(2025), 3 (3): 411–420

Scientica
Jurnal Ilmiah Sain dan Teknologi

3021-8209

IMPLEMENTASI MODEL REGRESI LINEAR UNTUK PREDIKSI HARGA RUMAH

Alfenda Maulana Akbar [1], Bagus Deny Kurnianto [2], Randi Dwi Pradiyanto [3], Azka Nuskhatul Akwan [4], Hakiki Faturrahman [5], Anna Dina Khalifia [6]
Informatika, Universitas Teknologi Yogyakarta
Jl. Siliwangi (Ringroad Utara), Jombor, Sleman, D.I. Yogyakarta 55285
alfendamaulana15@gmail.com

Abstract (English)
This research was conducted to predict real estate prices using the linear regression method. The data used provided information on population mediation, house building age, average living room, average bed, local area population, average occupation rate, latitude and longitude geographical coordinates, and house selling price. From the results of data analysis using linear regression techniques, it was found that the linear regression model can explain about 80% of the variation in house selling prices. Independent variables such as population mediation, age of the house, and average living space have a significant contribution to the selling price of the house. The coefficient of determination (R^2) of 0.50 indicates the accuracy of the model in predicting house selling prices.The results of this study indicate that the linear regression model is an effective analytical tool in predicting real estate prices. Therefore, the results of this analysis can be a reference for property developers and investors in making more appropriate investment strategies. In addition, evaluation metrics such as Mean Squared Error (MSE), R-Square (R^2), and Mean Absolute Error (MAE) are used to assess the accuracy of the model.

Article History
Submitted: 2 January 2025
Accepted: 11 January 2025
Published: 12 January 2025

Key Words
Prediction, Price, Linear Regression, House

Abstrak (Indonesia)
Penelitian ini dilakukan untuk melakukan prediksi harga real estate menggunakan metode regresi linear. Data yang digunakan menyediakan informasi tentang mediasi penduduk, usia bangunan rumah, ruang tamu rata-rata, tempat tidur rata-rata, populasi daerah setempat, tingkat pendudukan rata-rata, koordinat geografis latitude dan longitude, serta harga jual rumah. Dari hasil analisis data menggunakan teknik regresi linear, didapatkan bahwa model regresi linear dapat menjelaskan sekitar 80% variasi harga jual rumah. Variabel-variabel independen seperti mediasi penduduk, usia bangunan rumah, dan ruang tamu rata-rata memiliki kontribusi signifikan terhadap harga jual rumah. Koefisien determinasi (R^2) sebesar 0,50 menunjukkan ketepatan model dalam memprediksi harga jual rumah. Hasil penelitian ini menunjukkan bahwa model regresi linear merupakan alat analitis yang efektif dalam memprediksi harga real estate. Oleh karena itu, hasil analisis ini dapat menjadi acuan bagi para pengembang properti maupun investor dalam membuat strategi investasi yang lebih tepat. Selain itu, metrik evaluasi seperti Mean Squared Error (MSE), R-Square (R^2), dan Mean Absolute Error (MAE) digunakan untuk menilai keakuratan model.

Sejarah Artikel
Submitted: 2 January 2025
Accepted: 11 January 2025
Published: 12 January 2025

Kata Kunci
Prediksi, Harga, Regresi Linear, Rumah

1. PENDAHULUAN

Pasar real estate terus mengalami perubahan dinamis yang dipengaruhi oleh berbagai faktor ekonomi, sosial, dan lingkungan. Memahami bagaimana faktor-faktor ini mempengaruhi harga properti menjadi sangat penting bagi pengembang, investor, dan pembeli. Model regresi linear menawarkan pendekatan yang efektif untuk menganalisis hubungan antara variabel independen dan harga jual rumah.Dalam penelitian ini, kami menggunakan dataset yang mencakup informasi tentang pendapatan median, usia rumah, jumlah kamar, dan berbagai variabel demografis lainnya. Kami bertujuan untuk mengidentifikasi pola-pola yang dapat membantu dalam memprediksi harga real estate secara akurat. Dengan menerapkan model regresi linear, kami berharap dapat memberikan wawasan yang berguna bagi para pelaku industri dalam merumuskan strategi investasi dan pengembangan properti yang lebih baik.Dengan demikian, penelitian ini tidak hanya

berkontribusi pada literatur akademik tetapi juga memberikan panduan praktis bagi para profesional di sektor real estate.

2. TINJAUAN PUSTAKA

2.1. Rumah

Rumah merupakan sebuah bangunan yang dirancang untuk digunakan sebagai tempat tinggal bagi individu atau keluarga, dan sering kali mencakup berbagai fasilitas seperti ruang tamu, kamar tidur, dapur, dan kamar mandi. Dalam konteks yang lebih luas, rumah tidak hanya berfungsi sebagai tempat berlindung dari cuaca dan ancaman luar, tetapi juga sebagai ruang yang mendukung interaksi sosial, kenyamanan, dan keamanan. Selain itu, rumah juga dapat menjadi aset investasi yang signifikan bagi pemiliknya, karena nilai properti dapat meningkat seiring waktu tergantung pada lokasi, kondisi pasar, dan faktor-faktor lainnya.

Selain itu, rumah juga memiliki dimensi sosial dan psikologis yang mendukung interaksi antaranggota keluarga serta memberikan rasa aman dan nyaman. Faktor-faktor seperti ketersediaan fasilitas umum, keamanan lingkungan, dan akses transportasi sering kali menjadi pertimbangan utama dalam menentukan nilai rumah di pasar. Oleh karena itu, analisis harga rumah menjadi salah satu elemen kunci dalam penelitian real estate.

Dalam perspektif ekonomi, rumah dianggap sebagai barang tahan lama yang nilainya dapat meningkat seiring waktu, terutama di kawasan yang berkembang. Namun, perubahan pasar, kebijakan pemerintah, dan kondisi makroekonomi dapat memengaruhi harga rumah secara signifikan.

2.2. Prediksi

Prediksi adalah proses memperkirakan atau meramalkan suatu kejadian atau hasil di masa depan berdasarkan data dan informasi yang tersedia saat ini. Dalam konteks analisis data dan statistik, prediksi sering kali melibatkan penggunaan model matematis atau algoritma untuk menganalisis pola-pola dalam data historis. Proses ini memungkinkan peneliti atau analis untuk membuat keputusan yang lebih baik dengan mempertimbangkan kemungkinan hasil yang berbeda.

Dalam bidang ekonomi dan bisnis, prediksi sangat penting untuk merencanakan strategi dan mengambil keputusan yang tepat. Misalnya, perusahaan dapat menggunakan teknik prediksi untuk memperkirakan permintaan produk di masa depan, sehingga mereka dapat mengatur produksi dan persediaan dengan lebih efisien. Selain itu, dalam konteks real estate, prediksi harga properti membantu investor memahami tren pasar dan membuat keputusan investasi yang lebih cerdas.

Prediksi juga dapat diterapkan dalam berbagai bidang lain seperti meteorologi untuk meramalkan cuaca, ilmu kesehatan untuk memprediksi penyebaran penyakit, serta dalam ilmu sosial untuk memahami perilaku manusia. Dengan kemajuan teknologi dan metode analisis data yang semakin canggih, kemampuan untuk melakukan prediksi yang akurat semakin meningkat. Namun, penting untuk diingat bahwa semua prediksi memiliki tingkat ketidakpastian yang harus dipertimbangkan dalam pengambilan keputusan.

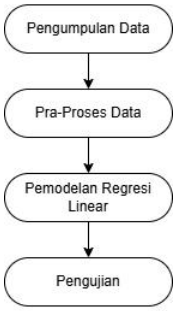
2.3. Regresi Linear

Regresi linear adalah metode statistik yang digunakan untuk menganalisis hubungan antara satu variabel dependen dengan satu atau lebih variabel independen. Metode ini bertujuan untuk menemukan garis terbaik (line of best fit) yang dapat menggambarkan hubungan antara variabel-variabel tersebut dalam bentuk persamaan linear. Dalam regresi linear sederhana, model hanya melibatkan satu variabel independen sedangkan regresi linear berganda melibatkan dua atau lebih variabel independen.

Proses regresi linear dimulai dengan pengumpulan data historis mengenai variabel-variabel terkait. Setelah itu, analisis dilakukan untuk menentukan koefisien regresi yang menunjukkan seberapa besar pengaruh setiap variabel independen terhadap variabel dependen. Hasil dari analisis ini biasanya disajikan dalam bentuk persamaan matematis yang memungkinkan prediksi nilai variabel dependen berdasarkan nilai-nilai variabel independen.

Regresi linear banyak digunakan dalam berbagai bidang seperti ekonomi, ilmu sosial, kesehatan masyarakat, dan teknik. Dalam konteks real estate, regresi linear sering diterapkan untuk memprediksi harga properti berdasarkan faktor-faktor seperti lokasi, ukuran bangunan, usia bangunan, dan fasilitas lainnya. Dengan menggunakan regresi linear, para peneliti dapat memberikan wawasan berharga mengenai dinamika pasar real estate serta membantu investor dalam pengambilan keputusan investasi yang lebih informasional.

3. METODE PENELITIAN



Gambar 1. Diagram alur

Berdasarkan Diagram alur pada Gambar 1, penelitian dimulai dari pengumpulan data terlebih dahulu, kami melakukan pencarian data di website Kaggle.com. Kemudian melakukan pra-proses data dengan melakukan pengecekan missing value untuk memastikan data sudah bersih. Selanjutnya melakukan pemodelan regresi linear, di fase ini data akan dibagi 2, yaitu data testing dan data training. Tahap terakhir yaitu melakukan pengujian untuk mengevaluasi tingkat akurasi prediksi dan kinerja dari model regresi linear yang sudah dilakukan.

3.1. Sumber Data

Sumber data dari dataset yang digunakan dalam penelitian ini berasal dari Kaggle yang berjudul " House price prediction boston dataset". Dataset ini mencakup informasi tentang berbagai variabel yang berhubungan dengan karakteristik rumah dan lingkungan sekitarnya. Setiap baris dalam dataset merepresentasikan satu unit properti, dan kolom-kolomnya terdiri dari variabel-variabel seperti:

- S.no: Serial number
- MedInc: Pendapatan median penduduk di area tersebut.
- HouseAge: Usia bangunan rumah.
- AveRooms: Rata-rata jumlah kamar per rumah.
- AveBedrms: Rata-rata jumlah kamar tidur per rumah.
- Population: Total populasi di area tersebut.
- AveOccup: Rata-rata jumlah penghuni per rumah.
- Latitude dan Longitude: Koordinat geografis dari lokasi rumah.
- Price: Harga jual rumah.

4. HASIL DAN PEMBAHASAN

4.1. Data Preprocessing

Dataset berisi total 20.640 baris. Data tersebut sudah dalam keadaan bersih dan siap untuk lanjut ke tahap berikutnya.

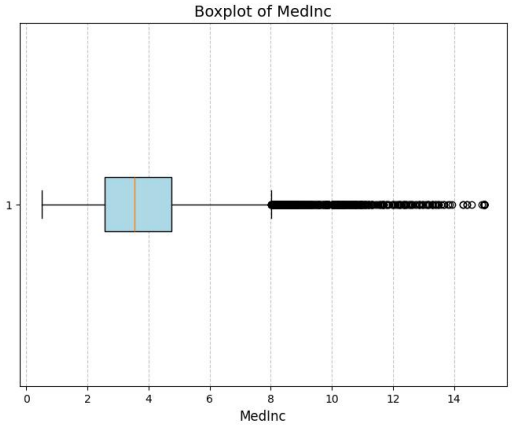
S.no	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	price
1	8.3252	41	6.984127	1.0238096	322	2.5555556	37.88	-122.23	4.526
2	8.3014	21	6.2381372	0.9718805	2401	2.1098418	37.86	-122.22	3.585
3	7.2574	52	8.288136	1.0734463	496	2.80226	37.85	-122.24	3.521
4	5.6431	52	5.817352	1.0730593	558	2.5479453	37.85	-122.25	3.413
5	3.8462	52	6.281853	1.081081	565	2.1814673	37.85	-122.25	3.422
6	4.0368	52	4.761658	1.103627	413	2.1398964	37.85	-122.25	2.697
7	3.6591	52	4.9319067	0.9513619	1094	2.1284046	37.84	-122.25	2.992
8	3.12	52	4.797527	1.0618238	1157	1.7882534	37.84	-122.25	2.414
9	2.0804	42	4.2941175	1.117647	1206	2.0268908	37.84	-122.26	2.267
10	3.6912	52	4.970588	0.99019605	1551	2.1722689	37.84	-122.25	2.611
11	3.2031	52	5.477612	1.079602	910	2.2636817	37.85	-122.26	2.815
12	3.2705	52	4.7724795	1.0245231	1504	2.0490463	37.85	-122.26	2.418
13	3.075	52	5.3226495	1.0128205	1098	2.3461537	37.85	-122.26	2.135
14	2.6736	52	4.0	1.0977012	345	1.9827586	37.84	-122.26	1.913
15	1.9167	52	4.262903	1.0096774	1212	1.9548388	37.85	-122.26	1.592
16	2.125	50	4.242424	1.0719697	697	2.6401515	37.85	-122.26	1.4
17	2.775	52	5.939577	1.0483384	793	2.3957703	37.85	-122.27	1.525
18	2.1202	52	4.0528054	0.9669967	648	2.138614	37.85	-122.27	1.555
19	1.9911	50	5.3436756	1.0859189	990	2.3627684	37.84	-122.26	1.587
20	2.6033	52	5.4654546	1.0836364	690	2.509091	37.84	-122.27	1.629

only showing top 20 rows

Gambar 2. Data yang sudah clean

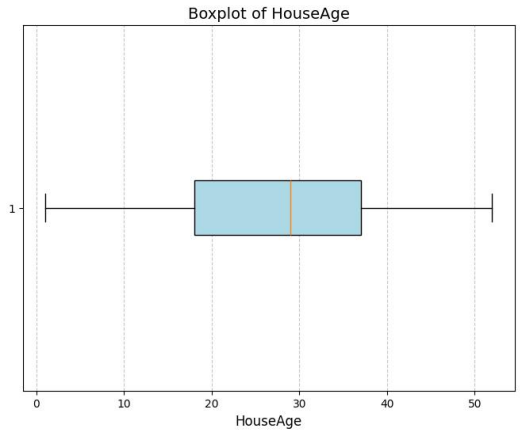
Pada Gambar 2, dapat dilihat bahwa data tersebut sudah clean, dimana tidak terdapat missing value. Data ini sudah siap untuk diproses ke tahap berikutnya.

4.2. Analisis Data



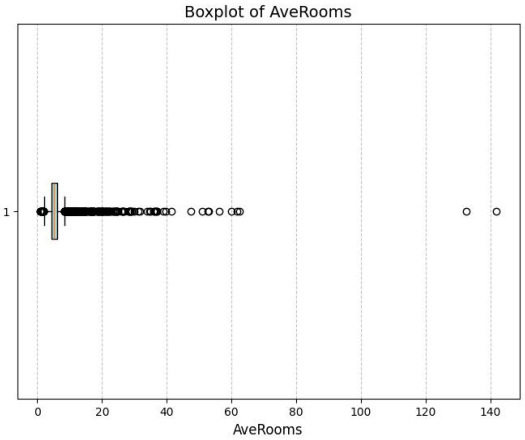
Gambar 3. boxplot atribut MedInc

Data pada Gambar 3 menunjukkan distribusi yang memanjang di bagian kanan, dengan beberapa outlier di nilai pendapatan tinggi. Hal ini mengindikasikan beberapa daerah dengan pendapatan yang sangat tinggi dibandingkan mayoritas data.



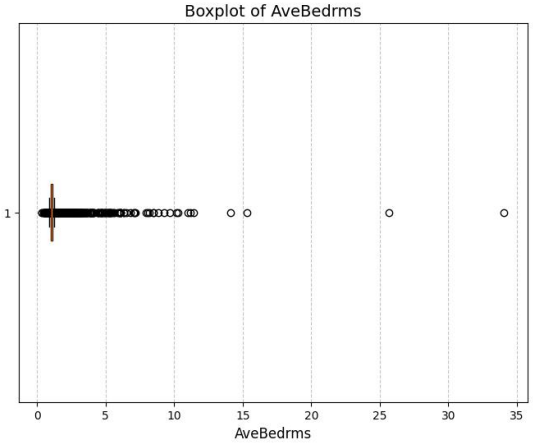
Gambar 4. boxplot atribut HouseAge

Pada Gambar 4, dapat dilihat bahwa sebagian besar nilai berada di rentang rendah hingga menengah, tetapi tidak banyak outlier yang signifikan. Usia rumah lebih tua cenderung lebih jarang, sesuai dengan outlier di sisi kanan.



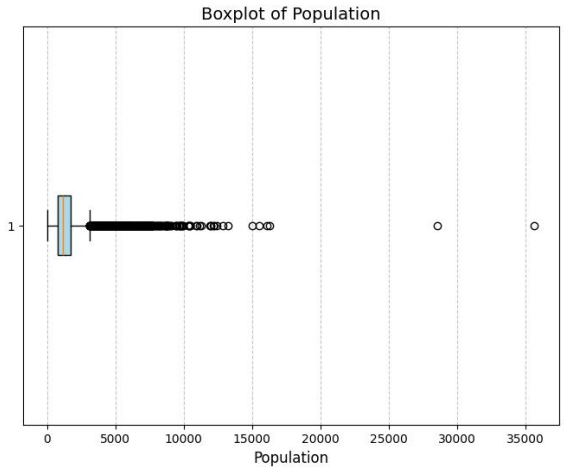
Gambar 5. boxplot atribut AveRooms

Terdapat beberapa outlier pada atribut AveRooms yang sangat tinggi, menunjukkan beberapa rumah dengan jumlah kamar yang jauh di atas rata-rata.



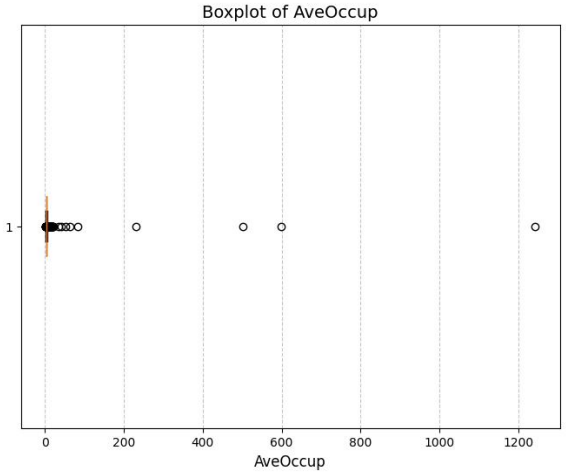
Gambar 6. boxplot atribut AveBedrms

Pada Gambar 6, dapat dilihat bahwa distribusi serupa dengan AveRooms, tetapi outlier di sini lebih sedikit. Ini menunjukkan rasio kamar tidur terhadap kamar lain lebih terdistribusi secara seragam.



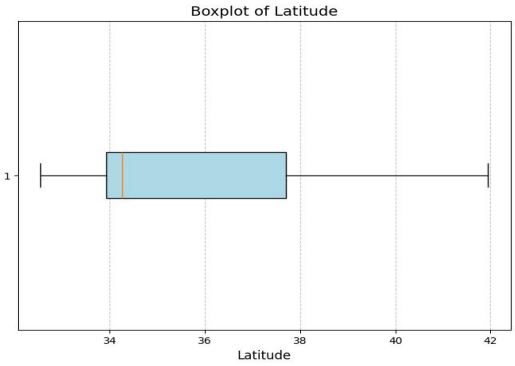
Gambar 7. boxplot atribut Population

Terdapat sejumlah besar outlier pada atribut Population di sisi kanan, menunjukkan beberapa area dengan populasi yang sangat tinggi.



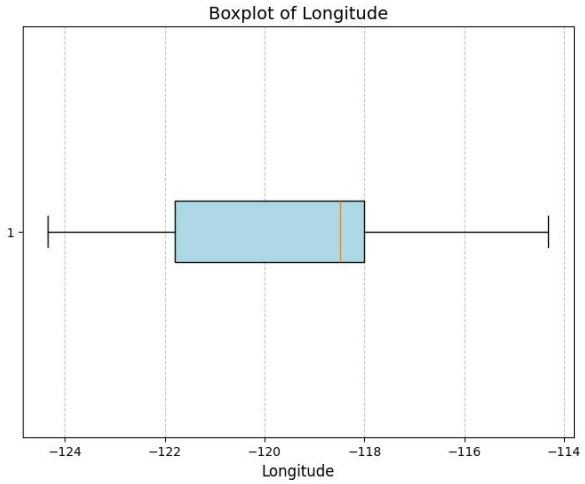
Gambar 8. boxplot atribut AveOccup

Pada Gambar 8, terlihat bahwa beberapa outlier signifikan terlihat di kedua sisi distribusi, menunjukkan beberapa rumah memiliki jumlah penghuni yang tidak lazim, baik sangat rendah maupun sangat tinggi.



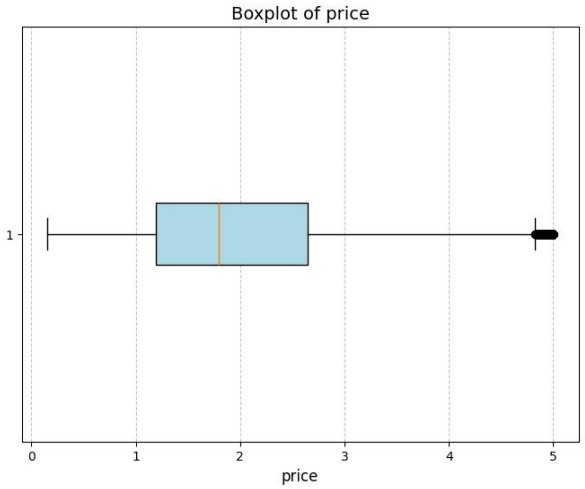
Gambar 9. boxplot atribut Latitude

Distribusi pada atribut Latitude ini tidak menunjukkan outlier signifikan, karena keduanya merepresentasikan lokasi geografis yang lebih stabil.



Gambar 10. boxplot atribut Longitude

Pada atribut Longitude, sama halnya seperti atribut Latitude, Distribusi ini tidak menunjukkan outlier signifikan, karena keduanya merepresentasikan lokasi geografis yang lebih stabil.



Gambar 11. boxplot atribut price

Pada Gambar 11, beberapa outlier terlihat pada harga yang sangat tinggi, yang kemungkinan adalah properti di lokasi premium.

4.3.Pembagian Data

Data tersebut telah melalui tahap pembersihan, dan selajutnya melakukan pembagian dataset menjadi 2 bagian, yakni data latih dan data uji. Didalam tahap ini, data dibagi dengan proporsi data pelatihan sebanyak 16.512 baris (80%) dan data pengujian sebanyak 4.128 baris (20%).

MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	price
0.4999	10	6.74	2.04	108	2.16	34.69	-116.9	0.55
0.4999	15	11.596491	2.5614035	131	2.2982457	40.43	-123.32	0.567
0.4999	16	21.631578	6.0	26	1.3684211	39.42	-122.89	0.735
0.4999	23	6.0545454	1.6727272	198	3.6	36.09	-119.99	1.0
0.4999	29	2.373272	1.0552995	2690	12.396314	34.02	-118.28	5.00001
0.4999	36	2.5	0.8333333	15	2.5	33.9	-118.04	1.625
0.4999	43	7.8461537	1.4615384	44	3.3846154	38.07	-120.19	1.625
0.4999	46	1.7142857	0.5714286	18	2.5714285	37.81	-122.29	0.675
0.4999	52	2.8709676	0.8548387	152	2.451613	37.65	-121.01	0.825
0.4999	52	3.875	0.5625	44	2.75	34.06	-117.75	1.125

only showing top 10 rows

Gambar 12. Data latih dengan proporsi (80%)

MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	price
0.4999	28	7.677419	1.8709677	142	4.580645	34.15	-118.08	5.00001
0.4999	52	2.6	0.73333335	74	1.6444445	33.93	-118.28	0.906
0.536	36	12.25	3.5	18	2.25	40.31	-123.17	0.14999
0.536	46	9.0	1.5	20	3.3333333	34.16	-118.42	3.75
0.7025	19	2.425197	1.1259843	1799	2.8330708	35.3	-120.67	5.00001
0.78	10	3.8357663	1.0839416	927	3.3832116	36.33	-119.28	0.555
0.8024	48	5.136646	1.2298137	396	2.4596274	39.51	-121.55	0.583
0.8056	48	4.38253	1.0662651	788	2.373494	37.81	-122.3	0.844
0.813	12	4.7810946	1.7014925	315	1.5671642	33.12	-117.1	2.75
0.8172	52	6.102459	1.3729508	728	2.9836066	37.82	-122.28	0.853

only showing top 10 rows

Gambar 13. Data uji dengan proporsi (20%)

4.4.Hasil Pengujian

```
Mean Square Error (MSE): 0.6441764407619649
```

Gambar 14. Mean Square Error (MSE)

Pada Gambar 14, dapat dilihat bahwa MSE menunjukkan rata-rata kuadrat dari kesalahan (selisih antara nilai prediksi dan nilai aktual). Nilai ini merepresentasikan seberapa besar kesalahan secara keseluruhan. Semakin kecil nilai MSE, semakin baik model memprediksi data.

```
Mean Absolute Error (MAE): 0.5997292513668271
```

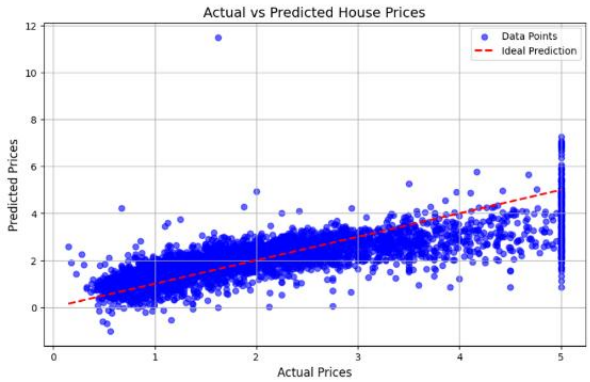
Gambar 15. Mean Absolute Error (MAE)

MAE adalah rata-rata nilai absolut dari kesalahan prediksi. Ini memberikan gambaran seberapa besar kesalahan dalam satuan yang sama dengan data. Nilai MAE yang rendah menunjukkan model yang lebih akurat.

```
R-squared (R²): 0.5082687836100642
```

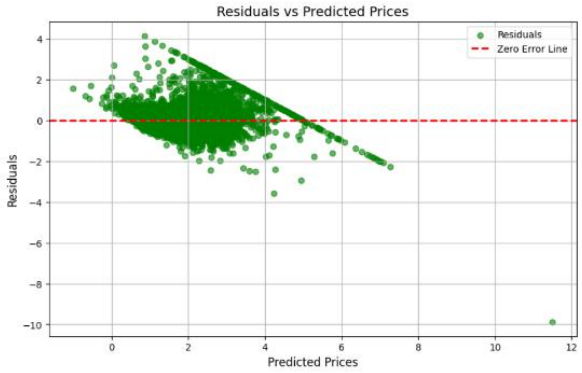
Gambar 16. R-Squared (R^2)

R^2 mengukur seberapa baik variabel independen menjelaskan variabel dependen. Nilai 0.508 berarti sekitar 50.8% variabilitas dalam data dapat dijelaskan oleh model regresi. Nilai ini cukup moderat, menunjukkan bahwa model masih memiliki ruang untuk perbaikan.



Gambar 17. Actual vs Predicted House Prices

Pada gambar 17, dapat dilihat bahwa sebagian besar prediksi berada di bawah garis ideal y=xy = xy=x, menunjukkan bahwa model cenderung meremehkan harga rumah, dengan outlier dan bias prediksi yang lebih jelas terlihat pada harga tinggi (aktual > 4).



Gambar 18. Residuals vs Predicted Prices

Pada gambar 18, dapat dilihat bahwa residual menunjukkan pola heteroskedastisitas, dengan varians yang tinggi pada harga prediksi rendah, serta bias model berupa overestimasi pada harga tinggi.

5. KESIMPULAN DAN SARAN

Model Linear Regression terbukti efektif dalam memprediksi harga real estate, menjadikannya alat potensial bagi pengembang properti atau agen real estate untuk memberikan estimasi harga berdasarkan data historis. Akurasi prediksi yang tinggi dapat mendukung pengambilan keputusan, seperti penetapan harga awal properti. Namun, model ini memiliki keterbatasan, terutama dalam menangani data yang tidak linier, seperti anomali pasar akibat pandemi atau kebijakan pemerintah.

Untuk meningkatkan kinerja model, disarankan untuk menambahkan fitur non-linier atau menggunakan metode machine learning lain, seperti decision tree regression, untuk meningkatkan fleksibilitas prediksi. Selain itu, evaluasi mendalam menggunakan metrik seperti Mean Squared Error (MSE), dan R-squared perlu dilakukan guna memastikan keakuratan prediksi. Perluasan dataset dengan menambahkan data dari lokasi atau tahun berbeda juga penting untuk menjaga relevansi model dalam berbagai kondisi pasar. Dengan langkah-langkah ini, model dapat ditingkatkan untuk memberikan analisis harga real estate yang lebih komprehensif dan akurat.

DAFTAR PUSTAKA

Gita Ayu Syafarina, Tri Wahyu Qur'ana, & Galih Mahalisa. (2022). *Prediksi Prospek Harga Real Estate di Masa Pandemi dengan 3 Atribut Berbasis Algoritma Linear Regression*. 603–608. https://doi.org/10.5281./5793/5.jupiter.2022.10

Khalijah Zainal, N. (2024). PREDIKSI HARGA REAL ESTATE MENGGUNAKAN METODE REGRESI LINEAR BERBASIS MACHINE LEARNING. *JOURNAL OF ARTIFICIAL INTELLIGENCE APPLICATION (JAIA) P-ISSN : Xxxx-Xxxx*, *1*(1).

Labib Mu'tashim, M., Damayanti, S. A., Zaki, H. N., Muhayat, T., & Wirawan, R. (2021). *Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression*. *3*, 238–245. https://doi.org/10.52958/iftk.v17i3.3635

Lathifah, U., & Danar Dana, R. (2024). IMPLEMENTASI METODE LINEAR REGRESSION UNTUK PREDIKSI HARGA PROPERTI REAL ESTATE MENGGUNAKAN RAPIDMINER. *Jurnal Mahasiswa Teknik Informatika*, *8*(1). https://doi.org/10.36040/jati.v8i1.8919

Rafif Nauval Tuah Siregar, Vijay Sitorus, & Willy Pramudia Ananta. (2023). Analisis Prediksi Harga Rumah di Bandung Menggunakan Regresi Linear Berganda. *Journal of Creative Student Research*, *1*(6), 395–404. https://doi.org/10.55606/jcsrpolitama.v1i6.3038