Analisis Perbandingan Metode Simple Linear Regression, Multiple Linear Regression, dan Random Forest Regression dalam Memprediksi Harga Rumah

Catherine Olivia
Program Studi Sistem Informasi
Fakultas Teknik dan Informatika
Universitas Multimedia Nusantara
Tangerang, Banten
E-mail: catherine.olivia1@student.umn.ac.id

Abstract - Home is one of the basic needs in human life. Many people flock to buy a house. However, problems occur when finances are not sufficient for the price of a house. Furthermore, house prices continue to increase every year. Many factors influence the price increase. Therefore, this research aims to predict house prices based on existing factors such as house size, number of bedrooms, number of bathrooms, and house age using Machine Learning algorithms namely Simple Linear Regression, Multiple Linear Regression, and Random Forest Regression. Evaluation is seen based on the R-squared value, Root Mean Squared Error (MSE), and Mean Absolute Error (MAE) of each model. The division of data composition with a division of 80% for training data and 20% for testing data. In this study, the prediction results using the Multiple Linear Regression method produced the highest accuracy rate of 57.83% compared to the Simple Linear Regression and Random Forest Regression methods.

Keywords: House Price Prediction, Machine Learning, Regression, Simple Linear Regression, Multiple Linear Regression, Random Forest Regression, Comparison.

Abstrak – Rumah menjadi salah satu kebutuhan pokok dalam kehidupan manusia. Banyak orang berbondong-bondong untuk membeli sebuah rumah. Namun, masalah terjadi ketika finansial belum mencukupi harga dari sebuah rumah tersebut. Lebih lanjut, harga rumah yang terus meningkat setiap tahunnya. Banyak faktor yang mempengaruhi peningkatan harga tersebut. Oleh karena itu, penelitian ini bertujuan untuk melakukan prediksi harga rumah berdasarkan faktor-faktor yang ada seperti ukuran luas rumah, jumah kamar tidur, jumlah kamar mandi, dan umur rumah dengan menggunakan algoritma Machine Learning yaitu Simple Linear Regression, Multiple Linear Regression, dan Random Forest Regression. Evaluasi dilihat berdasarkan nilai R-squared, Root Mean Squared Error (MSE), dan Mean Absolute Error (MAE) masing-masing model. Adapun pembagian komposisi data dengan pembagian 80% untuk training data dan 20% untuk testing data. Pada penelitian ini, hasil prediksi menggunakan metode Multiple Linear Regression menghasilkan tingkat akurasi yang paling tinggi sebesar 57.83% dibandingkan dengan metode Simple Linear Regression dan Random Forest Regression.

Keywords: Prediksi Harga Rumah, Machine Learning, Regression, Simple Linear Regression, Multiple Linear Regression, Random Forest Regression, Perbandingan.

I. PENDAHULUAN

Rumah merupakan salah satu kebutuhan pokok mendasar yang tidak dapat dipisahkan dalam kehidupan manusia. Rumah merupakan kebutuhan primer, sebagai tempat berlindung dan tempat beristirahat dari penatnya aktivitas seharian [3]. Setiap tahunnya kebutuhan akan rumah terus meningkat. Banyak orang yang berlomba-lomba untuk membeli sebuah rumah yang sesuai dengan kebutuhan mereka. Terkait hal tersebut, terdapat faktor-faktor yang mempengaruhi setiap manusia dalam mengambil keputusan untuk membeli sebuah rumah, salah satunya adalah faktor finansial [1]. Finansial merupakan salah satu aspek yang mempengaruhi keberlanjutan kehidupan seseorang yang diantaranya adalah memiliki tempat tinggal yang nyaman [2]. Finansial berhubungan erat dengan perencanaan dan pengelolaan keuangan seseorang.

Lebih lanjut, indikator finansial yang menjadi pertimbangan banyak orang adalah harga sebuah rumah. Seiring berjalannya waktu, harga rumah terus meningkat dan berubah-ubah setiap tahunnya. Banyak faktor yang menjadi penyebab peningkatan harga rumah tersebut. Oleh karena itu, peneliti ingin berfokus pada penelitian terkait masalah harga rumah berdasarkan faktor yang ada. Penelitian ini bertujuan untuk memberikan informasi kepada masyarakat terkait prediksi harga rumah di masa yang akan datang untuk persiapan finansial setiap orang.

Prediksi tentang harga sebuah rumah akan dilakukan dengan menggunakan *machine learning*. Algoritma *machine learning* merupakan salah satu metode jaringan syaraf tiruan yang sering digunakan untuk memprediksi data [4]. Algoritma dari machine learning memiliki beberapa jenis yaitu *supervised learning, unsupervised learning, semi-supervised learning dan reinforcement learning* [5]. Peneliti akan menggunakan jenis algoritma *supervised learning* untuk memprediksi harga rumah. Algoritma *supervised learning* adalah algoritma yang bergantung pada data input berlabel untuk mempelajari fungsi yang menghasilkan output yang sesuai ketika diberi data baru tanpa label [6]. Salah satu jenis dari algoritma supervised learning adalah analisa regresi. Analisa regresi merupakan salah satu alat analisis statistika yang memanfaatkan hubungan antara dua variabel atau lebih dengan tujuan membuat perkiraan atau prediksi untuk nilai suatu variabel, jika nilai variabel lain berhubungan dengannya diketahui [7]. Terdapat banyak jenis analisa regresi yang dapat digunakan untuk melakukan prediksi, seperti *Simple Linear Regression, Multiple Linear Regression,* dan *Random Forest Regression*. *Simple Linear Regression Analisis* adalah analisis regresi linear yang hanya melibatkan dua variabel, yaitu satu variabel independen dan satu variabel dependen [7]. *Multiple Linear Regression* adalah analisis regresi yang menjelaskan hubungan antara peubah respon (variabel dependen) dengan faktor-faktor yang mempengaruhi lebih dari satu prediktor (variabel independen) [8]. Sedangkan, *Random Forest*

Regression adalah suatu algoritma yang sering digunakan pada klasifikasi data dengan jumlah besar karena tingkat akurasi dari prediksinya yang tinggi, dan berdasarkan pada banyaknya pohon [9]. Peneliti memilih ketiga jenis tersebut untuk digunakan dalam memprediksi harga rumah sebagai solusi terkait masalah yang menjadi fokus peneliti.

II. METODE PENELITIAN

A. Sumber Data

Data yang digunakan dalam penelitian kali ini adalah data sintetis yang merupakan kumpulan data harga-harga rumah dalam bentuk USD yang didapat bersumber dari kaggle yang bersifat open source dengan jumlah 6 kolom dan 50.000 baris melalui link berikut. https://www.kaggle.com/datasets/muhammadbinimran/housing-price-prediction-data.

B. Metode Penelitian

Peneliti menggunakan empat tahap dalam melakukan penelitian yaitu penarikan dataset, data preprocessing, pengujian model prediksi, dan evaluasi. Berikut adalah penjelasan dari masing-masing tahapan yang dilakukan oleh peneliti :

1. Penarikan dataset. Dataset yang digunakan dalam melakukan penelitian adalah dataset prediksi harga rumah yang bersumber dari kaggle dengan rincian 6 kolom dan 50.000 baris data.
2. Data *Preprocessing*. Pre-processing data adalah langkah penting untuk mencapai kinerja klasifikasi yang baik sebelum mengevaluasi data pada algoritma Machine Learning [13]. Pada tahap ini, peneliti melakukan pemrosesan data yang dimulai dari mengatasi *missing values*, mengatasi *outliers*, menormalisasikan data, melakukan *formatting*, melakukan *binning*, melakukan *encoding*, hingga melakukan *grouping*.
3. Data *Visualisation*. Visualisasi data adalah proses penyajian data dalam bentuk grafik yang membuat informasi mudah dimengerti, hal ini membantu menjelaskan tentang fakta dan menentukan arah tindakan [14]. Dalam tahap ini, peneliti memberikan visualisasi data terkait dengan berbagai jenis visualisasi menggunakan *heatmap, pie chart, boxplot, histograms*, dan *scatter plot.*
4. Pengujian Model Prediksi. Kemudian, tahap modelling yaitu melakukan prediksi data dengan menggunakan tiga jenis model prediksi yaitu *Simple Linear Regression, Multiple Linear Regression,* dan *Random Forest Regression.*
5. Evaluasi. Tahapan terakhir yang dilakukan peneliti adalah tahapan evaluasi. Evaluasi model atau proses mengukur seberapa baik kinerja model dalam memprediksi nilai target dari data testing pada penelitian ini dilakukan dengan menggunakan confusion matrix [13]. Pada proses evaluasi ini, ketiga model prediksi yang sudah diuji akan dibandingkan dengan mengukur *R-squared, Residual Mean Squared Error (MSE), dan Mean Absolute Error (MAE)* masing-masing model sehingga dapat dibandingkan untuk mendapatkan hasil kesimpulan model prediksi mana yang cukup baik digunakan. RMSE adalah perhitunganakar darikuadraterror (data aktual -data prediksi)yang dibagi dengan jumlah data. [16]. Sedangkan, *Mean Absolute Error adalah metode untuk mengevaluasi metode peramalan menggunakan jumlah dari kesalahan-kesalahan yang absolut* [15].

C. Metode Pengujian Dataset

Penelitian ini melakukan perbandingan dengan menggunakan tiga metode algoritma, diantaranya metode *Simple Linear Regression* yang digunakan untuk mengetahui pengaruh antara satu buah variabel bebas terhadap satu buah variabel terikat [10]. Bentuk umum dari persamaan *Simple Linear Regression* sebagai berikut:

$$\bar{y} = a + bx$$

Keterangan: $\bar{y}$ = Nilai yang diramalkan, x = Variabel bebas, a = Parameter Intercept yaitu perpotongan dengan sumbu vertikal, b = Parameter slope koefisien regresi untuk variabel bebas.

Dengan Y adalah variabel terikat dan X adalah variabel bebas. Koefisien a adalah konstanta (intercept) yang merupakan titik potong antara garis regresi dengan sumbu Y pada kordinat kartesius [10].

Selanjutnya jenis metode algoritma yang digunakan adalah *Multiple Linear Regression*. Model *Multiple Linear Regression* merupakan metode yang menggunakan jumlah variabel bebas lebih dari satu dan satu variabel terikat untuk mengetahui pengaruh antar variabel [11]. Adapun persamaannya sebagai berikut :

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn$$

Dengan Y = prediksi nilai variabel dependen, X = variabel independen, β0= konstanta, βn= Bobot (koefisien) regresi untuk variabelindependent [2].

Kemudian, perbandingan lainnya juga dilakukan dengan metode *Random Forest Regression. Random Forest* merupakan suatu set dari decision trees yang dibangun dengan sampel yang dipilih secara acak tetapi memiliki peraturan membelah simpul yang berbeda [12]. Berikut rumus dari *Random Forest*:

$$f(x) = Average(f_1(x), f_2(x), \ldots, f_n(x))$$

Dengan keterangan, $f(x)$ = hasil prediksi, $f_1(x), f_2(x), \ldots, f_n(x)$ = hasil prediksi dari setiap pohon keputusan, dan x = input [13].

D. Metode Pengolahan Dataset

Preprocessing merupakan tahap yang penting dlaam penambangan data [19]. Adapun *preprocessing* dataset yang dilakukan oleh peneliti dengan detail sebagai berikut:

1. Proses Awal sebagai tahapan awal yang dilakukan adalah pembacaan file dataset.

```
# Membaca data
hp = pd.read_csv('housing_price_dataset.csv')
hp.head (10)
```

	SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price
0	2126	4	1	Rural	1969	215355.283618
1	2459	3	2	Rural	1980	195014.221626
2	1860	2	1	Suburb	1970	306891.012076
3	2294	2	1	Urban	1996	206786.787153
4	2130	5	2	Suburb	2001	272436.239065
5	2095	2	3	Suburb	2020	198208.803907
6	2724	2	1	Suburb	1993	343429.319110
7	2044	4	3	Rural	1957	184992.321268
8	2638	4	3	Urban	1959	377998.588152
9	1121	5	2	Urban	2004	95961.926014

2. *Handling Missing Values*. *Missing* Values merupakan data yang hilang [17]. Peneliti tidak melakukan proses handling missing values karena di dalam data tidak terdapat values yang hilang.

```
hp.isnull().sum()
# Tidak ada missing values
```

```
SquareFeet      0
Bedrooms        0
Bathrooms       0
Neighborhood    0
YearBuilt       0
Price           0
dtype: int64
```

3. *Encoding*. *Encoding* adalah metode untuk menyiapkan data sebelum diberikan kepada suatu model [20]. Peneliti melakukan encoding dalam persiapan data dengan cara membuat kolom baru yang bernama Neighborhood Code yang merupakan bentuk numerik dari data kategorikal kolom Neighborhood. Dimana numerik ini memiliki index dari 0-2, dengan keterangan 0 = rural, 1 = subrural, dan 2 = urban.

```
hp['Neighborhood'] = hp['Neighborhood'].astype('category')
hp['Neighborhood code'] = hp['Neighborhood'].cat.codes
hp.head()
```

	SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price	YearNow	Age	Neighborhood code
0	2126	4	1	Rural	1969	215355.283618	2023	54	0
1	2459	3	2	Rural	1980	195014.221626	2023	43	0
2	1860	2	1	Suburb	1970	306891.012076	2023	53	1
3	2294	2	1	Urban	1996	206786.787153	2023	27	2
4	2130	5	2	Suburb	2001	272436.239065	2023	22	1

4. *Binning*. Binning adalah proses data numerik yang berlanjut dibagi-bagi menjadi input kategorikal [18]. Peneliti melakukan dua kali proses *binning*. Pertama, melakukan *binning* pada kolom *Price* dengan membuat kolom baru bernama *Price Category*, dimana jika *Price* > 224000 = mahal dan jika *price* < 224000 = murah.

```
hp['Price Category']=np.where(
hp['Price'] > 224000, 'mahal', np.where(
hp['Price'] < 224000, 'murah', -1))
hp.head()
```

	SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price	YearNow	Age	Neighborhood code	Price Category
0	2126	4	1	Rural	1969	215355.283618	2023	54	0	murah
1	2459	3	2	Rural	1980	195014.221626	2023	43	0	murah
2	1860	2	1	Suburb	1970	306891.012076	2023	53	1	mahal
3	2294	2	1	Urban	1996	206786.787153	2023	27	2	murah
4	2130	5	2	Suburb	2001	272436.239065	2023	22	1	mahal

Kedua, *binning* dilakukan pada kolom *Age* dengan membuat kolom baru bernama *Age Category* dimana *Age* sendiri akan dibagi menjadi 3 kategori dengan menggunakan library *numpy*. *Binning* ini dilakukan untuk mengetahui kategori rumah yang sudah tua, masih muda, atau baru.

```
bins = np.linspace(min(hp['Age']),
                   max(hp['Age']), 4)
group_names = ['Baru','Muda','Tua']

hp['Age Category'] = pd.cut(hp['Age'],
                            bins,
                            labels = group_names,
                            include_lowest = True)
hp.head(5)
```

	SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price	YearNow	Age	Neighborhood code	Price Category	Age Category
0	2126	4	1	Rural	1969	215355.283618	2023	54	0	murah	Tua
1	2459	3	2	Rural	1980	195014.221626	2023	43	0	murah	Muda
2	1860	2	1	Suburb	1970	306891.012076	2023	53	1	mahal	Tua
3	2294	2	1	Urban	1996	206786.787153	2023	27	2	murah	Muda
4	2130	5	2	Suburb	2001	272436.239065	2023	22	1	mahal	Baru

5. *Formatting*. Peneliti melakukan *formatting* dengan mengubah data type *Price Category* yang bertipe objek menjadi *category*.

```
hp["Price Category"] = hp["Price Category"].astype("category")
hp.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   SquareFeet          50000 non-null  int64
 1   Bedrooms            50000 non-null  int64
 2   Bathrooms           50000 non-null  int64
 3   YearBuilt           50000 non-null  int64
 4   YearNow             50000 non-null  int64
 5   Age                 50000 non-null  int64
 6   Age Category        50000 non-null  category
 7   Neighborhood        50000 non-null  category
 8   Neighborhood code   50000 non-null  int8
 9   Price               50000 non-null  float64
 10  Price Category      50000 non-null  category
dtypes: category(3), float64(1), int64(6), int8(1)
memory usage: 2.9 MB
```

Selain itu, peneliti juga melakukan *formatting* data kolom *Price* dengan mengambil 2 angka di belakang koma untuk mempermudah proses analisa.

```
hp['Price'] = hp['Price'].apply(lambda x: round(x, 2))
hp.head()
```

	SquareFeet	Bedrooms	Bathrooms	YearBuilt	YearNow	Age	Age Category	Neighborhood	Neighborhood code	Price	Price Category
0	2126	4	1	1969	2023	54	Tua	Rural	0	215355.28	murah
1	2459	3	2	1980	2023	43	Muda	Rural	0	195014.22	murah
2	1860	2	1	1970	2023	53	Tua	Suburb	1	306891.01	mahal
3	2294	2	1	1996	2023	27	Muda	Urban	2	206786.79	murah
4	2130	5	2	2001	2023	22	Baru	Suburb	1	272436.24	mahal

6. Normalisasi. Peneliti melakukan normalisasi pada kolom numerik dengan menggunakan ZScore karena dataset tidak berdistribusi dengan normal. Selain itu, ZScore juga bersifat fleksibel dan cocok dengan dataset yang tidak normal.

```
# Melakukan normalisasi dengan Zscore karena dataset tidak berdistribusi dengan normal. Selain itu, Zscore juga ber
from sklearn.preprocessing import StandardScaler

HPforNorm = hp[['SquareFeet','Bedrooms','Bathrooms', 'YearBuilt', 'YearNow', 'Age', 'Neighborhood code', 'Price']]
SSc = StandardScaler()
hp_norm = SSc.fit_transform(HPforNorm)
for i in range(len(hp_norm)): #data yang sudah dinormalisasikan masuk ke dalam kolomnya masing-masing di HPforNorm
    HPforNorm.loc[i, "SquareFeet"] = hp_norm[i, 0]
    HPforNorm.loc[i, "Bedrooms"] = hp_norm[i, 1]
    HPforNorm.loc[i, "Bathrooms"] = hp_norm[i, 2]
    HPforNorm.loc[i, "YearBuilt"] = hp_norm[i, 3]
    HPforNorm.loc[i, "YearNow"] = hp_norm[i, 4]
    HPforNorm.loc[i, "Age"] = hp_norm[i, 5]
    HPforNorm.loc[i, "Neighborhood code"] = hp_norm[i, 6]
    HPforNorm.loc[i, "Price"] = hp_norm[i, 7]

HPforNorm.head()
```

	SquareFeet	Bedrooms	Bathrooms	YearBuilt	YearNow	Age	Neighborhood code	Price
0	0.297861	0.449067	-1.220113	-0.791751	0	0.791751	-1.223957	-0.124401
1	0.786480	-0.446738	0.005614	-0.260842	0	0.260842	-1.223957	-0.391551
2	-0.254340	-1.342543	-1.220113	-0.743486	0	0.743486	0.001790	1.077785
3	0.499777	-1.342543	-1.220113	0.511390	0	-0.511390	1.227536	-0.236936
4	0.214811	1.344872	0.005614	0.752713	0	-0.752713	0.001790	0.625272

7. *Handling Outliers*. Penanganan data *outliers* menggunakan metode IQR karena distribusi dataset yang tidak normal. Data yang terdapat *outlier* adalah data *Price* sehingga data *outlier* tersebut ditangani oleh IQR dan menghasilkan data bersih dengan shape *(49941, 11)*.

```
# Penanganan Outlier dengan menggunakan IQR
Q1 = HP["Price"].quantile(0.25) # mencari Q1
Q3 = HP["Price"].quantile(0.75) # mencari Q3
IQR = Q3 - Q1  # mencari IQR

# data outlier
outlier_indices = HP[((HP["Price"] < (Q1-1.5*IQR)) | (HP["Price"] > (Q3+1.5*IQR)))].index

# menghapus row data yang outlier
HP.drop(outlier_indices, inplace=True)

# menampilkan jumlah baris dan kolom HP yang sudah bersih
HP.shape
```

```
(49941, 11)
```

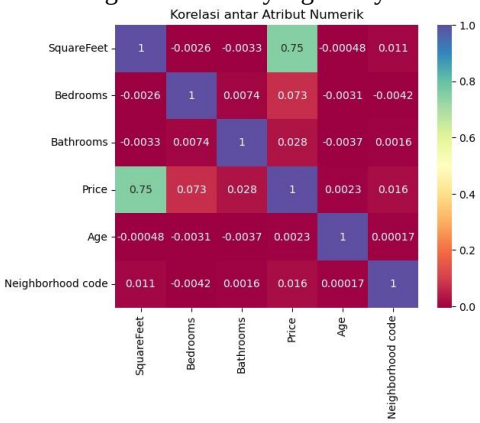
8. *Grouping*, dilakukan untuk mengetahui komposisi rumah yang diwakilkan oleh umur rumah (*Age*) berdasarkan kategori rumah (*Age Category*) yang dikelompokkan menjadi 3 bagian, yaitu kategori rumah Baru, Muda, dan Tua.

```
# Detail
hp.groupby('Age Category')['Age'].count()
```

```
Age Category
Baru    16523
Muda    16745
Tua     16732
Name: Age, dtype: int64
```

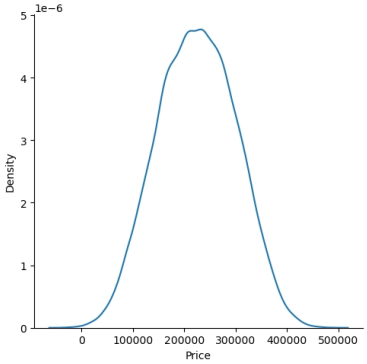
E. *Visualisasi Data*

1. Korelasi Enam Atribut Numerik
   Visualisasi di bawah ini menggambarkan korelasi antar 6 atribut numerik dataset *Housing Price*. Berdasarkan visualisasi tersebut, pembaca dapat mengetahui bahwa atribut *SquareFeet* memiliki hubungan yang paling kuat dengan *Price* karena hasil korelasinya sebesar 0.75 yang paling mendekati angka 1 diantara yang lainnya.

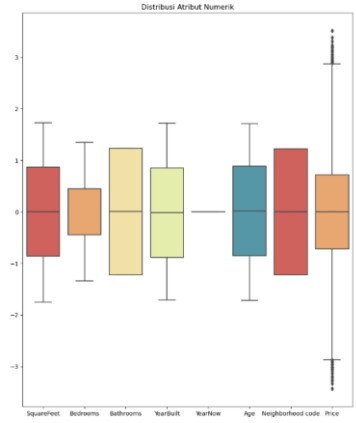


2. Distribusi Jumlah Rumah berdasarkan Kategori Rumah
   Berikut merupakan visualisasi data yang menggambarkan komposisi jumlah rumah yang diwakilkan oleh umur rumah (*Age*) berdasarkan kategori rumah (*Age Category*) yang dikelompokkan menjadi 3 bagian. Berdasarkan visualisasi ini, dapat disimpulkan bahwa 33,05% termasuk ke dalam kategori rumah Baru, 33,49% termasuk ke dalam kategori rumah Muda, dan sisanya yang sebesar 33,46% termasuk ke dalam kategori rumah Tua.

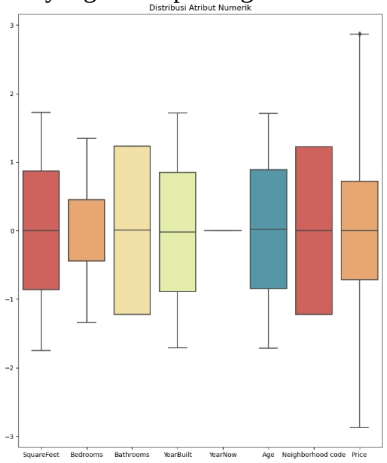
3. Distribusi Normalisasi Dataset *Housing Price*
   Berikut merupakan visualisasi data yang menggambarkan distribusi dari dataset *Housing Price*. Visualisasi ini menunjukkan bahwa data tidak berdistribusi dengan normal karena terdapat dua puncak dibagian atas bel.



4. Boxplot yang Menampilkan *Outliers*
   Visualisasi dibawah ini merupakan visualisasi dari dataset yang menampilkan data *outlier* setelah dinormalisasikan dan dapat terlihat seluruh boxplot tiap kolom muncul dalam visualisasi. Berdasarkan visualisasi tersebut, dapat diketahui bahwa kolom yang memiliki *outliers* adalah kolom *Price*.

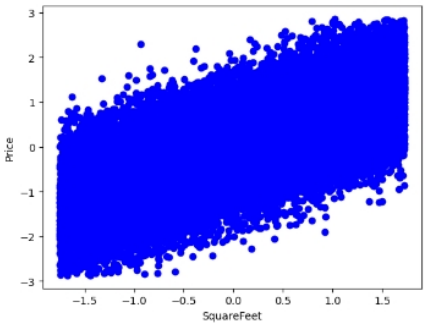


5. Boxplot setelah Data *Outliers* ditangani
   Berikut adalah visualisasi dari dataset yang numerik setelah dilakukan penanganan *outlier*. Visualisasi tersebut menunjukkan bahwa *outlier* dari kolom *Price* telah hilang dan meninggalkan sedikit sisa data *outlier* yang masih penting.



6. *Scatter Plot* Harga Rumah berdasarkan Luas Rumah
   Gambar dibawah ini merupakan visualisasi menggunakan *scatter plot* yang menggambarkan

korelasi hubungan antara *SquareFeet* (luas rumah) dengan *Price* (harga rumah).



III. HASIL DAN PEMBAHASAN

Pada proses pengujian, ketiga dataset masing-masing diuji dengan model dan proses yang sama. Hasil dari *processing* dataset dilakukan pembagian data dengan jumlah 80 : 20. 80% data digunakan untuk data *training*. Sedangkan, 20% data digunakan untuk data *testing*.

A. *Pengujian model prediksi dengan Simple Linear Regression*

Simple Linear Regression dibuat dengan tujuan memprediksi harga rumah (*Price*) berdasarkan luas rumah (*SquareFeet*).

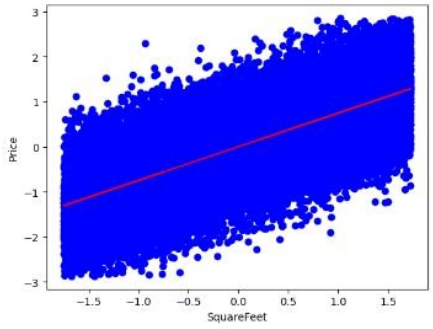
1. *Modeling*

Modeling Simple Linear Regression menggunakan data train x dan y dengan komposisi 80% dari masing-masing data itu sendiri. Modeling ini menghasilkan nilai koefisien yang menghitung hubungan antara variabel SquareFeet dengan Price, dimana menghasilkan angka sebesar 0.75 mendekati 1 yang berarti hubungan antara kedua variabel tersebut kuat. Sedangkan nilai intercept digunakan untuk menghitung rata-rata variabel Price jika variabel independen SquareFeet bernilai 0, yang artinya berdasarkan hasil tersebut nilai variabel Price akan bernilai 0.00017 jika nilai independen SquareFeet-nya 0.

```
# modeling
from sklearn import linear_model
regr = linear_model.LinearRegression()
train_x = np.asanyarray(train[['SquareFeet']])
train_y = np.asanyarray(train[['Price']])
regr.fit(train_x,train_y)
print('Coefficients:', regr.coef_)
print('Intercept: ', regr.intercept_)
```

```
Coefficients: [[0.74683489]]
Intercept:  [0.00017118]
```

Berikut visualisasi dari modeling yang dilakukan:



2. Evaluasi

Evaluasi model *Simple Linear Regression* menggunakan data test x dan test y. Data test x merupakan data independen yang berdiri sendiri tidak dipengaruhi oleh yang lain dengan komposisi 20% dari data itu sendiri. Sedangkan, data test y adalah data dependen yang dipengaruhi oleh data lain dan merupakan variabel data yang ingin diprediksi dengan komposisi 20% juga dari data itu sendiri. Dalam proses ini, test x merupakan data dari variabel SquareFeet, sedangkan test y merupakan data dari variabel Price. Evaluasi ini menghasilkan nilai Mean Absolute Error (MAE) sebesar 52%, Residual Sum of Squares (MSE) sebesar 0.43% dan nilai R-Squared sebesar 57%.

```
# EVALUATION

from sklearn.metrics import r2_score

test_x = np.asanyarray(test[['SquareFeet']])
test_y = np.asanyarray(test[['Price']])
test_y_= regr.predict(test_x)

print('Mean absoulute error: %.2f' % np.mean(np.absolute(test_y_ - test_y)))
print('Residual sum of squares (MSE): %.2f' % np.mean((test_y_ - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y,test_y_))
```

```
Mean absoulute error: 0.52
Residual sum of squares (MSE): 0.43
R2-score: 0.57
```

Selain itu, peneliti juga melakukan uji akurasi model yang menghasilkan akurasi model *Simple Linear Regression* sebesar 56,11%.

```
# Skor akurasi model sebesar 56,11% berdasarkan Simple Regression yang sudah dilakukan
regr.score(train_x, train_y)
```

```
0.5611033495026813
```

3. Uji Coba Prediksi Data lain

Selanjutnya, peneliti juga mencoba untuk memprediksi sebuah data baru dengan nilai dari SquareFeet sebesar 1800 yang menghasilkan prediksi harga rumah sebesar 1344.30 USD.

```
# Prediksi harga rumah impian Beverly
regr.predict([[1800]])
```

```
array([[1344.30298108]])
```

B. *Pengujian model prediksi dengan Multiple Linear Regression*

Multiple Linear Regression dibuat dengan tujuan memprediksi harga rumah (*Price*) berdasarkan luas rumah (*SquareFeet*), jumlah kamar tidur (*Bedrooms*), jumlah kamar mandi (*Bathrooms*), dan umur rumah (*Age*).

1. *Modeling*

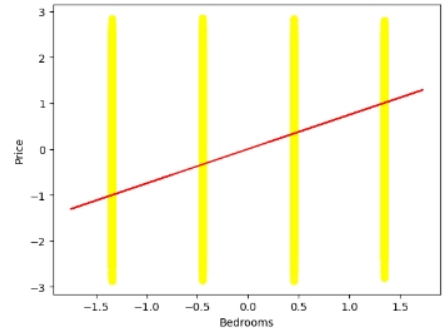
Modeling Multiple Linear Regression menggunakan data train x dan y dengan komposisi 80% dari masing-masing data itu sendiri. Modeling ini menghasilkan nilai koefisien yang menghitung hubungan antara variabel SquareFeet dengan Price, dimana menghasilkan angka sebesar 0.75 dan variabel Bedrooms dengan Price sebesar 0.73 yang keduanya mendekati 1, yang berarti hubungan antara masing-masing kedua variabel tersebut kuat. Selain itu, hasil nilai koefisien Bathrooms dengan Price dan Age dengan Price cenderung tidak kuat karena hasil berturut-turutnya sebesar 0.03 dan 0.002 yang jauh dari angka 1. Sedangkan nilai intercept digunakan untuk menghitung rata-rata variabel Price jika variabel independen SquareFeet, Bedrooms, Bathrooms, dan Age bernilai 0, yang artinya berdasarkan hasil tersebut nilai variabel

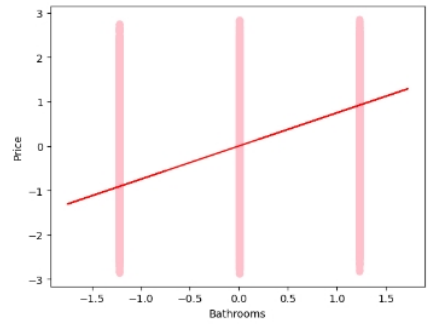
Price akan bernilai -2.6478 jika nilai independen SquareFeet, Bedrooms, Bathrooms, dan Age -nya 0.

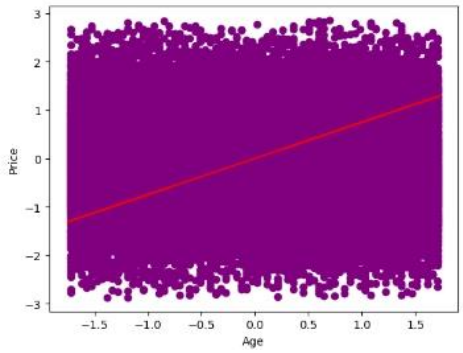
```
from sklearn import linear_model
regr = linear_model.LinearRegression()
x = np.asanyarray(train[['SquareFeet', 'Bedrooms', 'Bathrooms','Age']])
y = np.asanyarray(train[['Price']])
regr.fit(x,y)

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

```
Coefficients:  [[0.74718649 0.07335981 0.03021329 0.0029526 ]]
Intercept:  [-2.6478589e-05]
```

Berikut visualisasi dari modeling yang sudah dilakukan:







2. Evaluasi

Evaluasi model *Multiple Linear Regression* menggunakan data test x dan test y. Data test x merupakan data independen yang berdiri sendiri tidak dipengaruhi oleh yang lain dengan komposisi 20% dari data itu sendiri. Sedangkan, data test y adalah data dependen yang dipengaruhi oleh data lain dan merupakan variabel data yang ingin diprediksi dengan komposisi 20% juga dari data itu sendiri. Dalam proses ini, test x merupakan data dari variabel SquareFeet, Bedrooms, Bathrooms, dan Age, sedangkan test y merupakan data dari variabel Price. Evaluasi ini menghasilkan nilai Mean Absolute Error (MAE) sebesar 52%, Residual Sum of Squares (MSE) sebesar 42% dan nilai R-Squared sebesar 58%.

```
# EVALUATION
y_hat = regr.predict(test[['SquareFeet', 'Bedrooms', 'Bathrooms','Age']])
x = np.asanyarray(test[['SquareFeet', 'Bedrooms', 'Bathrooms','Age']])
y = np.asanyarray(test[['Price']])

print("Residual sum of squares: %.2f" % np.mean((y_hat - y) ** 2))
print("R2-score: %.2f" % r2_score(y, y_hat))
print('Mean absoulute error: %.2f' % np.mean(np.absolute(y_hat - y)))
```

```
Residual sum of squares: 0.42
R2-score: 0.58
Mean absoulute error: 0.52
```

Selain itu, peneliti juga melakukan uji akurasi model yang menghasilkan akurasi model *Simple Linear Regression* sebesar 57,83%.

```
# Skor akurasi model sebesar 57,83% berdasarkan Multiple Regression yang sudah dilakukan
regr.score(x,y)
```

```
0.5783426146065824
```

3. Uji Coba Prediksi Data lain

Selanjutnya, peneliti juga mencoba untuk memprediksi sebuah data baru dengan nilai dari SquareFeet sebesar 1800, Bedrooms sebanyak 4 kamar, Bathrooms sebanyak 2 buah, dan Age 10 tahun, yang menghasilkan prediksi harga rumah sebesar 1345.32 USD.

```
# Prediksi harga rumah impian Beverly
regr.predict([[1800,4,2,10]])
```

```
array([[1345.31905335]])
```

C. Pengujian model prediksi dengan Random Forest Regression

Random Forest Regression dibuat dengan tujuan memprediksi harga rumah (*Price*) berdasarkan variabel luas rumah (*SquareFeet*), jumlah kamar tidur (*Bedrooms*), jumlah kamar mandi (*Bathrooms*), dan umur rumah (*Age*).

1. *Modeling*

Modeling Random Forest Regression menggunakan data train x dan y dengan komposisi 80% dari masing-masing data itu sendiri. Modeling ini menghasilkan nilai Feature Importance yang digunakan untuk memahami hubungan antara variabel SquareFeet, Bedrooms, Bathrooms, dan Age dengan variabel Price. Feature Importance antara SquareFeet dan Price menghasilkan nilai sebesar 0.76 yang berarti kedua variabel tersebut memiliki hubungan yang kuat. Sedangkan variabel Bedrooms dengan Price menghasilkan nilai sebesar 0.031, Bathrooms dengan Price sebesar 0.036, dan Age dengan Price sebesar 0.16, yang berarti bahwa ketiga variabel ini tidak memiliki hubungan yang kuat dengan variabel Price.

```
X = HP[['SquareFeet', 'Bedrooms', 'Bathrooms','Age']]
y = HP['Price']
```

```
# Membagi data train (80%) dan data test (20%)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.8)
```

```
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor()
regressor.fit(X_train, y_train)

# Feature Importances menunjukkan seberapa penting nilai setiap kolom dalam mempengaruhi kolom harga
print('Feature Importances: ', regressor.feature_importances_)
```

```
Feature Importances:  [0.76191839 0.03164608 0.03699445 0.16944108]
```

2. Evaluasi

Evaluasi model *Random Forest Regression* menggunakan data test x dan test y. Data test x merupakan data independen yang berdiri sendiri tidak dipengaruhi oleh yang lain dengan komposisi 20% dari data itu sendiri. Sedangkan, data test y adalah data dependen yang dipengaruhi oleh data lain dan merupakan variabel data yang ingin diprediksi dengan komposisi 20% juga dari data itu sendiri. Dalam proses ini, test x merupakan data dari

variabel SquareFeet, Bedrooms, Bathrooms, dan Age, sedangkan test y merupakan data dari variabel Price. Evaluasi ini menghasilkan nilai Mean Absolute Error (MAE) sebesar 57%, Residual Sum of Squares (MSE) sebesar 71% dan nilai R-Squared sebesar 49%.

```
#EVALUATION
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
mse = mean_squared_error (y_test, y_pred)
rmse = np.sqrt(mse)
mae = mean_absolute_error (y_test, y_pred)
r2 = r2_score (y_test, y_pred)
print("Residual sum of squares: %.2f" % rmse)
print("R2-score: %.2f" % r2)
print('Mean absoulute error: %.2f' % mae)
```

```
Residual sum of squares: 0.71
R2-score: 0.49
Mean absoulute error: 0.57
```

Selain itu, peneliti juga melakukan uji akurasi model yang menghasilkan akurasi model *Random Forest Regression* sebesar 49,47%.

```
# Skor akurasi model sebesar 49.47% berdasarkan Random Forest Regression Model yang sudah dilakukan
regressor.score(X_test, y_test)
```

```
0.4947721999457073
```

3. Uji Coba Prediksi Data lain

Selanjutnya, peneliti juga mencoba untuk memprediksi sebuah data baru dengan nilai dari SquareFeet sebesar 1800, Bedrooms sebanyak 4 kamar, Bathrooms sebanyak 2 buah, dan Age 10 tahun, yang menghasilkan prediksi harga rumah sebesar 1.29 USD.

```
# Prediksi harga rumah impian Beverly
regressor.predict([[1800,4,2,10]])
```

```
C:\Users\cathe\anaconda3\Lib\site-packa
mForestRegressor was fitted with featur
  warnings.warn(
```

```
array([1.29653768])
```

KESIMPULAN

Berdasarkan uji perbandingan yang telah dilakukan antar tiga model yaitu *Simple Linear Regression, Multiple Linear Regression* dan *Random Forest Regression* pada dataset prediksi harga tumah ini, dapat dipastikan bahwa model *Multiple Linear Regression* memiliki hasil keakuratan yang paling tinggi daripada kedua model tersebut. Hal ini dibuktikan dengan nilai *accuracy* model *Multiple Linear Regression* sebesar 57.83% yang berarti model ini memiliki nilai akurasi yang lebih tinggi daripada nilai akurasi kedua model lain.

REFERENSI

[1] D. D. Wijaya and N. Anastasia, "Pertimbangan generasi milenial pada kepemilikan rumah dan kendala finansial," *Jurnal Manajemen Aset Dan Penilai*, vol. 1, no. 2, Dec. 2021, doi: 10.56960/jmap.v1i2.23.

[2] E. Fitri, "Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," *Journal of Applied Computer Science and Technology*, vol. 4, no. 1, pp. 58–64, Jul. 2023, doi: 10.52158/jacost.v4i1.491.

[3] A. Saiful, "Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression," *JATISI: Jurnal Teknik Informatika Dan Sistem Informasi*, vol. 8, no. 1, pp. 41–50, Mar. 2021, doi: 10.35957/jatisi.v8i1.701.

[4] I. M. Muhamad, "Algoritma Machine Learning untuk penentuan Model Prediksi Produksi Telur Ayam Petelur di Sumatera," Jun. 30, 2022. https://djournals.com/jieee/article/view/382

[5] A. Indrasetianingsih, F. Fitriani, and P. J. Kusuma, "Klasifikasi Indeks Pembangunan gender di Indonesia tahun 2020 menggunakan supervised Machine learning algorithms," *Inferensi: Jurnal Statistika*, vol. 4, no. 2, p. 129, Sep. 2021, doi: 10.12962/j27213862.v4i2.10940.

[6] K. Kristiawan, D. D. Somali, T. Jaya, and A. Widjaja, "Deteksi Buah Menggunakan Supervised Learning dan Ekstraksi Fitur untuk Pemeriksa Harga," *JuTISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, vol. 6, no. 3, Dec. 2020, doi: 10.28932/jutisi.v6i3.3029.

[7] L. B. C. Tanujaya, B. Susanto, and A. Saragih, "The comparison of logistic regression methods and random Forest for Spotify Audio Mode featurre Classification," *Indonesian Journal of Data and Science*, vol. 1, no. 3, Dec. 2020, doi: 10.33096/ijodas.v1i3.16.

[8] E. Triyanto, H. Sismoro, and A. D. Laksito, "IMPLEMENTASI ALGORITMA REGRESI LINEAR BERGANDA UNTUK MEMPREDIKSI PRODUKSI PADI DI KABUPATEN BANTUL," *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, vol. 4, no. 2, pp. 66–75, Jul. 2019, doi: 10.36341/rabit.v4i2.666.

[9] L. B. C. Tanujaya, B. Susanto, and A. Saragih, "The comparison of logistic regression methods and random Forest for Spotify Audio Mode featurre Classification," *Indonesian Journal of Data and Science*, vol. 1, no. 3, Dec. 2020, doi: 10.33096/ijodas.v1i3.16.

[10] N. M. Lalapa, "Implementasi metode regresi linear sederhana untuk prediksi harga cabai rawit," *ejurnal.unisan.ac.id*, Dec. 2023, doi: 10.37195/balok.v2i2.121.

[11] S. K. C. Pulungan, S. K. C. Pulungan, S. M. Pasaribu, S. M. Pasaribu, S. M. Pasaribu, and H. Cipta, "Estimasi penerima alat bantu penyandang disabilitas di Dinas Sosial Kota Medan menggunakan metode regresi linier berganda," *Indonesia Berdaya*, vol. 4, no. 2, pp. 525–534, Jan. 2023, doi: 10.47679/ib.2023450.

[12] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *Jurnal Informasi Dan Teknologi*, vol. 5, no. 3, pp. 58–64, Sep. 2023, doi: 10.60083/jidt.v5i3.393.

[13] J. M. A. S. Dachi, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *prin.or.id*, Jul. 2023, doi: 10.55606/jurrimipa.v2i2.1470.

[14] F. S. D. Alfia and A. Agussalim, "Literature Review Visualisasi Data dan Sistem Informasi Geografis," *COMSERVA*, vol. 2, no. 8, pp. 1494–1500, Dec. 2022, doi: 10.59141/comserva.v2i8.493.

[15] H. D. E. Sinaga and N. Irawati, "PERBANDINGAN DOUBLE MOVING AVERAGE DENGAN DOUBLE EXPONENTIAL SMOOTHING PADA PERAMALAN BAHAN MEDIS HABIS PAKAI," *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, vol. 4, no. 2, pp. 197–204, Jun. 2018, doi: 10.33330/jurteksi.v4i2.60.

[16] N. Afrianto, D. H. Fudholi, and S. S. Rani, "Prediksi Harga Saham Menggunakan BiLSTM dengan Faktor Sentimen Publik," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 6, no. 1, pp. 41–46, Feb. 2022, doi: 10.29207/resti.v6i1.3676.

[17] M. R. A. Prasetya, A. M. Priyatno, and Nurhaeni, "Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining," *Jurnal Informasi Dan Teknologi*, pp. 52–62, Jun. 2023, doi: 10.37034/jidt.v5i2.324.

[18] A. R. Shaumi, "Penerapan Data Mining menggunakan Metode Teknik Classification untuk Melihat Potensi Kepatuhan Wajib Pajak Bumi dan Bangunan," *ioinformatic.org*, 2022, doi: 10.53842/juki.v4i2.131.

[19] M. D. Purbolaksono, M. I. Tantowi, A. Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol. 5, no. 2, pp. 393–399, Apr. 2021, doi: 10.29207/resti.v5i2.3008.

[20] Winata, Welly, Lily Puspa Dewi, and Alvin Nathaniel Tjondrowiguno. "Prediksi Skor Pertandingan Sepak Bola menggunakan Neuroevolution of Augmenting Topologies dan Backpropagation." Jurnal Infra 8, no. 1 (2020): 46-52.