JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 2, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i2.6370

PREDIKSI HARGA RUMAH MENGGUNAKAN MULTIPLE LINEAR REGRESSION (STUDI KASUS : KABUPATEN KARAWANG PADA WEBSITE LAMUDI.CO.ID)

Silviana Anggellica[1*], Betha Nurina Sari[2], Iqbal Maulana[3]

[1,2,3]Informatika, Universitas Singaperbangsa Karawang; Jl. H.S. Ronggo Waluyo Telukjambe Timur Karawang

Received: 5 Maret 2025
Accepted: 27 Maret 2025
Published: 14 April 2025

Keywords:
Prediksi Harga;
Multiple Linear Regression;
OLS.

Corespondent Email:
2110631170036@student.unsika.ac.id

Abstrak. Keinginan alami adalah perilaku alami dalam bentuk melengkapi kehidupan yang layak. Termasuk keinginan untuk memiliki rumah. Sebagai pembeli yang akan membeli rumah dihadapkan pada situasi sulit untuk menebak harga. Maka, dibutuhkan sebuah model yang akurat untuk memperkirakan harga. Penelitian bertujuan untuk pemahaman terkait detail keterkaitan antara variabel independen dengan variabel dependen di dalam data dari situs lamudi, pengujian secara asumsi klasik dan pengujian parameter dilakukan, serta memastikan penggunaan model dengan pendekatan *multiple linear regression* dapat memperkirakan berdasarkan evaluasi R-squared, *Mean Squared Error*, dan *Root Mean Squared Error*. Metode penelitian yang digunakan adalah KDD, penyeleksian data dilakukan dengan menyatukan hasil penemuan yang didapatkan dari *web scraping*, terdapat pemilihan fitur dengan SelectKBest ANOVA F yang diuji coba kemudian dipilih yang terbaik. Pembagian data uji sebanyak 20% dan model OLS untuk pengujian asumsi klasik, maka didapat R-squared 0.51, MSE 0.53, dan RMSE 0.73 dan disimpulkan bahwa *error* relatif kecil. Sehingga nilai perkiraan berbeda sedikit dengan nilai aktual sebanyak 0.31.

Abstract. *Natural desires are natural behavior in the form of completing a decent life. Including the desire to own a house. As a buyer who wants to buy a house, you are faced with the difficult situation of guessing the price. So, an accurate model is needed to estimate prices. The research aims to understand the details of the relationship between the independent variable and the dependent variable in the data from the Lamudi site, classical assumption testing and parameter testing are carried out, as well as ensuring the use of a model with a multiple linear regression can estimate based on R-squared evaluation, Mean Squared Error, And Root Mean Squared Error. The research method used is KDD, data selection is carried out by combining the findings obtained from web scraping, there is feature selection with SelectKBest ANOVA F which is tested and then the best is selected. By dividing the test data by 20% and using the OLS model for testing classical assumptions, we get an R-squared of 0.51, MSE 0.53, and RMSE 0.73 and it is concluded that error relatively small. So the estimated value differs slightly from the actual value by 0.31.*

1. PENDAHULUAN

Keinginan alami adalah perilaku alami dalam bentuk melengkapi kehidupan yang

layak. Termasuk keinginan untuk memiliki rumah. Prinsip memiliki rumah adalah dengan membeli rumah. Membeli rumah dikaitkan dengan harga. Di Kabupaten Karawang adanya siklus kenaikan jumlah penduduk yang terjadi setiap tahun [1]. Sebagai pembeli yang akan membeli rumah dihadapkan pada situasi sulit untuk menebak harga. Maka, dibutuhkan sebuah model yang akurat untuk memperkirakan harga.

Pada penelitian sebelumnya yang diteliti oleh [2]. Penelitian dengan pendekatan *multiple linear regression* dan pendekatan *support vector regression* akan digunakan dan dijadikan pembanding, pendekatan yang mana yang menghasilkan evaluasi *error* yang baik. Penggunaan variabel independen diantaranya alamat, tanah (luas), bangunan (luas), kamar tidur (jumlah), kamar mandi (jumlah), dan harga sebagai variabel dependen dipergunakan pada penelitian ini. Diperoleh evaluasi *error* terhadap RMSE (*Root Mean Squared Error*) sebanyak 148.3 pada pendekatan pertama, dan sebanyak 153,4 pada pendekatan kedua. Walaupun pendekatan pertama atau *multiple linear regression* lebih baik dibanding *support vector regression* tetap akan dikatakan cukup tinggi, hal ini akan berpengaruh terhadap nilai prediksi nanti. Penelitian yang dilakukan tidak menggunakan pengujian berupa asumsi klasik. Apabila menggunakan pengujian ini, maka diketahui model dapat dikatakan tepat dalam prediksi [3]. Sehingga kurang efisien apabila tidak menggunakan ini.

Berdasarkan uraian di atas, penelitian yang digunakan menggunakan pendekatan yang serupa. Namun, yang membedakan penelitian ini adalah penggunaan pengujian asumsi klasik untuk mengetahui model tepat diprediksi dalam model OLS [3]. Maka topik penelitian ini adalah "Prediksi Harga Rumah Menggunakan *Multiple Linear Regression* (Studi Kasus : Kabupaten Karawang pada Website Lamudi.co.id)". Dengan tujuan untuk memperkirakan harga dengan melakukan pengujian yang telah direncanakan.

2. TINJAUAN PUSTAKA

2.1. Harga Rumah

Harga dikaitkan dengan berapa dana yang dikeluarkan untuk memiliki sebuah bangunan hunian [4]. Sedangkan harga rumah dikaitkan dengan konstruksi, dibangun dengan baik untuk menciptakan peningkatan harga.

2.2. Prediksi

Prediksi adalah cara kerja untuk mengevaluasi model dengan tujuan untuk memperkirakan hasil dengan akurat. Prediksi membutuhkan pengutamaan baru pada keahlian model untuk memodelkan variabel dependen yang terdapat di dalam data [5]. Prediksi yang berhubungan dengan harga membantu dalam memperkirakan nilai di masa depan.

2.3. KDD *(Knowledge Discovery in Database)*

KDD merupakan cara kerja yang beraturan untuk mengidentifikasi pola yang nantinya akan menghasilkan penjelas yang terdapat di dalam basis data dengan menggunakan algoritma [6]. Data mining memiliki pusat yang berproses mengenai *Knowledge Discovery in Databas*e [7].

2.4. Peranan Multiple Linear Regression

Dengan menggunakan multiple linear regression berarti akan menerapkan banyaknya variabel independen, lebih dari satu variabel. Model ini akan mendeteksi keterkaitan variabel - variabel pemberi dampak dengan variabel dependen [8]. Dengan begitu, dapat disimpulkan model ini kedalam bentuk rumusan sebagai berikut :

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

Dari persamaan 1 dapat diuraikan sebagai berikut :

Y	: Variabel pemberi dampak (independen)
$x_1, x_2, x_3, \ldots, x_n$	: Variabel terdampak (dependen)
α	: Konstanta
$\beta_1, \beta_2, \beta_3, .., \beta_n$	: Koefisien regresi yang menguji *independent variable* berpengaruh terhadap *dependent variable*.
ε	: *Error*

2.5. BLUE

Dengan menggunakan model yang berjenis *multiple linear regression* akan dikatakan baik

apabila sesuai dengan kualifikasi dari *Best* BLUE (*Linear Unbiased Estimator*). Dalam menggunakan BLUE, model harus menggunakan asumsi klasik. Dalam menguji asumsi klasik terbagi ke dalam empat pengujian, yaitu :

a. Residual Normality
Tahapan ini untuk menemukan uji asumsi yang tepat, diperlukan menganalisis data untuk menemukan ketepatan di dalam pendistribusian data. Menggunakan *Kolmogorov-Smirnov*, apabila adanya terima H_0 maka berdistribusi normal dengan (*asym sig 2 tailed*)>0.05. Kemudian sebaliknya [9].

b. Multicollinearity
Perumusan untuk menguji multikolinearitas adalah apabila H_0 diterima maka tidak menghasilkan multikolinearitas dengan VIF<10. Sedangkan apabila H1 ditolak maka adanya multikolinearitas dengan VIF>10 [9].

c. Heteroscedasticity

d. Apabila menggunakan *scatter plot* dan hasil yang diperoleh seperti kemunculan bulatan – bulatan kecil yang menyebar secara acak di sekitaran area garis 0, serta tidak terbentuk pola abstract maka dinyatakan tidak terjadi *heteroscedasticity* [10].

e. Autocorrelation
Perumusan menggunakan Durbin - Watson adalah apabila d_{hitung} yang berada diantara dL dan 4-dU (dL<d_{hitung}<4-dU) maka gagak menolak H_0 artinya tidak terjadi autokorelasi [11].

2.6. Parameter Test

Tahapan ini menggunakan pengujian T dan Pengujian F [12].

a. Pengujian T
Apabila t_{hitung} lebih besar dibandingkan t_{tabel} dengan signifikansi 5%, dengan begitu adanya penolakan H_0 diartikan sebagai variabel independen yang berpengaruh kepada variabel dependen. Sedangkan apabila t_{hitung} lebih kecil dibandingkan t_{tabel} dengan signifikansi 5%, dengan begitu adanya penerimaan H_0 diartikan sebagai variabel independen tidak memiliki pengaruh kepada variabel dependen [13]. Dalam pencarian t_{tabel} berdasarkan *degree of freedom* dengan jumlah data (n) – variabel (K).

b. Pengujian F
Apabila f_{hitung} lebih kecil dibandingkan f_{tabel} dengan signifikansi 5%, dengan begitu adanya penerimaan H_0 diartikan sebagai variabel independen tidak memiliki pengaruh kepada variabel dependen. Dan apabila f_{hitung} lebih besar dibandingkan f_{tabel} dengan signifikansi 5%, dengan begitu adanya penolakan H_0 diartikan sebagai variabel independen memiliki pengaruh terhadap variabel dependen [14]. Dalam pencarian f_{tabel} berdasarkan df1 dengan K (variabel) - 1. Beserta df2 dengan n - K.

2.7. MSE *(Mean Squared Error)*

MSE menghasilkan perhitungan yang diharuskan mendapatkan nilai yang rendah dikarenakan sebagai acuan dimana model prediksi dapat memprediksi nilai yang sebenarnya [15].

2.8. RMSE *(Root Mean Squared Error)*

RMSE adalah perhitungan untuk menimbang seberapa baik model memperkirakan. Apabila mempertimbangkan ketepatan perkiraan dengan data yang pasti harus menghasilkan nilai mendekati angka 0 atau mendekati nilai rendah karena semakin rendah yang dihasilkan dapat mempertimbangkan ketepatan perkiraan yang akan dicapai [15].

2.9. ANOVA

Seleksi fitur ini bekerja dengan memilih *feature* yang berada pada data yang dimiliki, menggunakan *feature* dependen untuk melakukan perhitungan *f-value*. Kemudian, seleksi fitur ini akan memilih berdasarkan *f-*

value untuk memilih pasangan independen yang nantinya akan melalui uji coba untuk memilih *feature* yang terbaik [16].

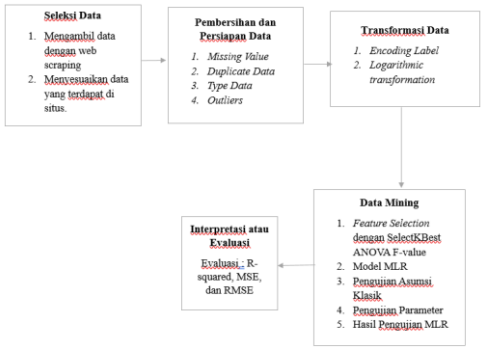
2.10. Web Scraping

Web scraping adalah cara kerja khusus dengan mendapatkan informasi dari *website*, pengambilan data ini tidak secara manual. Menggunakan teknik ini berguna untuk mendapatkan content dari situs, terutama cara mendapatkan link image dengan menyeleksi setiap image di dalam situs tersebut [17].

3. METODE PENELITIAN

3.1. Metodologi Penelitian

Penelitian ini menggunakan metodologi KDD yang terdiri dari lima tahapan, yaitu seleksi data, pembersihan dan persiapan data, transformasi data, dan interpretasi atau evaluasi yang dapat dilihat pada gambar 3.1.



Gambar 3.1 Metodologi Penelitian

3.2. Rancangan Penelitian

3.2.1. Data Seleksi

Mendapatkan data dengan melakukan teknik yang disebut dengan *web scraping* di situs lamudi dan melakukan penyaringan pada wilayah Karawang. Data yang didapatkan diintegrasikan ke dalam format CSV.

3.2.2. Pembersihan dan Persiapan Data

Data yang sudah diintegrasikan, kemudian dibersihkan dan dipersiapkan sebelum pemodelan.

3.2.3. Transformasi Data

Data yang telah dibersihkan, akan memasuki ketahapan selanjutnya yaitu ditransformasikan.

3.2.4. Data Mining

Di tahapan ini adanya pemilihan fitur untuk dilakukannya berbagai percobaan dan memilih percobaan yang terbaik untuk pengujian.

3.2.5. Interpretasi atau Evaluasi

Akhir tahapan adalah untuk mengevaluasi model.

4. HASIL DAN PEMBAHASAN

4.1. Data Selection

Proses ini berupa pencarian, data didapatkan dari situs lamudi dengan cara melakukan *web scrapin*g pada situs tersebut. Di dapatkan sebanyak 690 dari dua puluh empat halaman. Variabel independen yang digunakan adalah "lok" berdasarkan kecamatan atau kelurahan, "jkt" berdasarkan ruang (kamar tidur), "jkm" berdasarkan ruang (kamar mandi), "ltan" berdasarkan tanah (luas), "lban" berdasarkan bangunan (luas), dan harga sebagai variabel dependen yang dapat dilihat pada gambar 4.1.

	lok	jkt	jkm	ltan	lban	harga
0	Karawang Barat	4.0	2	84	110	1025000000
1	Telukjambe Barat	27.0	27	376	440	5500000000
2	Telukjambe Timur	28.0	28	376	440	5450000000
3	Karawang Timur	3.0	2	643	160	2500000000
4	Telukjambe Timur	3.0	1	77	154	685000000
...	...	...	...	...	...	...
685	Telukjambe Timur	2.0	1	90	54	749000000
686	Telukjambe Timur	3.0	2	200	125	1750000000
687	Karawang Barat	8.0	8	314	280	1700000000
688	Karawang Timur	3.0	2	72	100	360000000
689	Karawang Timur	2.0	1	78	72	386000000

690 rows × 6 columns

Gambar 4.1 Dataset yang digunakan

4.2. Pembersihan dan Persiapan Data

Proses ini dikaitkan dengan proses membersihkan data seperti menangani *missing value*. Berdasarkan gambar 4.2 bahwa terdapat missing value pada "jkt".

	Variabel	Jumlah Data Kosong
0	lok	0
1	jkt	2
2	jkm	0
3	ltan	0
4	lban	0
5	harga	0

Gambar 4.2 Menampilkan *Missing Value*

Dilakukan penanganan dengan nilai tengah untuk mengisi daerah yang kosong. Hasil dari penanganan dapat dilihat pada gambar 4.3.

	Variabel	Jumlah Data Kosong
0	lok	0
1	jkt	0
2	jkm	0
3	ltan	0
4	lban	0
5	harga	0

Gambar 4.3 Penanganan *Missing Value*

Proses yang kedua adalah menghilangkan *duplicated data*, pada gambar 4.4 merupakan banyaknya data yang *duplicated*, untuk hal ini diatasi dengan menghilang data menggunakan *drop_duplicates*. Maka dari 690 menjadi 572.

	lok	jkt	jkm	ltan	lban	harga
58	Karawang Barat	4.0	2	84	110	1025000000
59	Telukjambe Barat	27.0	27	376	440	5500000000
75	Karawang Timur	3.0	2	643	160	2500000000
92	Karawang Timur	3.0	2	114	108	1700000000
96	Telukjambe Timur	3.0	1	74	95	850000000
...	...	...	...	...	...	...
666	Klari	2.0	1	60	35	166000000
675	Telukjambe Timur	2.0	1	62	62	320000000
682	Karawang Timur	2.0	1	60	30	166000000
684	Telukjambe Timur	2.0	1	90	54	749000000
685	Telukjambe Timur	2.0	1	90	54	749000000

118 rows × 6 columns

Gambar 4.4 Banyaknya *Duplicated Data*

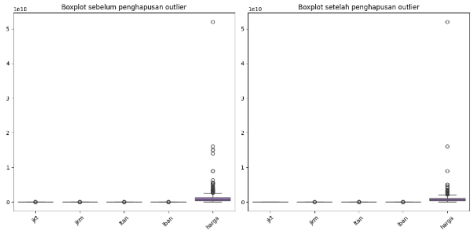
Proses yang ketiga adalah perubahan *type data*, dari *float* menjadi *int* pada "jkt" yang dapat dilihat pada gambar 4.5.

	lok	jkt	jkm	ltan	lban	harga
0	Karawang Barat	4	2	84	110	1025000000
1	Telukjambe Barat	27	27	376	440	5500000000
2	Telukjambe Timur	28	28	376	440	5450000000
3	Karawang Timur	3	2	643	160	2500000000
4	Telukjambe Timur	3	1	77	154	685000000
...	...	...	...	...	...	...
683	Karawang Timur	2	2	72	36	423500000
686	Telukjambe Timur	3	2	200	125	1750000000
687	Karawang Barat	8	8	314	280	1700000000
688	Karawang Timur	3	2	72	100	360000000
689	Karawang Timur	2	1	78	72	386000000

572 rows × 6 columns

Gambar 4.5 Perubahan *Type Data*

Proses yang keempat adalah untuk mengatasi adanya *outliers*. *Inter Quartile Range* (IQR) adalah teknik untuk mengatasi *outliers* pada penelitian ini. Pada gambar 4.6 merupakan gambar *before* dan *after* setelah ditangani dengan teknik ini. Penggunaan IQR ini menghasilkan 500 dari 572.



Gambar 4.6 Mengatasi *Outliers*

4.3. Transformation Data

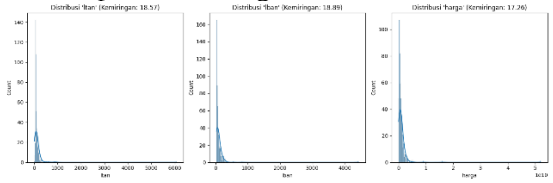
Tahapan yang dilakukan berupa penggunaan *encoding label* beserta *logarithmic transformation*. Transformasi pertama, adanya "lok" yang berbentuk kategorik maka perlu diubah menjadi bentuk numerik dengan menggunakan *encoding label* secara otomatis yang dapat dilihat pada gambar 4.7. Pada baris pertama "lok" berupa Karawang Barat diubah menjadi angka delapan di "lok1" begitupun dengan baris akhir Karawang Timur diubah menjadi angka sepuluh.

	lok	jkt	jkm	ltan	lban	harga	lok1
0	Karawang Barat	4	2	84	110	1025000000	8
3	Karawang Timur	3	2	643	160	2500000000	10
4	Telukjambe Timur	3	1	77	154	685000000	30
6	Karawang Timur	2	1	60	36	230000000	10
7	Karawang Timur	3	2	72	144	590000000	10
...	...	...	...	...	...	...	...
681	Karawang Barat	2	1	72	45	575000000	8
683	Karawang Timur	2	2	72	36	423500000	10
686	Telukjambe Timur	3	2	200	125	1750000000	30
688	Karawang Timur	3	2	72	100	360000000	10
689	Karawang Timur	2	1	78	72	386000000	10

500 rows × 7 columns

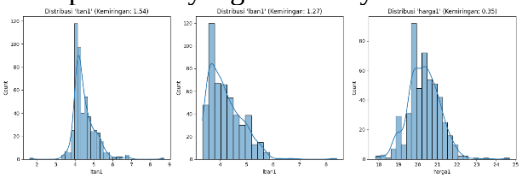
Gambar 4.7 Penggunaan *Encoding Label*

Transformasi kedua, penggunaan *logarithmic transformation*. Transformasi ini dilakukan untuk meminimalkan kemiringan (*skew*) pada data atau membuat distribusi yang merata. Pada gambar 4.8 merupakan bentuk awal tanpa dibuat dengan transformasi ini.



Gambar 4. 8 Sebelum Menggunakan Log

Pada gambar 4.8 bahwa variabel "ltan", "lban" dan "harga" memiliki *skew* masing – masing sebanyak 18.5 pada "ltan", 18.8 pada "lban" dan 17.2 pada "harga". Maka disimpulkan untuk menggunakan transformasi ini dengan tujuan data dari tidak merata menjadi merata yang dapat dilihat pada gambar 4.9. Penelitian yang dilakukan oleh [18]. Menggunakan transformasi ini untuk mengatasi *outlier* pada data yang dimilikinya.



Gambar 4.9 Setelah Menggunakan Log

Gambar 4.9 menghasilkan perubahan berupa pengurangan *skew* dengan penambahan variabel seperti "ltan1", "lban1" dan "harga1" yang telah dilakukan *logarithmic transformation* yang dapat dicermati pada gambar 4.10. *Skew* pada "ltan1" sebanyak 1.54, *skew* pada "lban1" sebanyak 1.27, *skew* pada "harga1" sebanyak 0.35.

	lok	jkt	jkm	ltan	lban	harga	lok1	ltan1	lban1	harga1
0	Karawang Barat	4	2	84	110	1025000000	8	4.442651	4.709530	20.747958
3	Karawang Timur	3	2	643	160	2500000000	11	6.467699	5.081404	21.639557
4	Telukjambe Timur	3	1	77	154	685000000	33	4.356709	5.043425	20.344929
6	Karawang Timur	2	1	60	36	230000000	11	4.110874	3.610918	19.253590
7	Karawang Timur	3	2	72	144	590000000	11	4.290459	4.976734	20.195633
...	...	...	...	...	...	...	...	...	...	...
681	Karawang Barat	2	1	72	45	575000000	8	4.290459	3.828641	20.169881
683	Karawang Timur	2	2	72	36	423500000	11	4.290459	3.610918	19.864064
686	Telukjambe Timur	3	2	200	125	1750000000	33	5.303305	4.836282	21.282882
688	Karawang Timur	3	2	72	100	360000000	11	4.290459	4.615121	19.701615
689	Karawang Timur	2	1	78	72	386000000	11	4.369448	4.290459	19.771348

500 rows × 10 columns

Gambar 4.10 Menampilkan Variabel Baru

4.4. Data Mining

Adanya percobaan yang dilakukan untuk menghasilkan yang sempurna. Percobaan yang dilakukan berdasarkan 9 percobaan, 7 diantaranya menggunakan *feature selection* SelectKBest ANOVA F. Percobaan ke-1 ketika melakukan pengujian asumsi klasik didapatkan hasil tidak normal, terdapat dua multikolinearitas pada variabel "ltan" memiliki nilai VIF sebanyak 16.6 sedangkan pada variabel "lban" memiliki nilai VIF sebanyak 16.8. Beserta salah satu variabel "jkt" tidak memiliki pengaruh terhadap variabel dependen. Oleh sebab itu, dilakukan percobaan ke-2 dengan menggunakan *logarithmic transformation*. Hasilnya VIF menjadi lebih kecil, namun terjadi autokorelasi, dan variabel "lok1" tidak berkaitan dengan variabel dependen. Pemilihan fitur dilakukan dengan mencoba 9 percobaan, diantaranya sebagai berikut :

1. Penggunaan fitur seperti lok1, jkt, jkm, ltan, lban, dan harga. Menghasilkan r-squared sebanyak 0.95, MSE sebanyak 2817354.4, sedangkan RMSE sebanyak 1678.4. Dengan hasil pengujian yaitu tidak berdistribusi normal, multikolinearitas, heteroskedastisitas, dan 1 variabel tidak berkaitan dengan variabel dependen.
2. Penggunaan fitur seperti lok1, jkt, jkm, ltan1, lban1, dan harga1. Menghasilkan r-squared sebanyak 0.58, MSE sebanyak 0.41, sedangkan RMSE sebanyak 0.64. Dengan hasil pengujian yaitu tidak berdistribusi normal, autokorelasi, dan 1 variabel tidak berkaitan dengan variabel dependen.
3. Penggunaan fitur seperti jkt, jkm, dan harga1. Menghasilkan r-squared sebanyak 0.33, MSE sebanyak 0.65, sedangkan RMSE sebanyak 0.80. Dengan hasil pengujian yaitu tidak berdistribusi normal, dan heteroskedastisitas.
4. Penggunaan fitur seperti jkm, ltan1 dan harga1. Menghasilkan r-squared sebanyak 0.51, MSE sebanyak 0.53, sedangkan RMSE sebanyak 0.73. Dengan hasil pengujian yaitu tidak berdistribusi normal.
5. Penggunaan fitur seperti jkm, lban1, dan harga1. Menghasilkan r-squared sebanyak 0.57, MSE sebanyak 0.50, sedangkan RMSE sebanyak 0.70. Dengan hasil pengujian yaitu tidak berdistribusi normal, dan autokorelasi.
6. Penggunaan fitur seperti jkt, ltan1, dan harga1. Menghasilkan r-squared sebanyak 0.44, MSE sebanyak 0.44, sedangkan RMSE sebanyak 0.66. Dengan hasil pengujian yaitu tidak berdistribusi normal, dan heteroskedastisitas.
7. Penggunaan fitur seperti jkt, jkm, ltan1, dan harga1. Menghasilkan r-squared sebanyak 0.51, MSE sebanyak 0.49, sedangkan RMSE sebanyak 0.70. Dengan hasil pengujian yaitu tidak berdistribusi normal, dan heteroskedastisitas.
8. Penggunaan fitur seperti jkt, jkm, lban1, dan harga1. Menghasilkan r-squared sebanyak 0.57, MSE sebanyak 0.50, sedangkan RMSE sebanyak 0.70. Dengan hasil pengujian yaitu tidak normal, heteroskedastisitas, dan autokorelasi.
9. Penggunaan fitur seperti jkt, jkm, ltan1, lban1, dan harga1. Menghasilkan r-squared sebanyak 0.58, MSE sebanyak 0.47, sedangkan RMSE sebanyak 0.69. Dengan hasil pengujian yaitu tidak berdistribusi normal, dan autokorelasi.

Dengan memutuskan hasil pengujian yang diperoleh, serta berdasarkan evaluasi yang

didapatkan di setiap percobaan. Maka, terpilih percobaan ke-4 karena hanya terdapat 1 permasalahan dan dapat ditangani.

4.4.1. Multiple Linear Regression

Gambar 4.11 adalah mendefinisikan variabel sesuai dengan fitur dari percobaan ke-4.

```
X = dfr[['jkm', 'ltan1']]
y = dfr['harga1']
```

Gambar 4.11 Mendefinisikan Variabel

Membagi data uji sebanyak 20% di ambil dari pengujian, dan 80% digunakan untuk pelatihan yang dapat dicermati pada gambar 4.12.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 4.12 Pembagian Dataset

Membangun model regresi linear yang dapat dicermati pada gambar 4.13.

```
regr1 = LinearRegression()
regr1.fit(X_train, y_train)
```

Gambar 4.13 Regresi Linear

Perbandingan antara nilai prediksi dengan nilai aktual dari uji data (y_test) pada model regresi di dalam y_pred, hasilnya dapat dicermati pada gambar 4.14.

	Prediction	Test
0	19.809032	19.485391
1	21.068101	21.247994
2	20.481990	20.863028
3	20.444065	18.792244
4	20.655797	21.202601

Gambar 4.14 Membandingkan Nilai

Membuat model untuk OLS, pemakaian model ini digunakan di dalam pengujian dapat dicermati pada gambar 4.15.

```
X = sm.add_constant(X_train)
mod1 = sm.OLS(y_train, X).fit()
```

Gambar 4.15 OLS

4.4.2. Pengujian Asumsi Klasik

Pengujian *normality residual* dengan menggunakan Kolmogorov – Smirnov. Hasilnya dapat dicermati pada gambar 4.16.

```
KstestResult(statistic=0.08917670262886934, pvalue=0.003227122106655795, statistic_location=-0.17213011152572832, statistic_sign=-1)
```

Gambar 4.16 Normalitas Residual

Gambar 4.15 didapatkan hasil uji normalitas residual sebanyak 0.003 yang artinya lebih kecil dari 0.05 atau menerima H_1 yang berarti tidak berdistribusi normal. Dalam mengatasi hal ini digunakan asumsi *central limit theorem* dengan tujuan data berdistribusi normal. Karena, *central limit theorem* menyatakan bahwa sampel data yang dimiliki di atas 30 adalah normal [19].

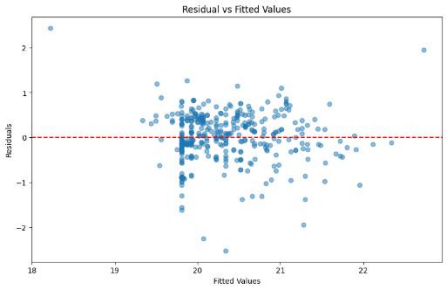
Pengujian yang kedua adalah uji multikolinearitas untuk mendapatkan nilai VIF dapat dicermati pada gambar 4.17.

	Variabel	VIF
0	const	58.192619
1	jkm	1.102252
2	ltan1	1.102252

Gambar 4.17 Multikolinearitas

Gambar 4.16 adalah menghasilkan nilai VIF untuk variabel "jkm" dan variabel "ltan1". Pada variabel "jkm" didapatkan VIF sebanyak 1.102 yang berarti lebih kecil dari 10. Sama halnya dengan variabel "ltan1". Maka dapat disimpulkan menerima H_0 atau tidak terjadi multikolinearitas.

Pengujian yang ketiga adalah uji heteroskedastisitas dengan menggunakan *scatter plot* yang hasilnya dapat cermati pada gambar 4.18. Berdasarkan gambar 4.18 bahwa plot tersebut mengidentifikasi adanya bulatan – bulatan kecil yang menyebar secara acak disekitaran area garis 0, yang artinya tidak terjadi heteroskedastisitas.



Gambar 4.18 Heteroskedastisitas

Pengujian yang keempat adalah uji autokorelasi dengan menggunakan Durbin – Watson yang dapat dicermati pada gambar 4.18. Gambar 4.18 bahwa tidak terjadi autokorelasi dikarenakan sesuai dengan kriteria rumusan yaitu dL<d_{hitung}<4-dU atau 1.84<1.88<4-1.86. Karena, banyaknya data sekitar 500 dan jumlah variabel yang digunakan adalah 3. Maka, berdasarkan tabel Durbin – Watson didapatkan nilai dL sebanyak 1.84 sedangkan nilai dU sebanyak 1.86. Dengan d_{hitung} yang terlihat pada gambar 4.19 sebanyak 1.88.

```
dw = durbin_watson(mod2)
print(dw)
```

```
1.883205535064141
```

Gambar 4.19 Autokorelasi

4.4.3. Pengujian Parameter

Pengujian parameter yang pertama adalah pengujian T. Berdasarkan rumusan bahwa apabila $t_{hitung} > t_{tabel}$ maka adanya penolakan H_0 artinya variabel independen berkaitan dengan variabel dependen.

```
# Uji T
print(mod1.tvalues)
```

```
const    80.045147
jkm      10.965153
ltan1    13.060695
dtype: float64
```

Gambar 4.20 Pengujian T

Berdasarkan gambar 4.20 dapat dinyatakan kedua variabel berpengaruh terhadap variabel dependen. Karena nilai t_{hitung} pada masing – masing variabel "jkm" dan "ltan" adalah lebih besar dari t_{tabel}. Untuk mencari t_{tabel} yaitu dengan menggunakan signifikansi 5% dan df-nya adalah 497 dari n – K yang dapat dilihat pada gambar 4.21. Maka didapatkan t_{tabel} sebanyak 1.96.

```
=TINV(0.05,497)
```

Gambar 4.21 Menemukan Ttabel

Pengujian parameter yang kedua adalah pengujian F. Berdasarkan rumusan bahwa apabila $f_{hitung} > f_{tabel}$ maka adanya penolakan H_0 artinya variabel independen berkaitan dengan variabel dependen.

```
# Uji F
print('F-statistic: ', mod1.fvalue)
```

```
F-statistic:  208.35571767802594
```

Gambar 4.22 Pengujian F

Berdasarkan gambar 4.22 dapat dinyatakan kedua variabel keterkaitan (pengaruh) terhadap variabel dependen, karena adanya penolakan H_0. Untuk mencari f_{tabel} yaitu dengan menggunakan signifikansi 5% dan df1-nya adalah 2 dari K – 1 sedangkan df2-nya adalah 497 dari n – K yang dapat dilihat pada gambar 4.23. Maka didapatkan f_{tabel} sebanyak 3.013.

```
=FINV(0.05, 2, 497)
```

Gambar 4.23 Menemukan Ftabel

4.4.4. Hasil Pengujian Regresi

Hasil pengujian regresi yang didapatkan menggunakan *tools* yaitu Google Colaboratory dapat dilihat pada gambar 4.24.

```
intercept = mod1.params['const']
X1_coef = mod1.params['jkm']
X2_coef = mod1.params['ltan1']


print(f"Intercept: {intercept}")
print(f"Coefficient 'jkm' (X1): {X1_coef}")
print(f"Coefficient 'ltan1' (X2): {X2_coef}")
```

```
Intercept: 16.78909384294763
Coefficient 'jkm' (X1): 0.41230407258651924
Coefficient 'ltan1' (X2): 0.6343260733055074
```

Gambar 4.24 Hasil Uji Regresi

Berdasarkan gambar 4.23 diperoleh hasil :

α = 16.789

x_1 = 0.412

x_2 = 0.634

Maka dapat disimpulkan berdasarkan persamaan yaitu sebagai berikut :

$$Y = 16.789 + 0.412x_1 + 0.634x_2$$

4.5. Interpretasi atau Evaluasi

Pada tahapan ini menggunakan evaluasi R-squared, MSE, dan RMSE yang dapat dilihat pada gambar 4.25.

```
R-squared:  0.5121120550231832
MSE:  0.5359375754474082
RMSE:  0.7320775747469719
```

Gambar 4.25 Evaluasi

Berdasarkan gambar 4.25 menghasilkan r-squared sebanyak 0.51, MSE sebanyak 053, dan RMSE sebanyak 0.73 yang berarti mendapatkan nilai rendah dan kurang untuk memprediksi secara tepat. Untuk membuktikan seberapa tepat model ini dalam memprediksi dapat dicermati pada gambar 4.26.

```
Masukkan nilai jkm: 2
Masukkan nilai ltan1: 4.442651
Prediksi Harga Rumah:  [20.43179135]
```

Gambar 4.26 Prediksi Harga Rumah

Berdasarkan gambar 4.26 bahwa "jkm" adalah jumlah kamar mandi, sedangkan "ltan1" adalah luas tanah dengan nilai logaritma dari "ltan", sama dengan "harga1" termasuk nilai logaritma dari "harga". Prediksi yang diperoleh sebanyak 20.431, "jkm" yang berjumlah 2 dan "ltan1" yang berjumlah 4.442 diambil dari data dari baris pertama yang dapat dilihat pada gambar 4.10. Sedangkan harga aktual dalam bentuk logaritma adalah 20.747. Maka

didapatkan selisih sebesar 0.31, tidak begitu jauh dari harga aktualnya.

5. KESIMPULAN

a. Penelitian ini memilih percobaan ke-4 dengan fitur yang digunakan "jkm" dan "ltan1", serta penggunaan 1 variabel dependen yaitu "harga1". Maka dapat dinyatakan adanya variabel independen adanya pengaruh secara signifikan secara parsial (pengujian T) secara serempak dalam memperkirakan harga di Kabupaten Karawang yang didapatkan dari situs lamudi dengan teknik *web scraping*.
b. Penelitian ini menghasilkan evaluasi yang terdiri dari R-squared, MSE, dan RMSE. Dari R-squared diperoleh sebanyak 0.51, dari MSE diperoleh sebanyak 0.53, sedangkan dari RMSE diperoleh sebanyak 0.73.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak yang terkait, terutama orang tua yang telah memberikan dukungan terhadap penelitian ini, dan dosen pembimbing yang telah membimbing penulis dalam melakukan penelitian.

DAFTAR PUSTAKA

[1] M. A. Nurdien and B. J. Rijal Zulfikar Habibie, "Perlindungan Hukum Terhadap Hak Penghuni Kawasan Perumahan Atas Tempat Pemakaman di Kabupaten Karawang," *JURNAL SYNTAX IMPERATIF : Jurnal Ilmu Sosial dan Pendidikan*, vol. 2, no. 6, pp. 566–575, Jan. 2022, doi: 10.36418/syntax-imperatif.v2i6.142.

[2] A. Handani, A. Siregar, and T. Mudzakir, "Prediksi Harga Rumah Di Karawang Menggunakan Algoritma Multiple Linear Regression dan Support Vector Regression," *Scientific Student Journal for Information, Technology and Science* , vol. 5, no. 2, pp. 33–40, 2024.

[3] A. Widyastuti, "Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Metode Multiple Linear Regression," *Jurnal Ilmiah Teknologi Informasi dan Sains*, vol. 4, no. 1, pp. 30–35, 2024, [Online]. Available: http://ejurnal.unim.ac.id/index.php/submit

[4] R. Khoirudin, U. Khasanah, and U. Ahmad Dahlan, "Dampak Kebijakan LTV Terhadap Harga Properti Berdasarkan Pendekatan Spasial," *JURNAL ILMIAH KOHESI*, vol. 6, no. 1, pp. 148–157, Jan. 2022, [Online]. Available: https://kohesi.sciencemakarioz.org/index.php/JIK/article/view/342/342

[5] M. D. Verhagen, "A Pragmatist's Guide to Using Prediction in the Social Sciences," *Socius*, vol. 8, pp. 1–17, Feb. 2022, doi: 10.1177/23780231221081702.

[6] N. Marito Putry and B. Nurina Sari, "Komparasi Algoritma KNN dan Naive Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Melitus," *Jurnal Sains dan Manajemen*, vol. 10, no. 1, pp. 45–57, 2022.

[7] M. S. Ma'arif, H. J. Jaman, and A. S. Irawan, "Analisis Sentimen Ulasan Aplikasi Investasi Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, pp. 2004–2012, Aug. 2024, doi: 10.23960/jitet.

[8] Rafif Nauval Tuah Siregar, Vijay Sitorus, and Willy Pramudia Ananta, "Analisis Prediksi Harga Rumah di Bandung Menggunakan Regresi Linear Berganda," *Journal of Creative Student Research*, vol. 1, no. 6, pp. 395–404, Dec. 2023, doi: 10.55606/jcsrpolitama.v1i6.3038.

[9] R. Romadhoni, R. Yanti, T. Nasution, and K. Anam, "Analisis Faktor Hasil Produksi Kelapa Sawit Menggunakan Regresi Linier Berganda Studi Kasus : Koperasi Unit Desa (KUD) Setia Kawan Desa Koto Damai," *Formosa Journal of Science and Technology (FJST)*, vol. 1, no. 4, pp. 217–234, 2022, [Online]. Available: https://journal.formosapublisher.org/index.php/fjst

[10] D. Purba, W. Tarigan, M. Sinaga, and V. Tarigan, "Pelatihan Penggunaan Software SPSS Dalam Pengolahan Regressi Linear Berganda Untuk Mahasiswa Fakultas Ekonomi Universitas Simalungun Di Masa Pandemi Covid 19," *Jurnal Karya Abadi*, vol. 5, no. 2, pp. 202–208, 2021.

[11] F. Azizah and D. M. Athoillah, "Analisis Dampak Covid-19 Terhadap Indeks Harga Konsumen dengan K-Means dan Regresi Berganda," *Indonesian Journal of Applied Statistics*, vol. 4, no. 1, pp. 21–33, May 2021, doi: https://doi.org/10.13057/ijas.v4i1.46329.

[12] N. A. AHMAD and R. Raupong, "Estimation Of Parameter Regression Panel Data Model Using Least Square Dummy Variable

Method," *Jurnal Matematika, Statistika dan Komputasi*, vol. 20, no. 1, pp. 221–228, Sep. 2023, doi: 10.20956/j.v20i1.27530.

[13] S. Setiawati, "Analisis Pengaruh Kebijakan Deviden Terhadap Nilai Perusahaan Pada Perusahaan Farmasi Di BEI," *Jurnal Inovasi Penelitian*, vol. 1, no. 8, pp. 1581–1590, 2021.

[14] T. N. Padilah and R. I. Adam, "Analisis Regresi Linear Berganda Dalam Estimasi Produktivitas Tanaman Padi Di Kabupaten Karawang," *Jurnal Pendidikan Matematika dan Matematika*, vol. 5, pp. 117–128, 2019.

[15] I. Amansyah, J. Indra, E. Nurlaelasari, and A. R. Juwita, "Prediksi Penjualan Kendaraan Menggunakan Regresi Linear: Studi Kasus pada Industri Otomotif di Indonesia," *INNOVATIVE: Journal Of Social Science Research* , vol. 4, no. 4, pp. 1199–1216, 2024, Accessed: Mar. 01, 2025. [Online]. Available: https://j-innovative.org/index.php/Innovative

[16] M. F. Thoriq, W. J. Pranoto, and F. Faldi, "Penerapan Seleksi Fitur Analysis of Variance Pada Algoritma Random Forest Classifier Dalam Klasifikasi Nilai Mahasiswa," *Explore: Jurnal Sistem Informasi dan Telematika*, vol. 14, no. 2, p. 185, Dec. 2023, doi: 10.36448/jsit.v14i2.3187.

[17] A. Fauzia Putri, G. Manik, F. Nabila, and N. Chamidah, "Implementasi Scraping Google Scholar Menggunakan HTML DOM Untuk Pengumpulan Data Artikel Dosen UPN Veteran Jakarta Berbasis Web," in *Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya 2021*, Jakarta: Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta, Apr. 2021, pp. 668–678.

[18] M. Fihan Ashidiq, L. Muflikhah, and B. D. Setiawan, "Deteksi Nefropati Diabetik Pada Pasien Diabetes Melitus Menggunakan Regresi Logistik," *urnal Pengembangan Teknologi Informasi dan Ilmu Komputer* , vol. 9, no. 2, pp. 2548–964, Feb. 2025, [Online]. Available: http://j-ptiik.ub.ac.id

[19] H. Arsham, "Systems Simulation: The Shortest Route to Applications," *Dr. Hossein Arsham*, pp. 1–67, Oct. 2020, Accessed: Mar. 01, 2025. [Online]. Available: https://www.researchgate.net/publication/344638606