprediksi nilai data yang tidak diketahui dengan menggunakan nilai data lain yang terkait dan diketahui [4]. Selain itu, GridSearchCV memaksimalkan penggunaan algoritma ini untuk membuat model yang dihasilkan lebih akurat dan optimal. GridSearchCV membantu dalam menemukan kombinasi hyperparameter terbaik untuk model Random Forest. Hal ini dapat sangat bermanfaat bagi penjual, pembeli, dan agen properti karena dapat membantu mereka membuat keputusan yang lebih cerdas dan mengoptimalkan rencana mereka di pasar properti yang dinamis seperti Surabaya.

Studi sebelumnya menunjukkan bahwa algoritma Random Forest berguna dalam berbagai situasi prediktif. Misalnya, penelitian yang dilakukan pada tahun 2019 oleh Ilham Kurniawan, Duwi Cahya Putri Buani, Abdussomad, Widya Apriliah, dan Rizal Amegia Saputra menggunakan algoritma Random Forest menunjukkan bahwa pembelajaran mesin metode ini lebih akurat daripada algoritma seperti Support Vector Machine (SVM) dan Naive Bayes [5]. Penggunaan algoritma Random Forest dan metodologi data mining CRISP-DM memiliki kemiripan dengan penelitian kami. Di sisi lain, ada perbedaan dalam fokus penelitian kami yaitu pada analisis harga properti dengan melihat data asli, menghapus data dengan outlier, dan menghapus data dengan outlier dari kolom harga. Melalui penelitian ini diharapkan akan memberikan prediksi harga properti yang lebih akurat dan informatif, sehingga dapat membantu Generasi Z membuat keputusan pembelian rumah yang lebih baik. Penelitian ini didasari oleh studi sebelumnya yang menunjukkan bahwa algoritma Random Forest dan metodologi CRISP-DM efektif.

II. METODE PENELITIAN

Dalam penelitian ini akan dilakukan analisis korelasi untuk memahami hubungan antara berbagai variabel fitur dengan harga rumah sebagai variabel target. Metode analisis korelasi yang digunakan adalah korelasi Pearson, yang mengukur kekuatan dan arah hubungan linear antara dua variabel. Koefisien korelasi Pearson, yang bernilai antara -1 hingga 1, menunjukkan arah dan kekuatan relasi tersebut [6]. Analisis ini penting untuk mengidentifikasi variabel mana yang memiliki pengaruh paling signifikan terhadap harga rumah serta memahami bagaimana variabel - variabel tersebut saling berinteraksi. Rumus untuk menghitung koefisien korelasi Pearson adalah sebagai berikut:

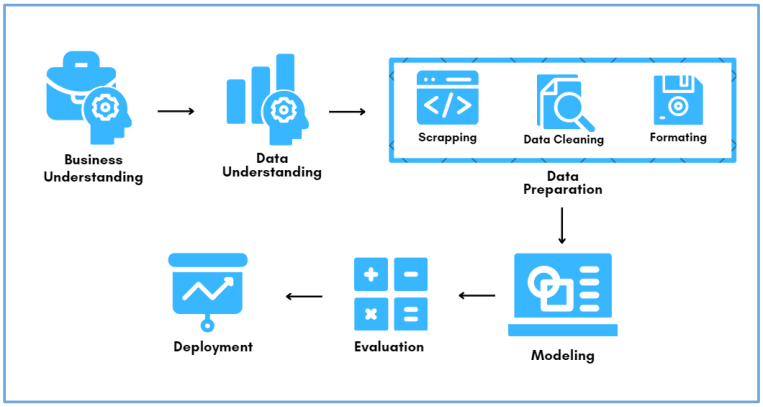
E-ISSN 2808-5841
P-ISSN 2808-7283

$$r = \frac{\sum_{i=1}^{n} (Xi - \underline{X})(Yi - \underline{Y})}{\sqrt{\sum_{i=1}^{n} (Xi - \underline{X})^2} \sqrt{\sum_{i=1}^{n} (Yi - \underline{Y})^2}}$$

Dimana:
r: koefisien korelasi Pearson.
Xi: nilai individu dari variabel X.
Yi: nilai individu dari variabel Y.
$\underline{X}$: rata-rata dari nilai variabel X.
$\underline{Y}$: rata-rata dari nilai variabel Y.
n: jumlah pasangan data.

Algoritma yang digunakan dalam penelitian ini adalah algoritma Random Forest, yang dirancang untuk menangani data kontinu. Untuk memastikan bahwa model dapat diuji pada data yang belum pernah dilihat sebelumnya, dataset yang telah dipersiapkan sebelumnya dibagi menjadi data latih dan data uji [7]. Model ini dioptimalkan menggunakan GridSearchCV, yaitu teknik yang digunakan untuk mencari kombinasi hyperparameter terbaik bagi algoritma dengan cara mengevaluasi performa model pada berbagai kombinasi hyperparameter yang telah ditentukan [8] . Proses ini memastikan model yang paling optimal dipilih berdasarkan kinerja terbaiknya pada data uji, sehingga memberikan hasil prediksi yang lebih akurat.

Dalam penelitian ini, kami menerapkan tiga skenario untuk membandingkan hasil pemodelan. Skenario pertama menggunakan data lengkap tanpa penghapusan outlier. Skenario kedua menggunakan data yang telah dihapus baris yang mengandung outlier di setidaknya satu kolom. Terakhir, skenario ketiga menggunakan data yang dihapus baris yang memiliki outlier di kolom harga, karena kolom harga merupakan target prediksi. Dengan menerapkan ketiga skenario ini, kami akan mengevaluasi pengaruh keberadaan outlier terhadap performa model dan menentukan strategi terbaik untuk menangani outlier dalam konteks prediksi harga rumah. Hasil analisis dari ketiga skenario ini akan dijadikan acuan untuk pemilihan skenario terbaik dalam memprediksi harga rumah.



Gambar 1. Desain Sistem CRISP-DM

Data sekunder yang digunakan dalam penelitian ini diperoleh melalui teknik scraping dari situs web 'Db Real Estate', sebuah platform jual beli properti dari seluruh Indonesia. Proses scraping dilakukan dengan menetapkan filter atau parameter pada kolom pencarian di situs tersebut, yaitu lokasi. Lima lokasi yang dipilih untuk scraping data adalah Surabaya Timur, Surabaya Selatan, Surabaya Tengah, Surabaya Utara, dan Surabaya Barat. Data yang diperoleh dari scraping lima lokasi ini

mencakup 8.940 baris, yang terdiri dari kolom Agen Properti, Judul, Harga, Luas Tanah, Luas Bangunan, Jumlah Kamar Tidur, Jumlah Kamar Mandi, dan Lokasi.

Pada tahapan business understanding, tujuan utama adalah memahami kebutuhan masyarakat terhadap informasi harga properti yang akurat di Kota Surabaya. Peningkatan harga properti yang tajam, dikombinasikan dengan pendapatan rata-rata masyarakat yang rendah, membuat akses terhadap informasi harga rumah menjadi sangat penting. Oleh karena itu, solusi yang diajukan adalah membangun model prediktif menggunakan algoritma Random Forest yang mampu memberikan estimasi harga properti secara akurat berdasarkan faktor-faktor yang berpengaruh seperti lokasi, luas tanah, jumlah kamar tidur, dan jumlah kamar mandi.

Tahap data understanding melibatkan pengumpulan data dan memahami karakteristik data yang akan digunakan. Jumlah total baris data yang diperoleh adalah 8.940, dengan persebaran harga rumah berkisar dari 45 juta hingga yang paling mahal mencapai 249 miliar rupiah. Tahap data preparation atau praproses data mencakup dua tahap: cleaning dan formatting. Data cleaning melibatkan penggantian nilai median setiap kolom pada data yang memiliki missing value, penghapusan tanda koma, penghapusan satuan, penjumlahan nilai, dan penggabungan data. Formatting atau transformasi data adalah proses penataan dan pengorganisasian data ke dalam format yang konsisten dan terstruktur agar mudah dianalisis dan diinterpretasi.

Pada tahap pemodelan, algoritma Random Forest diterapkan untuk memodelkan data menggunakan CRISP-DM (Cross Industry Standard Process for Data Mining). Dataset yang telah disiapkan sebelumnya dibagi menjadi data pelatihan dan data pengujian. Dengan menggunakan Random Forest, model dapat dikembangkan untuk mengidentifikasi pola dan keterkaitan antara fitur-fitur properti dengan harga properti. Model ini diharapkan dapat memberikan estimasi harga yang akurat dan membantu calon pembeli, penjual, serta makelar dalam membuat keputusan yang lebih baik dan lebih terinformasi di pasar properti Kota Surabaya.

Selanjutnya, evaluasi model dilakukan menggunakan Mean Absolute Error (MAE), Mean Squared Error (MSE), dan Root Mean Squared Error (RMSE) untuk mengukur seberapa baik model Random Forest dalam meramal harga properti. Pada tahap penerapan, model yang telah dievaluasi dan dioptimalkan digunakan dalam lingkungan operasional. Tahap ini bertujuan untuk memastikan bahwa model prediksi harga properti dapat diakses dan digunakan oleh calon pembeli, penjual, dan makelar. Implementasi penelitian ini berupa visualisasi yang menyajikan hasil analisis dan prediksi model secara jelas dan mudah dipahami, sehingga mendukung pengambilan keputusan yang lebih baik di pasar properti Kota Surabaya.

III. HASIL DAN PEMBAHASAN

3.1 Analisis Korelasi

Pada skenario pertama yang dapat dilihat pada Gambar 1, data digunakan sepenuhnya tanpa menghapus outlier. Korelasi antara harga dan dimensi tanah (0,895) serta dimensi bangunan (0,893) menunjukkan bahwa semakin besar tanah dan bangunan, semakin tinggi harga properti. Korelasi dengan jumlah kamar tidur (0,688) dan jumlah kamar mandi (0,675) juga menunjukkan pengaruh positif yang signifikan terhadap harga properti. Namun, lokasi di Surabaya Timur memiliki korelasi negatif (-0,056), menunjukkan harga properti di sana sedikit lebih rendah. Lokasi lain seperti Surabaya Selatan, Utara, Barat, dan Tengah memiliki pengaruh yang sangat lemah terhadap harga properti, dengan korelasi sangat rendah dan hampir mendekati nol.

E-ISSN 2808-5841
P-ISSN 2808-7283

	Harga	Dimensi Tanah	Dimensi Bangunan	Jumlah Kamar Tidur	Jumlah Kamar Mandi	Surabaya Timur	Surabaya Selatan	Surabaya Utara	Surabaya Barat	Surabaya Tengah
Harga	1.000000	0.895270	0.893233	0.688004	0.675315	-0.056216	0.014560	0.006170	0.046307	0.008451
Dimensi Tanah	0.895270	1.000000	0.940900	0.772707	0.544613	-0.178274	0.046174	0.019566	0.146848	0.026800
Dimensi Bangunan	0.893233	0.940900	1.000000	0.688569	0.642531	-0.105425	0.027306	0.011570	0.086840	0.015849
Jumlah Kamar Tidur	0.688004	0.772707	0.688569	1.000000	0.642029	-0.063585	0.016469	0.006979	0.052377	0.009559
Jumlah Kamar Mandi	0.675315	0.544613	0.642531	0.642029	1.000000	0.109926	-0.028472	-0.012064	-0.090549	-0.016525
Surabaya Timur	-0.056216	-0.178274	-0.105425	-0.063585	0.109926	1.000000	-0.259006	-0.109750	-0.823720	-0.150332
Surabaya Selatan	0.014560	0.046174	0.027306	0.016469	-0.028472	-0.259006	1.000000	-0.028047	-0.210504	-0.038418
Surabaya Utara	0.006170	0.019566	0.011570	0.006979	-0.012064	-0.109750	-0.028047	1.000000	-0.089198	-0.016279
Surabaya Barat	0.046307	0.146848	0.086840	0.052377	-0.090549	-0.823720	-0.210504	-0.089198	1.000000	-0.122181
Surabaya Tengah	0.008451	0.026800	0.015849	0.009559	-0.016525	-0.150332	-0.038418	-0.016279	-0.122181	1.000000

Gambar 2. Gambar Korelasi Skenario Pertama

	Harga	Dimensi Tanah	Dimensi Bangunan	Jumlah Kamar Tidur	Jumlah Kamar Mandi	Surabaya Timur	Surabaya Selatan	Surabaya Utara	Surabaya Barat	Surabaya Tengah
Harga	1.000000	0.826302	0.708654	0.810403	0.466670	-0.199892	NaN	NaN	0.199892	NaN
Dimensi Tanah	0.826302	1.000000	0.873538	0.776509	0.264701	-0.437711	NaN	NaN	0.437711	NaN
Dimensi Bangunan	0.708654	0.873538	1.000000	0.735918	0.421306	-0.411796	NaN	NaN	0.411796	NaN
Jumlah Kamar Tidur	0.810403	0.776509	0.735918	1.000000	0.587785	-0.146764	NaN	NaN	0.146764	NaN
Jumlah Kamar Mandi	0.466670	0.264701	0.421306	0.587785	1.000000	0.099701	NaN	NaN	-0.099701	NaN
Surabaya Timur	-0.199892	-0.437711	-0.411796	-0.146764	0.099701	1.000000	NaN	NaN	-1.000000	NaN
Surabaya Selatan	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Surabaya Utara	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Surabaya Barat	0.199892	0.437711	0.411796	0.146764	-0.099701	-1.000000	NaN	NaN	1.000000	NaN
Surabaya Tengah	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Gambar 3. Gambar menghapus outlier pada semua kolom

Pada skenario kedua, outlier dihapus jika berada di setidaknya satu kolom. Setelah menghapus outlier, korelasi antara harga dan dimensi tanah (0,826) serta dimensi bangunan (0,709) masih signifikan tetapi sedikit lebih rendah dibandingkan skenario pertama. Jumlah kamar tidur memiliki korelasi yang sangat tinggi (0,810) dengan harga, menunjukkan pengaruh yang kuat, sedangkan jumlah kamar mandi menunjukkan korelasi moderat (0,467). Lokasi di Surabaya Timur menunjukkan korelasi negatif (-0,200), sedangkan Surabaya Barat menunjukkan korelasi positif (0,1998). Data untuk lokasi lain tidak tersedia (NaN). Karena dalam dataset ini lokasi Surabaya Timur dan Surabaya Barat memiliki jumlah baris terbanyak, sistem menganggap lokasi selain dua lokasi tersebut sebagai outlier sehingga dihapus oleh sistem. Maka dari itu, hasil korelasinya menunjukkan NaN.

	Harga	Dimensi Tanah	Dimensi Bangunan	Jumlah Kamar Tidur	Jumlah Kamar Mandi	Surabaya Timur	Surabaya Selatan	Surabaya Utara	Surabaya Barat	Surabaya Tengah
Harga	1.000000	0.831475	0.703149	0.803251	0.454348	-0.197451	0.051141	0.021670	0.162644	0.029683
Dimensi Tanah	0.831475	1.000000	0.875413	0.784940	0.266556	-0.428619	0.111015	0.047041	0.353062	0.064435
Dimensi Bangunan	0.703149	0.875413	1.000000	0.741582	0.420805	-0.406583	0.105307	0.044623	0.334911	0.061123
Jumlah Kamar Tidur	0.803251	0.784940	0.741582	1.000000	0.586238	-0.147152	0.038113	0.016150	0.121212	0.022122
Jumlah Kamar Mandi	0.454348	0.266556	0.420805	0.586238	1.000000	0.103139	-0.026714	-0.011320	-0.084957	-0.015505
Surabaya Timur	-0.197451	-0.428619	-0.406583	-0.147152	0.103139	1.000000	-0.259006	-0.109750	-0.823720	-0.150332
Surabaya Selatan	0.051141	0.111015	0.105307	0.038113	-0.026714	-0.259006	1.000000	-0.028047	-0.210504	-0.038418
Surabaya Utara	0.021670	0.047041	0.044623	0.016150	-0.011320	-0.109750	-0.028047	1.000000	-0.089198	-0.016279
Surabaya Barat	0.162644	0.353062	0.334911	0.121212	-0.084957	-0.823720	-0.210504	-0.089198	1.000000	-0.122181
Surabaya Tengah	0.029683	0.064435	0.061123	0.022122	-0.015505	-0.150332	-0.038418	-0.016279	-0.122181	1.000000

Gambar 3. Gambar Korelasi Penghapusan Outlier Pada Harga

Pada skenario ketiga, outlier dihapus hanya pada kolom harga. Hal ini dilakukan karena harga merupakan kolom target, sehingga outlier pada kolom ini dianggap cukup mempengaruhi hasil korelasi. Setelah penghapusan outlier pada kolom harga, korelasi antara harga dan dimensi tanah (0,831) serta dimensi bangunan (0,703) masih signifikan dan mendekati hasil skenario pertama. Jumlah kamar tidur menunjukkan korelasi yang sangat tinggi (0,803), sedangkan jumlah kamar mandi memiliki korelasi moderat (0,454). Lokasi di Surabaya Timur memiliki korelasi negatif (-0,197), menunjukkan harga properti di sana lebih rendah. Lokasi di Surabaya Selatan, Utara, Barat, dan Tengah memiliki korelasi yang sangat lemah atau hampir tidak berpengaruh, hampir sama dengan korelasi pada skenario pertama.

Dari ketiga skenario, skenario 3 tampaknya merupakan analisis korelasi yang terbaik. Penghapusan outlier hanya pada kolom harga mengurangi pengaruh data ekstrim tanpa menghilangkan data penting dari kolom lain, sehingga memberikan gambaran yang lebih akurat dan bersih tentang hubungan antar variabel. Sebaliknya, penghapusan outlier setidaknya di satu kolom pada skenario 2 menurunkan korelasi dan menyebabkan hilangnya data untuk beberapa lokasi, sehingga tidak semua korelasi bisa dihitung. Data penuh pada skenario 1 memberikan gambaran korelasi yang kuat antara variabel-variabel properti dan harga tanpa ada pengaruh dari penghapusan data, tetapi ini mungkin termasuk outlier yang dapat mempengaruhi keakuratan model.

3.2. Hasil Prediksi

Hasil prediksi yang dapat dilihat pada gambar 4a hingga 4c dihasilkan oleh model yang menunjukkan bahwa beberapa prediksi sangat akurat dan sesuai dengan harga aktual. Sebagai contoh, ada banyak kasus di mana harga prediksi hampir sama atau bahkan identik dengan harga aktual, menunjukkan bahwa model ini mampu menangkap pola dan hubungan dalam data dengan baik. Namun, ada juga beberapa prediksi yang menunjukkan perbedaan selisih harga. Perbedaan ini cukup beragam, dengan beberapa hasil prediksi hanya berselisih sedikit dari harga aktual, sementara yang lain menunjukkan selisih yang cukup besar. Secara keseluruhan, hasil prediksi lebih banyak yang akurat dengan harga aktual. Namun, di antara ketiga skenario tersebut, hasil prediksi yang paling jauh terdapat pada skenario kedua dengan indeks ke-7100, menunjukkan selisih kurang lebih 300.000.000. Perlu diperhatikan bahwa hasil yang ditampilkan hanya beberapa, bukan keseluruhan hasil prediksi, sehingga evaluasi model lebih lanjut perlu dilakukan.

E-ISSN 2808-5841
P-ISSN 2808-7283

	Harga Aktual	Harga Prediksi
7135	3.950000e+09	3.950000e+09
3309	5.950000e+09	5.950000e+09
2199	5.950000e+09	5.950000e+09
6109	9.500000e+09	9.146140e+09
6730	3.950000e+09	3.950000e+09
...	...	...
8452	1.750000e+09	1.750000e+09
5627	3.600000e+09	3.600000e+09
432	7.750000e+08	7.750000e+08
7362	9.500000e+09	9.500000e+09
5649	4.500000e+06	4.500000e+06

2682 rows × 2 columns

Gambar 4a. Hasil Prediksi Skenario 1

	Harga Aktual	Harga Prediksi
3870	1.875000e+09	1.875000e+09
421	1.600000e+09	1.600000e+09
8157	9.500000e+09	9.500000e+09
6249	4.500000e+06	4.500000e+06
25	3.700000e+09	3.700000e+09
...	...	...
6446	8.800000e+09	8.800000e+09
7100	8.800000e+09	9.173945e+09
5980	3.950000e+09	3.950000e+09
1498	6.750000e+08	6.750000e+08
196	1.600000e+09	1.600000e+09

2264 rows × 2 columns

Gambar 4b. Hasil Prediksi Skenario 2

	Harga Aktual	Harga Prediksi
6827	3.600000e+09	3.600000e+09
5574	4.500000e+06	4.500000e+06
4586	8.800000e+09	8.800000e+09
1007	9.500000e+09	9.500000e+09
5749	9.500000e+09	9.143192e+09
...	...	...
6718	1.200000e+09	1.200000e+09
3952	4.500000e+09	4.500000e+09
5876	8.800000e+09	8.800000e+09
5762	3.600000e+09	3.600000e+09
4732	1.750000e+09	1.750000e+09

2504 rows × 2 columns

Gambar 4c. Hasil Prediksi Skenario 3

3.3 Evaluasi Model

Tabel 1. Hasil Performa Klasifikasi Data Mining

Skenario	MAE	MSE	RMSE
Data Lengkap	20.923.832,4	7398102822909874.0	86.012.224,8
Menghapus Seluruh Outlier	22.947.371,2	8146545649210365.0	86.012.224,8
Menghapus Outlier kolom Harga	23.434.187,9	8249797387578456.0	86.012.224,8

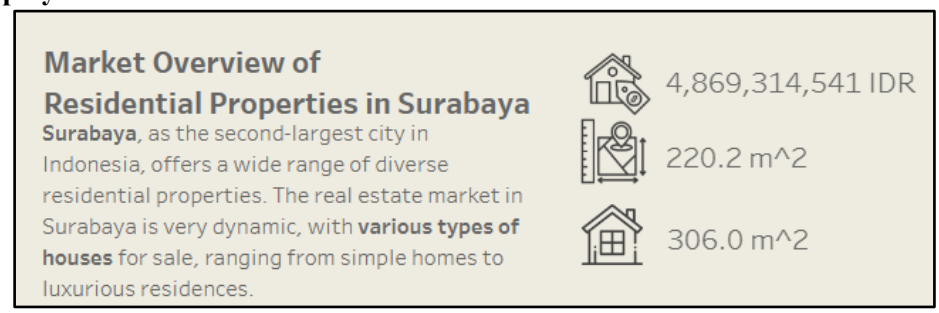
Pada skenario pertama, model dilatih dan dievaluasi menggunakan data lengkap tanpa menghapus outlier. Hasil evaluasi menunjukkan bahwa model memiliki MAE yang relatif rendah (20.923.832,4), menunjukkan bahwa rata-rata deviasi absolut antara prediksi dan nilai aktual tergolong cukup minimal. MSE juga cukup besar, menunjukkan bahwa ada beberapa kesalahan prediksi yang cukup besar, meskipun secara umum model tampak mampu menangkap pola dalam data dengan baik. RMSE yang tinggi (86.012.224,8) juga menunjukkan adanya beberapa prediksi dengan kesalahan yang signifikan.

Pada skenario kedua, outlier dihapus dari seluruh kolom. Hasil evaluasi menunjukkan peningkatan MAE menjadi 22.947.371,2, yang berarti rata-rata kesalahan absolut antara prediksi dan nilai aktual sedikit lebih besar dibandingkan dengan skenario pertama. MSE juga meningkat, menunjukkan bahwa model mungkin mengalami kesulitan dalam memprediksi nilai tanpa outlier tertentu. RMSE tetap sama, menunjukkan bahwa skala kesalahan signifikan tetap ada meskipun outlier telah dihapus. Hal ini mungkin disebabkan oleh hilangnya banyak data, khususnya data dari lokasi selain Surabaya Timur dan Surabaya Barat, yang terhapus karena dianggap outlier oleh sistem.

Pada skenario ketiga, outlier dihapus hanya pada kolom harga. Hasil evaluasi menunjukkan bahwa MAE meningkat menjadi 23.434.187,9, menunjukkan rata-rata kesalahan absolut yang lebih

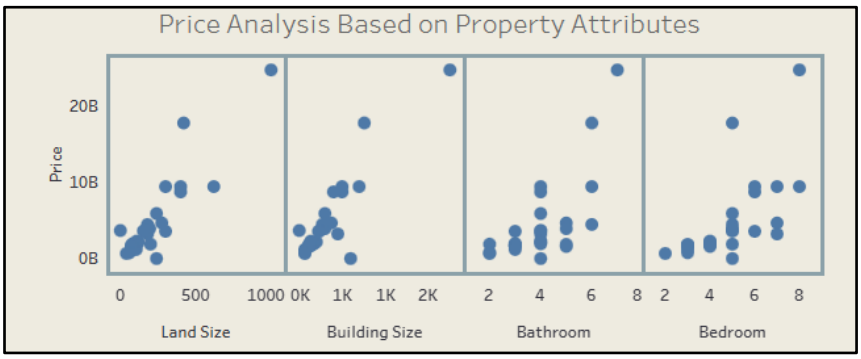
besar dibandingkan kedua skenario sebelumnya. MSE juga meningkat, yang menunjukkan bahwa model masih mengalami kesulitan dalam memprediksi nilai tanpa outlier pada kolom harga. RMSE tetap sama, menunjukkan bahwa kesalahan prediksi yang besar masih ada meskipun outlier pada kolom harga telah dihapus.

3.4. Deployment



Gambar 5. Visualisasi Tableau

Dari hasil deployment yang merupakan sebuah visualisasi yang kami buat menggunakan *tool* Tableau untuk mempermudah calon pembeli memahami aspek pasar properti di kota Surabaya. Sebagai kota metropolitan terbesar kedua di Indonesia, Surabaya menyediakan beragam jenis properti residensial dengan nilai pasar yang sangat signifikan yaitu mencapai Rp. 4.869.314.541. Luas rata-rata dari properti residensial di kota ini adalah 220,2 meter persegi, sementara luas rata-rata lahannya mencapai 306,0 meter persegi. Visualisasi tersebut menunjukkan scatter plot yang menghubungkan antara harga properti dengan ukuran lahan, ukuran bangunan, jumlah kamar mandi, dan jumlah kamar tidur.

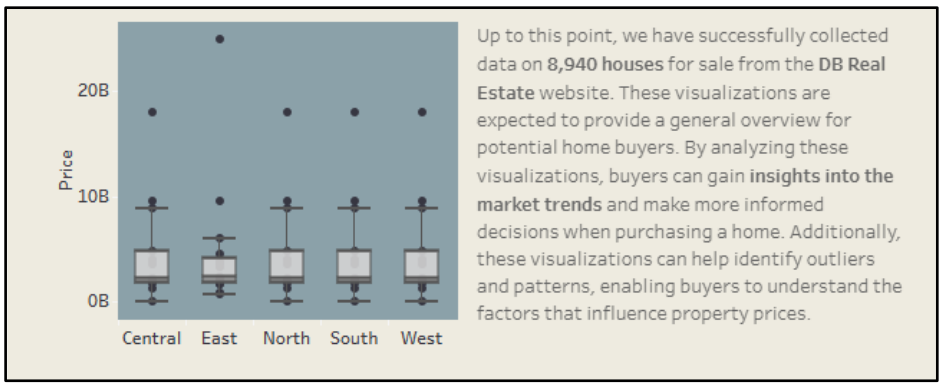


Gambar 6. Grafik Analisis Atribut Properti

Dari grafik pertama, terlihat bahwa harga properti cenderung meningkat dengan bertambahnya ukuran lahan. Properti dengan ukuran lahan lebih besar cenderung memiliki harga yang lebih tinggi, meskipun terdapat beberapa variasi harga di berbagai ukuran lahan. Grafik kedua menunjukkan bahwa ada korelasi antara ukuran bangunan dan harga properti. Properti dengan ukuran bangunan lebih besar (dalam kisaran 1K-2K meter persegi) memiliki harga yang lebih tinggi, meskipun beberapa properti dengan ukuran bangunan lebih kecil juga memiliki harga yang tinggi.

Grafik ketiga menunjukkan bahwa jumlah kamar mandi juga mempengaruhi harga properti. Properti dengan lebih banyak kamar mandi cenderung memiliki harga lebih tinggi, namun terdapat variasi signifikan dalam harga properti dengan jumlah kamar mandi yang sama. Grafik keempat memperlihatkan bahwa properti dengan lebih banyak kamar tidur cenderung memiliki harga lebih

tinggi. Namun, seperti halnya dengan jumlah kamar mandi, terdapat variasi harga yang signifikan pada properti dengan jumlah kamar tidur yang sama.



Gambar 7. Grafik Boxplot Data Harga Properti

Boxplot tersebut menyajikan data harga properti di lima wilayah Surabaya yang mencakup 8.940 rumah yang dijual pada bagian Surabaya Central, East, North, South, dan West. Harga properti di wilayah Central menunjukkan variasi yang cukup besar, dengan beberapa properti berharga sangat tinggi di atas 20 milyar IDR. Namun, sebagian besar harga properti berada di kisaran lebih rendah. Di wilayah East, harga properti lebih terkonsentrasi di kisaran bawah, dengan beberapa outlier di kisaran harga tinggi. Untuk wilayah North, South, dan West menunjukkan pola serupa dengan East, dengan harga properti mayoritas berada di kisaran bawah dan beberapa outlier.

IV. KESIMPULAN

Berdasarkan analisis yang dilakukan, faktor yang paling mempengaruhi harga rumah adalah dimensi tanah, diikuti oleh dimensi bangunan, jumlah kamar tidur, dan jumlah kamar mandi. Hasil analisis menunjukkan bahwa wilayah tempat rumah dijual tidak signifikan mempengaruhi harga rumah. Namun, faktor lain seperti lokasi di dalam atau di luar perumahan, kualitas konstruksi, umur bangunan, dan kondisi lingkungan sekitar juga perlu dipertimbangkan. Fasilitas umum seperti akses transportasi, kedekatan dengan sekolah, pusat perbelanjaan, dan rumah sakit juga berpengaruh terhadap harga rumah. Dalam pemilihan algoritma untuk memprediksi harga rumah, algoritma Random Forest yang dioptimalkan dengan GridSearchCV menunjukkan hasil terbaik dengan nilai error terkecil. Prediksi model ini sering kali akurat, meski ada beberapa perbedaan selisih harga antara prediksi dan harga aktual.

Evaluasi model dalam tiga skenario berbeda menunjukkan bahwa skenario 1, di mana data diolah utuh tanpa menghilangkan outlier, memberikan performa terbaik berdasarkan nilai MAE. Meskipun penghapusan outlier diharapkan memperbaiki data, model dengan data lengkap tetap lebih baik dalam kasus ini. Namun, meskipun MAE lebih rendah, nilai MSE dan RMSE yang tinggi menunjukkan adanya beberapa prediksi dengan kesalahan besar. Ini mungkin disebabkan oleh outlier yang mempengaruhi hasil prediksi secara signifikan. Dalam skenario ini, pendekatan terbaik adalah mempertahankan data utuh sambil mencari metode untuk menangani outlier, seperti transformasi data atau penggunaan algoritma yang lebih robust terhadap outlier, guna meningkatkan keakuratan prediksi.

REFERENSI

1. "Indonesia | Pertumbuhan Harga Rumah | 2003 – 2024 | Indikator Ekonomi." n.d. CEIC. Accessed June 16, 2024." Accessed: May 31, 2024. [Online]. Available: https://www.ceicdata.com/id/indicator/indonesia/house-prices-growth
2. "Ini Dia Rata-rata Pendapatan Gen Z, Ada yang Di Atas 10 Juta?" Accessed: May 31, 2024. [Online]. Available: https://data.goodstats.id/statistic/ini-dia-rata-rata-pendapatangen-z-ada-yang-di-atas-10-juta-QdAGg
3. "Harga Rumah Melambung Tinggi, Gen Z Susah Memiliki Rumah." Accessed: May 31, 2024. [Online]. Available: https://perkim.id/perumahan/harga-rumah-melambung-tinggigen-z-susah-memiliki-rumah/
4. E. S. Lestari and I. Astuti, "Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat," J. Ilm. FIFO, vol. 14, no. 2, p. 131, Nov. 2022, doi: 10.22441/fifo.2022.v14i2.003.
5. I. Kurniawan, D. C. P. Buani, A. Abdussomad, W. Apriliah, and R. A. Saputra, "Implementasi Algoritma Random Forest Untuk Menentukan Penerima Bantuan Raskin," J. Teknol. Inf. Dan Ilmu Komput., vol. 10, no. 2, pp. 421–428, Apr. 2023, doi: 10.25126/jtiik.20231026225.
6. "Korelasi Pearson." Accessed: May 27, 2024. [Online]. Available: https://ss.mipa.ub.ac.id/korelasi-pearson/
7. M. N. Tentua and S. Sumarmi, "Prediksi Promosi Pegawai Menggunakan Metode Extremely Randomized Trees," vol. 12, no. 2, 2023.
8. A. W. Ishlah, S. Sudarno, and P. Kartikasari, "IMPLEMENTASI GRIDSEARCHCV PADA SUPPORT VECTOR REGRESSION (SVR) UNTUK PERAMALAN HARGA SAHAM," J. Gaussian, vol. 12, no. 2, pp. 276–286, Jul. 2023, doi: 10.14710/j.gauss.12.2.276-286