PREDIKSI HARGA RUMAH DI MADRID DENGAN METODE MULTIPLE LINEAR REGRESSION

Delatifa Putri Sugandi, Putri Mellia Zahrani
School of Electrical Engineering
Telkom University, Indonesia
delatifaps@student.telkomuniversity.ac.id, putrimelliazahrani@telkomuniversity.ac.id

Abstrak

Rumah atau tempat tinggal merupakan salah satu kebutuhan primer manusia selain sandang dan pangan. Sebagian orang menginvestasikan sebagian hartanya pada properti atau Rumah. Dengan perkembangan teknologi kita dapat memprediksikan harga rumah di waktu mendatang dengan menggunakan data yang tersedia ada pada saat ini. Prediksi harga rumah dilakukan menggunakan metode Regresi Linear. Regresi Linear merupakan salah satu metode statistik yang memberikan hasil output prediksi dengan melakukan pengembangan hubungan matematis antar variabel. Terdapat 2 variabel yaitu variabel independen dan variabel dependen, dalam kasus ini variabel independennya adalah luas tanah, banyaknya kamar, banyaknya kamar mandi, banyaknya lantai, dan tipe rumah, sedangkan untuk variabel independennya adalah harga rumah. Dilakukan juga uji akurasi dan uji kinerja pemodelan menggunakan 3 jenis evaluasi yaitu RMSE, MSE, MAE dan R-2 Score.

Kata Kunci — Regresi Linear, Prediksi Harga Rumah,RMSE, MSE, MAE, R-2 Score.

I. PENDAHULUAN

Rumah adalah salah satu kebutuhan dari masyarakat yang tidak dapat untuk dihindari dikarenakan rumah merupakan kebutuhan primer, tempat untuk berlindung, tempat untuk beristirahat dari penatnya aktivitas harian. Tidak kalah dengan emas, rumah pun juga bisa dijadikan alat untuk berinvestasi di masa yang akan datang dikarenakan pergerakan harganya yang berubah sewaktu-waktu dan semakin banyak orang yang membutuhkan rumah, terlebih lagi dekat dengan lapangan kerja, pusat perkantoran, pusat perbelanjaan, sarana transportasi, dan lain sebagainya pasti akan mempengaruhi harga rumah tersebut dengan cepat [1].

Seiring berjalannya waktu, kebutuhan fisiologis manusia akan semakin bertambah, salah satunya adalah kebutuhan dalam membeli rumah. Pengusaha properti akan berlomba-lomba membangun properti khususnya rumah untuk sarana investasi. Hal ini akan membuat harga rumah semakin hari semakin naik dengan daya beli masyarakat yang melonjak tinggi. Tentunya akan membuat masyarakat dalam membeli rumah berfikir apakah rumah yang ia beli akan mempunyai nilai keuntungan yang baik atau tidak. Dalam berinvestasi, tak lepas dengan menebak naik turunnya harga agar tidak rugi dalam berinvestasi. Harga yang tidak pasti dan tidak terprediksi ini membuat investor atau pembeli rumah membutuhkan sebuah sistem untuk memprediksikan harga rumah berdasarkan letak rumah tersebut[2].

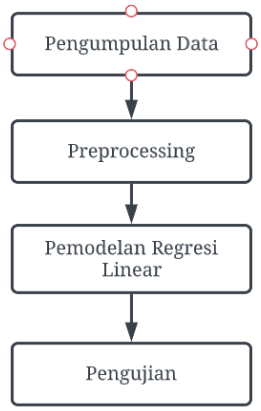
Seiring dengan perkembangan, prediksi digunakan sebagai alat bantu bahkan prediksi juga telah dijadikan suatu bahan pertimbangan dalam pengambilan keputusan pembelian. Dari penelitian yang dilakukan oleh Sefto P. (2016) dengan penelitian yang berjudul "Prediksi Harga Tanah menggunakan Algoritma Linear Regression" mengatakan bahwa linear regression dapat memberikan hasil pengukuran tingkat akurasi yang tinggi dan RMSE (Root Mean Square Error) terendah dengan data yang digunakan untuk training 70% dan sisanya data yang digunakan untuk testing 30%. Sindhu P. M. (2017) pada penelitiannya menghasilkan tingkat akurasi yang sangat baik dibandingkan dengan analisis deret waktu. Adapun penelitian lainnya yang menggunakan algoritma serupa adalah penelitian prediksi curah hujan di Kabupaten Majalengka, penelitian prediksi penjualan laptop, penelitian prediksi jumlah peminat mata kuliah, prediksi hasil panen. Berdasarkan penelitian tersebut dapat diambil kesimpulan bahwa melakukan prediksi menggunakan algoritma Linear Regression menghasilkan tingkat akurasi yang cukup baik jika digunakan untuk memprediksi[1].

Sehingga dalam penelitian ini, diusulkan penggunaan metode Regresi Linear

menggunakan semua variabel yang ada dengan harapan prediksi yang didapatkan memiliki tingkat keakuratan yang cukup. Implementasinya berupa sistem prediksi harga rumah di Madrid diharapkan mampu memberikan informasi harga rumah yang sesuai dengan keadaan yang diharapkan.

2. METODE PENELITIAN

Untuk memprediksikan harga rumah, maka kita memerlukan beberapa data seperti data harga rumah pada tahun dan tempat tertentu. Setelah itu, data diolah dengan menggunakan Metode Regresi Linear. Di bawah ini merupakan tahapan - tahapan bagaimana program ini dibuat :



1. Pengumpulan Data
   Pengumpulan data dilakukan dengan mendownload data pada situs https://kaggle.com. Data yang diambil 17 variabel dan 21739 baris, dari 17 variabel diambil 5 variabel X atau variabel dependen, dan 1 variabel Y atau variabel independen.
2. Preprocessing
   Dalam proses Preprocessing dilakukan pembersihan data ketika ada data yang missing value, format tidak konsisten dan outliner berlebihan. Data yang sudah diambil lalu di lakukan pengecekan tipe data, statistik data, dan banyaknya data yang null. Jika terdapat data yang null atau kosong maka dapat dilakukan pembersihan data.
3. Regresi Linear
   Regresi Linier adalah perhitungan berdasarkan sebuah variabel x yang merupakan variabel dependen dan variabel y yang merupakan variabel independen dan bisa membuat sebuah acuan untuk prediksi. Garis itu adalah kumpulan dari data yang menunjukkan peningkatan dan digunakan untuk proyeksi berdasarkan data maupun perbandingan tersebut, dengan rumus.

   $y = \alpha + b1X1 + b2X2 + \ldots + bn\,Xn$
4. Pengujian
   Untuk menentukan akurasi prediksi pemodelan Regresi Linear digunakan beberapa parameter diantaranya :

Root Mean Square Error (RMSE)

RMSE mengukur kuadratik dari rata-rata kesalahan atau error pada hasil actual dan predicted saat terjadi pemodelan. Formula RMSE didefinisikan sebagai berikut (Chai & Draxler, 2014) :

$$RMSE = \sqrt{\frac{\sum(Y' - Y)^2)}{n}}$$

Mean Square Error (MSE)

MSE merupakan error yang dikuadratkan, semakin besar error semakin besar juga nilai MSE yang dihasilkan. Cara perhitungan MSE yang dikuadratkan ini membuat data outlier sangat memiliki sensitivitas. Formula MSE didefinisikan sebagai berikut

$$MSE = \frac{\sum(Y' - Y^2)^2}{n}$$

Y' = Prediksi
Y = Aktual
n = Jumlah data

Mean Absolute Error (MAE)

MAE menentukan rata-rata kesalahan atau error pada hasil actual dan predicted saat pemodelan menggunakan Regresi Linear.

$$MAE = \frac{\sum|Y' - Y|^2}{n}$$

Y' = Prediksi
Y = Aktual
n = Jumlah data

R-Squared (R2)

R-squared bukan error namun metrik yang populer yang merepresentasikan sejauh mana data cocok dengan garis regresi yang didapatkan. Semakin besar R-squared akan semakin baik pencocokan garis terhadap data. Nilai terbaik adalah 1.0 dan dapat bernilai negatif.

3. HASIL DAN PEMBAHASAN

Data dalam penelitian ini diambil dari situs https://www.kaggle.com/ dengan judul dataset Madrid Houses Clean. Dataset terdiri dari 17 kolom dan 21739 baris.

	Unnamed: 0	id	sq_mt_built	n_rooms	n_bathrooms	n_floors	sq_mt_allotment	floor	buy_price
0	0	21742	64.0	2	1	1	0.0	3	85000
1	1	21741	70.0	3	1	1	0.0	4	129900
2	2	21740	94.0	2	2	1	0.0	1	144247
3	3	21739	64.0	2	1	1	0.0	-1	109900
4	4	21738	108.0	2	2	1	0.0	4	260000

is_renewal_needed	has_lift	is_exterior	energy_certificate	has_parking	neighborhood	district	house_type
False	False	True	4	False	135	21	1
True	True	True	0	False	132	21	1
False	True	True	0	False	134	21	1
False	True	True	0	False	134	21	1
False	True	True	0	True	133	21	1

Dari 17 variabel atau kolom hanya diambil 6 variabel diantaranya 5 variabel X atau dependen dan 1 variabel Y atau independen.

1. Luas tanah dan bangunan (sq_mt_built)
2. Banyaknya kamar (n_rooms)
3. Banyaknya kamar mandi (n_bathrooms)
4. Banyaknya lantai atau tinggi bangunan (n_floor)
5. Tipe rumah (house_type)
   Terdapat 5 kode tipe rumah dari mulai tipe 1 sampai tipe 5.

Lalu buy_price atau harga rumah merupakan variabel independen. Variabel ini bertipe integer dan harga dalam US Dollar.

	sq_mt_built	n_rooms	n_bathrooms	n_floors	buy_price	house_type
0	64.0	2	1	1	85000	1
1	70.0	3	1	1	129900	1
2	94.0	2	2	1	144247	1
3	64.0	2	1	1	109900	1
4	108.0	2	2	1	260000	1
...	...	...	...	...	...	...
21734	78.0	2	2	1	350000	5
21735	96.0	2	2	1	425000	1
21736	175.0	4	2	1	680000	1
21737	289.0	4	3	3	695000	2
21738	72.0	2	2	1	424000	1

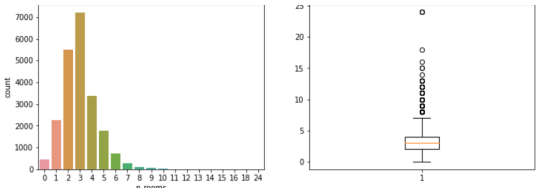
21739 rows × 6 columns

Selanjutnya dilakukan persiapan data untuk mengetahui statistik data dan missing values.

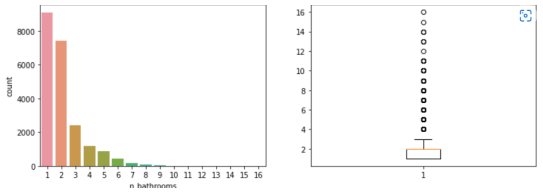
```
#Mencari dan menangani missing values
df.isnull().sum()
```

```
sq_mt_built     0
n_rooms         0
n_bathrooms     0
n_floors        0
buy_price       0
house_type      0
dtype: int64
```

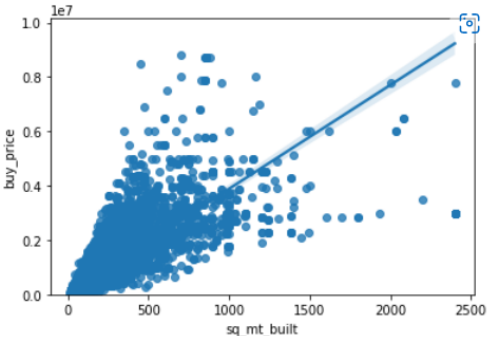
Berdasarkan hasil diatas dapat kita lihat bahwa data yang digunakan tidak memiliki data yang null atau kosong sehingga pemrosesan data dapat dilanjutkan ke proses selanjutnya. Data yang sudah bersih divisualisasikan dalam bentuk boxplot dan count plot atau histogram untuk mengetahui persebaran datanya.



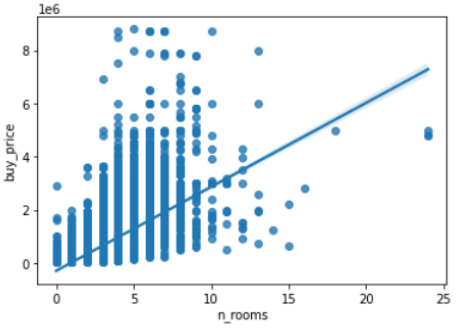
Berdasarkan gambar boxplot dan diagram diatas dapat kita lihat bahwa outlier data n_rooms tidak terlalu banyak.



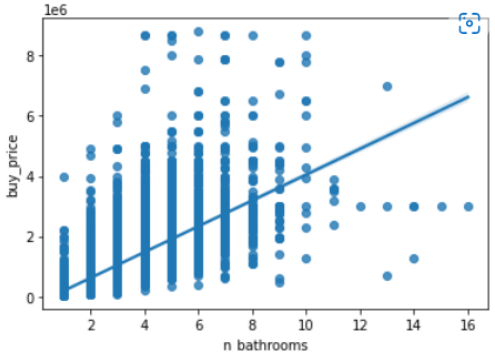
Berdasarkan gambar boxplot dan diagram diatas dapat kita lihat bahwa outlier data n_bathrooms tidak terlalu banyak. Begitu juga untuk data n_floor dan house_type tidak memiliki banyak outlier sehingga tidak akan berpengaruh pada hasil regresi data.



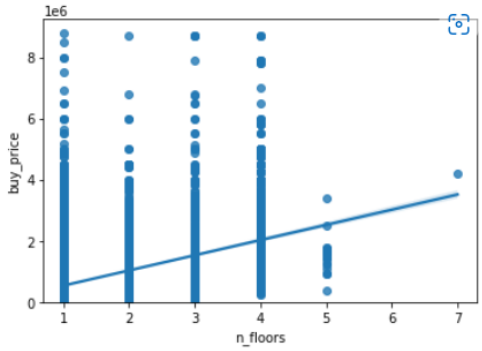
Grafik diatas merupakan grafik korelasi antara sq_mt_built dengan buy_price. Grafik tersebut menunjukan bahwa sq_mt_built berkorelasi kuat dengan buy_price. Sehingga ukuran bangunan sangat mempengaruhi besar harga rumah.



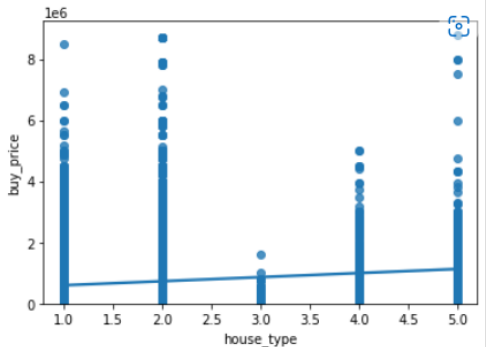
Grafik diatas merupakan grafik korelasi antara n_rooms dengan buy_price. Grafik tersebut menunjukan bahwa n_rooms berkorelasi kuat dengan buy_price. Sehingga jumlah kamar mempengaruhi besar nilai harga rumah.



Grafik diatas merupakan grafik korelasi antara n_bathrooms dengan buy_price. Grafik tersebut menunjukan bahwa n_bathrooms berkorelasi kuat dengan buy_price. Sehingga jumlah kamar mandi mempengaruhi besar nilai harga rumah.



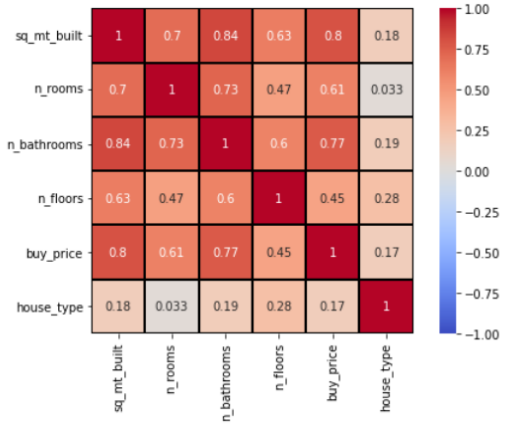
Grafik diatas merupakan grafik korelasi antara n_floors dengan buy_price. Grafik tersebut menunjukan bahwa n_floors berkorelasi lemah dengan buy_price. Sehingga jumlah lantai atau tinggi bangunan kurang mempengaruhi besar nilai harga rumah.



Grafik diatas merupakan grafik korelasi antara house_type dengan buy_price. Grafik tersebut menunjukan bahwa house_type berkorelasi lemah dengan buy_price. Sehingga tipe rumah kurang mempengaruhi besar nilai harga rumah.

Sehingga didapatkan bahwa nilai ukuran bangunan, jumlah kamar, dan jumlah kamar mandi sangat mempengaruhi besar nilai harga rumah. sedangkan data jumlah lantai pada rumah dan tipe rumah tidak mempengaruhi besar harga rumah.

Setelah melihat korelasi dari variabel dependen dan variabel independen. Selanjutnya kita dapat melihat keseluruhan nilai korelasi dengan menggunakan visualisasi heatmap.



Dari gambar diatas kita dapat melihat nilai korelasi antar semua variabel. Kita dapat melihat juga korelasi antara variabel independen dan variabel dependen. Berdasarkan gambar diatas korelasi antara sq_mt_built dan buy_price bernilai 0.8, n_rooms dengan buy_price bernilai 0.61, n_bathrooms dengan buy_price bernilai 0.77, n_floors dengan buy_price bernilai 0,45, dan house_type dengan buy_price bernilai 0,17. Dengan data diatas kita dapat mengetahui bahwa variabel yang paling berpengaruh terhadap variabel independen adalah sq_mt_built, n_rooms dan n_bathrooms.

Selain dengan menggunakan heatmap, kita juga dapat melihat nilai korelasi yang lebih mendetail dengan seperti gambar dibawah ini.

	sq_mt_built	buy_price
sq_mt_built	1.000000	0.804521
buy_price	0.804521	1.000000

Atau dapat dilihat juga dengan menggunakan korelasi pearson.

```
pearson_coef, p_value = stats.pearsonr(df['sq_mt_built'], df['buy_price'])
print("The Pearson Correlation Coefficient is", pearson_coef)
```

```
The Pearson Correlation Coefficient is 0.80452074637963
```

Dengan menggunakan Pearson Correlation Coefficient didapatkan hasil yang sama dengan perhitungan korelasi sebelumnya, nilai Pearson Correlation dari variabel sq_mt_built dengan buy_price bernilai 0.80452074637963.

Setelah itu kita lakukan training dan testing data dengan porsi 80:20 dan setelah itu menghitung nilai coefficient dan intercept.

```
#Kita coba buat kedalam dataframe agar kebih rapi
coef_dict = {
    'features': x.columns,
    'coef_value':lin_reg.coef_
}
coef = pd.DataFrame(coef_dict, columns=['features', 'coef_value'])
coef
```

	features	coef_value
0	sq_mt_built	2712.928905
1	n_rooms	8395.946603
2	n_bathrooms	200236.809132
3	n_floors	-166396.138341
4	house_type	34368.540708

```
print('Coefficients : ', lin_reg.coef_)
print('Intercept : ' , lin_reg.intercept_)

Coefficients :  [   2712.9289054    8395.94660284  200236.80913223 -166396.13834074
   34368.5407077 ]
Intercept :  -51923.824414750794
```

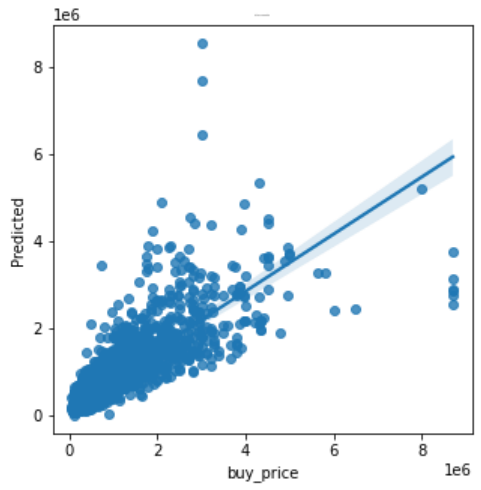
Setelah didapatkan Coeficient dan intercept seperti gambar diatas, maka dapat kita tuliskan persamaan regresinya. Sesuai dengan persamaan pada Multiple Linear Regression maka persamaan regresi untuk data ini adalah

y = 2712.9289054X1 + 8395.94660284X2 + 200236.80913223X3 - 166396.13834074X4 + 34368.5407077X5 -51923.824414750794

PENGUJIAN

Proses pengujian prediksi dilakukan dengan mengunakan data predict. Berikut adalah detail antara nilai asli dan nilai predict yang digunakan untuk proses pengujian.

	Actual	Predicted
8406	240000	258507.539325
5705	168000	206704.730236
5804	345000	534963.517695
3347	320000	217299.285971
17477	490000	361341.677844




```
#Prediksi harga rumah
#Sq_mt_built = 100
#n_rooms = 2
#n_bathrooms = 2
#n_floors = 1
#house_type = 1
lin_reg.predict([[100,2,2,1,1]])

/usr/local/lib/python3.7/dist-packag
  "X does not have valid feature nam
array([504606.97996271])
```

Dari gambar diatas kita misalkan ingin memprediksikan harga rumah dengan sq_mt_built 100, banyaknya kamar 2, kamar mandi 2, dan jumlah lantai 1, dengan tipe 1 maka akan didapatkan harga senilai 504.606, 97996271.

```
#Prediksi harga rumah
#Sq_mt_built = 70
#n_rooms = 3
#n_bathrooms = 1
#n_floors = 1
#house_type = 1
lin_reg.predict([[70.0,3,1,1,1]])

/usr/local/lib/python3.7/dist-packages/skle
  "X does not have valid feature names, but
array([231378.25027121])
```

Pengujian kedua, program diuji dengan menginputkan sq_mt_built 70, n_room 3, n_bathroooms 1, n_floors 1 dan house_type 1 dan dihasilkan buy_price sebesar 231378,25027121.

```
[19] from sklearn.metrics import r2_score

     test_y_ = lin_reg.predict(x_test)

     print("Mean absolute error (MAE): %.2f" % np.mean(np.absolute(test_y_ - y_test)))
     print("Residual sum of squares (MSE): %.2f" % np.mean((test_y_ - y_test) ** 2))
     print("Root Mean Squared Error (RMSE) : %.2f" % np.sqrt(np.mean((test_y_ - y_test) ** 2)))
     print("R2-score: %.2f" % r2_score(test_y_ , y_test) )

     Mean absolute error (MAE): 226806.92
     Residual sum of squares (MSE): 205299166576.75
     Root Mean Squared Error (RMSE) : 453099.51
     R2-score: 0.50
```

Didapatkan hasil MAE 226806,92 , MSE 205299166576,74 , RMSE 453099,51 dan nilai R2_score 0,50. Dengan nilai R2-score 0,5 dapat kita ketahui bahwa masih terdapat banyak error antara garis regresi dengan titik-titik data.

4. KESIMPULAN

Berdasarkan beberapa pengujian pada program prediksi harga rumah dengan pemodelan Regresi Linear didapatkan bahwa error dari perhitungan prediksi menggunakan Regresi Linear masih cukup besar dan masih terdapat banyak error antara titik data dengan gari regresi.

REFERENSI

[1] Andi Saiful, Septi Andryana, Aris Gunaryati. 2021, "*Prediksi Harga Rumah Menggunakan Web Scraping Dan Machine Learning Dengan Algoritma Linear Regression*", Jurnal Teknik Informatika dan Sistem Informasi, Vol. 8, No. 1, pp. 41.

[2] Evi Februion, Feny Novia, Yufis Azhar. 2021. "*Prediksi Harga Rumah Menggunakan General Regression Neural Network",* JURNAL INFORMATIKA, Vol.8 No.1 pp.59.

[3] Sholeh Muhammad, Yuliana Rachmawati, Eko Nur. 2022, "*Penerapan Regresi Linear Ganda Untuk Memprediksi Nilai Kuesioner Mahasiswa Dengan Menggunakan Phyton*", Jurnal Dinamika Informatika, Volume 11, No 1.

[4] Reza Mahendra, Anton Siswo Raharjo, Rifki Wijaya. 2022. "Prediksi Harga Rumah Di Kota Bandung Bagian Timur Dengan Menggunakan Metode Regresi". e-Proceeding of Engineering : Vol.7, No.3.