**Progresif:** Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

# Implementasi Algoritma *Support Vector Regression* dan *Linear Regression* Untuk Prediksi Harga Rumah

**Anggun Aulia Bastian[1*], Hanny Hikmayanti Handayani[2], Deden Wahiddin[3], Tatang Rohana[4]**
Teknik Informatika, Universitas Buana Perjuangan, Karawang, Indonesia
*e-mail *Corresponding Author:* If20.anggunbastian@mhs.ubpkarawang.ac.id

***Abstract***
*A house is one of the necessities of human life, but house prices tend to fluctuate every year. This is one of the causes of prospective buyers having difficulty in determining the budget and making decisions to purchase a house. So, it is necessary to conduct research to produce accurate house price predictions. The purpose of this research is to determine the best algorithm between the Support Vector Regression and Linear Regression algorithms in predicting house prices. Datasets are obtained from the results of scrapping on the house buying and selling website. This study uses a dataset from Telukjambe Timur Subdistrict with a total of 547 data with the parameters used are building area and land area and using a data division of 75:25. The results of the RMSE and MAPE evaluation show that the Support Vector Regression Algorithm is superior to Linear Regression with an RMSE value of 234,257 and a MAPE value of 21%.*
***Keywords:*** *House; Price; Prediction; Algorithm; Evaluation*

**Abstrak**
Rumah merupakan salah satu kebutuhan hidup manusia, namun harga rumah pada setiap tahunnya cenderung mengalami fluktuasi. Hal ini menjadi salahsatu penyebab calon pembeli kesulitan dalam menentukan *budget* dan mengambil keputusan untuk melakukan pembelian rumah. Sehingga, perlu dilakukan penelitian untuk menghasilkan prediksi harga rumah yang akurat. Adapun tujuan pada penelitian ini yaitu untuk menentukan Algoritma terbaik antara Algoritma *Support Vector Regression* dan *Linear Regression* dalam memprediksi harga rumah. Dataset diperoleh dari hasil *scrapping* pada *website* jual beli rumah. Penelitian ini menggunakan dataset dari Kecamatan Telukjambe Timur dengan jumlah sebanyak 547 data dengan parameter yang digunakan adalah luas bangunan dan luas tanah serta menggunakan pembagian data 75:25. Adapun hasil dari evaluasi RMSE dan MAPE menunjukkan bahwa Algoritma *Support Vector Regression* lebih unggul dari *Linear Regression* dengan nilai RMSE 234.257 dan nilai MAPE sebesar 21%.
**Kata kunci:** *Rumah; Harga; Prediksi; Algoritma; Evaluasi*

## 1. Pendahuluan

Rumah menjadi salahsatu kebutuhan hidup manusia sebagai tempat tinggal, tempat berlindung dan tempat beristirahat [1]. Seperti halnya investasi emas, kepemilikan rumah juga dianggap sebagai bentuk investasi yang potensial untuk jangka panjang. Namun, seiring dengan berjalannya waktu, harga rumah pada setiap tahunnya cenderung mengalami fluktuasi. Perubahan harga rumah dapat diukur dari beberapa aspek atau faktor pendukung yang dimiliki oleh rumah tersebut [2], [3]. Lokasi yang strategis, kondisi fisik rumah, termasuk usia rumah dan kualitas kontruksi menjadi faktor yang menyebabkan harga rumah mengalami fluktuasi. Seperti saat ini, permintaan rumah di Kabupaten Karawang yang semakin meningkat serta sejalan dengan pertumbuhan penduduk yang terus berlanjut. Urbanisasi dan migrasi penduduk dari luar kota merupakan salah satu penyebabnya. Menurut data dari Badan Pusat Statistik (BPS) Kabupaten Karawang, populasi wilayah karawang mencapai 2.529.882 jiwa pada tahun 2023 dengan tingkat pertumbuhan sebesar 0,42% [4]. Pertumbuhan penduduk juga berdampak pada kebutuhan masyarakat terhadap rumah yang layak sesuai dengan harga yang dapat dijangkau tergantung kebutuhan masing-masing [5].

Ketika akan melakukan pembelian rumah, calon pembeli umumnya mempunyai kriteria untuk membeli rumah baik dari spesifikasi bangunan serta fasilitas yang sejalan dengan keterbatasan *budget* yang dimiliki [1], [6]. Fluktuasi pada harga rumah menyebabkan calon pembeli sulit dalam menentukan budget dalam melakukan pembelian rumah. Sehingga, perlu dilakukan penelitian yang menghasilkan prediksi harga rumah secara akurat, hal ini diharapkan dapat memberikan acuan bagi calon pembeli dalam menentukan pilihan rumah yang sesuai dengan kebutuhan serta menyesuaikan budget yang telah ditentukan sebelumnya.

Prediksi adalah suatu metode dalam memperkirakan suatu nilai pada masa mendatang dengan mempertimbangkan data atau informasi pada masa lampau maupun pada masa sekarang [7]. Prediksi tidak harus menghasilkan nilai yang sesuai dengan apa yang sebenarnya terjadi, tetapi diupayakan untuk menghasilkan nilai yang seakurat mungkin dengan apa yang akan terjadi [8]. Dalam hal ini, nilai prediksi yang dihasilkan dapat dijadikan acuan untuk membuat suatu keputusan [9].

Algoritma *Support Vector Regression* merupakan model regresi dari pengembangan Algoritma *Support Vector Machine* (SVM) [10]. Pada kasus regresi, Algoritma ini menghasilkan nilai dalam bentuk bilangan real (riil) atau sekuensial (kontinu). Pendekatan SVR digunakan untuk memprediksi karena memiliki kemampuan untuk mengatasi *overfitting* pada akurasi data training saat melakukan prediksi [1], [10]. Algoritma ini bertujuan untuk menemukan garis pemisah (*hyperline*) dengan mengukur *margin* atau jarak terdekatnya dari pola data. Adapun Algoritma *Linear Regression* merupakan sebuah teknik untuk membuat prediksi berdasarkan hubungan antar variabel dependen (x) dan variabel independent (y) [1] . Terdapat dua jenis Algoritma Linear Regression diantaranya *Simple Linear Regression* dan Multiple Linear Regression. *Simple Linear Regression* adalah hubungan antara satu variabel dependen dengan satu variabel independen, sedangkan *Multiple Linear Regression* adalah hubungan antar satu variabel dependen dengan dua atau lebih variabel independennya [10].

Berdasarkan latarbelakang yang telah dipaparkan, pada penelitian ini akan menggunakan Algoritma *Support Vector Regression* dan *Linear Regression* untuk prediksi harga rumah. Penelitian ini bertujuan untuk menentukan algoritma terbaik antara Algoritma *Support Vector Regression* dan *Linear Regression* untuk prediksi harga rumah di Kabupaten Karawang.

## 2. Tinjauan Pustaka

Berdasarkan penelitian [11] untuk prediksi harga rumah menggunakan Algoritma K-Nearest Neighbor dan Naïve Bayes. Data diperoleh dari Kaggle dengan jumlah data sebanyak 4.601 dataset. Hasil dari penelitian ini menunjukkan bahwa Algoritma KNN lebih baik dibandingkan naïve bayes dengan nilai akurasi 0,5714. Namun, Algoritma tersebut kurang bisa dalam melakukan prediksi yang dimana nilai MAPE mencapai 43,52 yang termasuk kategori “cukup”.

Selanjutnya, pada penelitian [2] dengan topik yang sama [11] menggunakan Algoritma *Random Forrest Regression* dan *Multiple Linear Regression*. Pada penelitian ini, dataset diperoleh dari *website Kaggle.com* dengan jumlah sebanyak 1010 dataset dengan 7 variabel dengan pembagian data 80:20. Hasil dari penelitian ini menunjukkan bahwa Algoritma *Random Forrest Regression* menghasilkan nilai akurasi yang tinggi yaitu 81,86%. Kemudian, pada penelitian yang telah dilakukan [12] menggunakan Algoritma *Random Forest Regression* untuk prediksi harga rumah. Dataset diperoleh dari *website Kaggle.com* dengan jumlah sebanyak 1001 dataset rumah di wilayah Jakarta Selatan. Hasil dari penelitian ini memperoleh akurasi sebesar 75,10%.

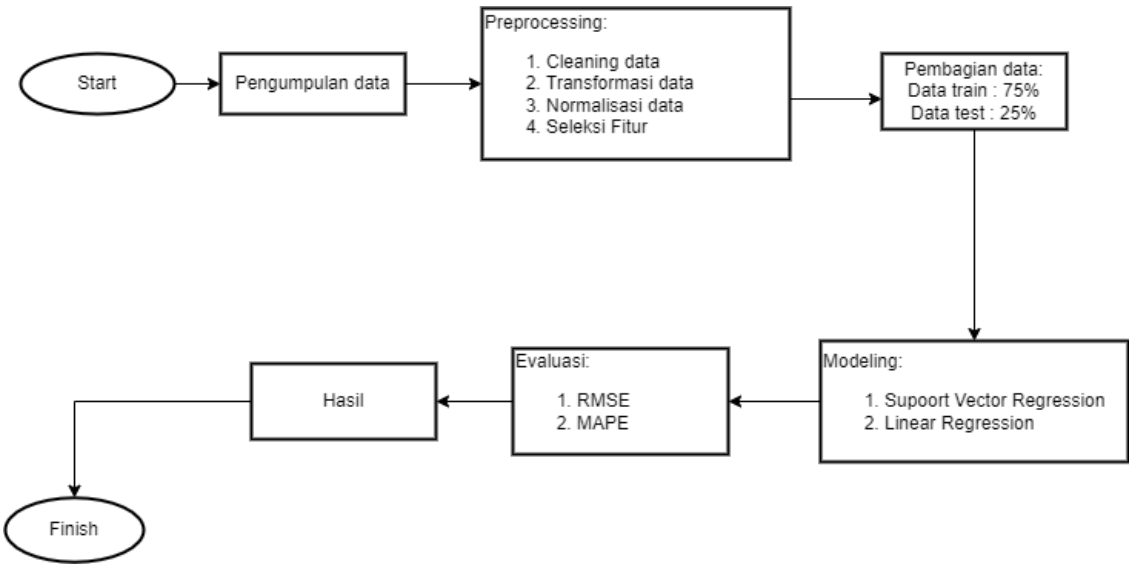
Selanjutnya pada penelitian yang dilakukan [13] dengan topik yang sama [2], [11] menggunakan Algoritma Regresi Linier, *Random Forest Regression* dan *Gradient Boosted Trees Regression*. Dataset yang digunakan diperoleh dari dataset yang sama [2], [12] dengan masing-masing memiliki 7 variabel. Hasil prediksi menunjukkan bahwa Algoritma Random Forres Regression memiliki nilai error yang lebih kecil yaitu 0.440Selanjutnya, pada penelitian [6] untuk prediksi harga rumah daerah Jabodetabek dengan menggunakan Algoritma *Multiple Linear Regression.* Dataset yang digunakan sebanyak 3553 data dengan pembagian data 80:20. Hasil dari penelitian ini menghasilkan nilai akurasi 85%.

Setelah melakukan *literatur review* pada penelitian sebelumnya, penelitian ini terdapat beberapa kesamaan dengan penelitian sebelumnya salahsatunya yaitu mengimplementasikan Algoritma *Support Vector Regression* dan *Linear Regression*. Adapun perbedaan dari penelitian sebelumnya yaitu pada penggunaan dataset, dalam penelitian ini dataset yang akan digunakan

diperoleh dari hasil *scraping* dari *website* jual beli rumah daerah Karawang. Selain itu, penelitian ini akan menerapkan tahap teknik *preprocessing* tambahan yaitu normalisasi data dan seleksi fitur. Hal ini diharapkan dapat prediksi yang lebih akurat.

## 3. Metodologi

Pada penelitian ini terdiri dari beberapa prosedur diantaranya pengumpulan data, *pre-processing*, pembagian dataset, *modeling* dan evaluasi. Adapun prosedur penelitian dapat dilihat pada gambar 1.




Gambar 1. Prosedur Penelitian

### 3.1. Pengumpulan data

Dataset yang digunakan yaitu data rumah di Kabupaten Karawang yang diperoleh dari hasil *scraping* pada *website rumah123.com* dan *lamudi.com* yang merupakan sebuah *website* jual beli rumah di Indonesia. Dengan melakukan teknik *scraping*, memungkinkan untuk mengumpulkan data dengan jumlah yang besar dan mendapatkan sumber data terkait variabel yang dibutuhkan. Adapun data yang diperoleh dari hasil scraping sebanyak 1653 data dengan 9 variabel.

### 3.2. Preprocessing

Tahap *preprocessing* merupakan tahapan penting dalam pengolahan suatu data. *Preprocessing* bertujuan untuk meningkatkan kualitas data agar dapat digunakan secara efektif. *Peprocessing* dimulai dengan melakukan *cleaning data* yang terdiri dari drop variabel yang tidak relevan. Kemudian dilanjutkan dengan melakukan pemeriksaan *missing value*, duplikat data, *outlier* dan *noise*. Setelah melakukan *cleaning* data, tahapan selanjutnya yaitu melakukan transformasi data. Pada tahap ini digunakan untuk mengubah data objek menjadi *integer* dengan menggunakan teknik *label encoder*. Selain itu, pada tahap ini juga melakukan perubahan data pada variabel *price* dengan cara setiap data dibagi menjadi satu juta, hal ini bertujuan untuk memudahkan dalam membaca data. Kemudian pada tahap selanjutnya normalisasi data. Tahap normalisasi pada penelitian ini menggunakan teknik *MinMaxScaller*.

### 3.3. Pembagian data

Tahap selanjutnya yaitu pembagian dataset. Penelitian ini akan menggunakan pembagian data 75% sebagai data latih dan 25% sebagai data uji.

### 3.4. Modeling

Pada penelitian ini prediksi dilakukan menggunakan Algoritma *Support Vector Regression* dan *Linear Regression*. Untuk mengimplementasikan Algoritma SVR menggunakan

*high dimensional feature space* atau disebut dengan kernel. Berikut merupakan rumus untuk kasus non linear yang ditunjukkan pada persamaan (1)

$f(x) = \sum_i^n (a_i + a_i^*) + K(x_i, x) + b$..... (1)

*Radial Basis Function* merupakan salahsatu kernel yang sering digunakan dalam penelitian [14]. Berikut merupakan rumus pada kernel RBF yang ditunjukkan pada persamaan (2) [15].

$K(x, y) = -|x.y|^2/2\sigma^2$..... (2)

Selain Algoritma *Support Vector Regression*, penelitian ini juga mengimplementasikan Algoritma *Linear Regression*. Pada penelitian ini akan menggunakan jenis Algoritma *Multiple Linear Regression*. Berikut merupakan perhitungan dari *Multiple Linear Regression* yang ditunjukkan pada persamaan (3)

$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n$..... (3)

### 3.5. Evaluasi

Setelah melakukan tahapan *modeling*, tahap selanjutnya yaitu melakukan evaluasi untuk menghitung tingkat akurasi antara nilai prediksi dan nilai aktual berdasarkan Algoritma yang diterapkan. Adapun pada penelitian akan menggunakan evalusasi *Root Mean Squared Error* (RMSE) dan *Mean Absolute Percentage Error* (MAPE). Berikut merupakan perhitungan RMSE yang ditunjukkan pada persamaan (4)

$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum(y_i - \overline{y})^2}$..... (4)

Adapun perhitungan MAPE yang ditunjukkan pada persamaan (5)

$MAPE = \frac{\sum_i^n=1\left|\frac{xi-fi}{xi}\right|}{n} \times 100$ ..... (5)

RMSE merupakan perhitungan nilai kuadrat *error* atau selisih antara nilai aktual dan nilai prediksi, yang dimana jika nilai RMSE yang dihasilkan tinggi maka keakuratan prediksi model cenderung jauh dari nilai yang sebenarnya [16]. Adapun evaluasi MAPE biasanya digunakan untuk pengukuran presentase kesalahan antara nilai aktual dengan hasil prediksi. Adapun skala akurasi pada evaluasi MAPE ditunjukkan pada tabel 1.

Tabel 1. Akurasi Prediksi MAPE [16]

| Presentase MAPE | Kategori |
|---|---|
| <10% | Sangat baik |
| <20 % | Baik |
| <50 % | Cukup |
| >50% | Buruk |

## 4. Hasil dan Pembahasan

### 4.1. Pengumpulan data

Pengumpulan data dilakukan dengan teknik *scraping* dari *website* jual beli rumah yaitu *rumah123.com* dan *lamudi.com*. Dalam melakukan *scraping* data *tools* yang digunakan adalah *Web Scraper extentions* dari *google chrome*. Data yang diperoleh dari hasil scraping sebanyak 1653 yang terdiri dari 9 variabel diantaranya *years, title, addres, building_area, surface_area, bedroom, bathroom, area_parking* dan *price*. Berikut adalah dataset hasil scraping yang ditunjukkan pada gambar 2.

| | years | title | address | surface_area | building_area | bedroom | bathroom | area_parking | price |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2023 | Rumah Asteria Galuh Mas Karawang.Harga Dibawah... | Telukjambe Timur, Karawang | 60 | 30 | 2 | 1 | 1 | 695000000 |
| 1 | 2024 | Rumah Siap Huni diGrand Taruma Karawang Barat | Karawang Barat, Karawang | 300 | 235 | 4 | 3 | 2 | 4970000000 |
| 2 | 2024 | Perumahan Strategis Tengah Kota, Dijual Cepat ... | Karawang Barat, Karawang | 144 | 83 | 3 | 2 | 4 | 4000000000 |
| 3 | 2024 | Lokasi Strategis, Dijual Cepat Rumah 2 Lantai ... | Telukjambe Timur, Karawang | 126 | 90 | 3 | 2 | 1 | 1600000000 |
| 4 | 2024 | Rumah Asri 2 Lantai di Cluster Grand Taruma Ka... | Telukjambe Timur, Karawang | 126 | 69 | 2 | 2 | 1 | 1100000000 |

Gambar 2. Dataset rumah

Berdasarkan gambar 2, pada variabel address terdiri dari 29 kecamatan. Namun, pada penelitian ini akan menggunakan satu kecamatan saja, hal ini dikarenakan pada dataset tersebut memiliki rentang harga yang terlalu jauh yang dapat menyebabkan hasil prediksi yang kurang maksimal. Sebelum melakukan preprocessing, dataset dikelompokkan berdasarkan kecamatan, kemudian dilanjutkan dengan menghitung jumlah dataset pada masing-masing kecamatan. Dari perhitungan tersebut, diperoleh bahwa kecamatan Telukjambe Timur memiliki jumlah data yang paling banyak. Sehingga, penelitian ini akan menggunakan dataset dari kecamatan Telukjambe Timur. Adapun jumlah dataset pada kecamatan Telukjambe Timur sebanyak 547 data. Berikut merupakan data hasil dari pengelompokkan berdasarkan alamat yang ditunjukkan pada gambar 3.

| | years | title | address | surface_area | building_area | bedroom | bathroom | area_parking | price |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2023 | Rumah Asteria Galuh Mas Karawang.Harga Dibawah... | Telukjambe Timur, Karawang | 60 | 30 | 2 | 1 | 1 | 695000000 |
| 3 | 2024 | Lokasi Strategis, Dijual Cepat Rumah 2 Lantai ... | Telukjambe Timur, Karawang | 126 | 90 | 3 | 2 | 1 | 1600000000 |
| 4 | 2024 | Rumah Asri 2 Lantai di Cluster Grand Taruma Ka... | Telukjambe Timur, Karawang | 126 | 69 | 2 | 2 | 1 | 1100000000 |
| 5 | 2023 | Rumah 2 lantai, minimalis, di Galuh Mas, Karaw... | Telukjambe Timur, Karawang | 72 | 68 | 3 | 2 | 1 | 1240000000 |
| 8 | 2024 | TOWN HOUSE THALASA DI GALUH MAS KARAWANG KARAW... | Telukjambe Timur, Karawang | 45 | 50 | 3 | 3 | 1 | 791000000 |

Gambar 3. Dataset rumah Kecamatan Telukjambe Timur

Setelah proses tersebut, langkah selanjutnya yaitu melihat informasi dari dataset dengan menggunakan *fungsi info()*. Fungsi ini digunakan untuk melihat informasi dari dataset seperti jumlah kolom dan *type* data dari masing-masing variabel. Berdasarkan dataset tersebut, terdapat *type* data *integer* sebanyak delapan kolom dan *type* data *object* sebanyak satu kolom. Berikut merupakan informasi dataset yang ditunjukkan pada gambar 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 547 entries, 0 to 546
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   years          547 non-null    int64
 1   title          547 non-null    object
 2   address        547 non-null    object
 3   surface_area   547 non-null    int64
 4   building_area  547 non-null    int64
 5   bedroom        547 non-null    int64
 6   bathroom       547 non-null    int64
 7   area_parking   547 non-null    int64
 8   price          547 non-null    int64
dtypes: int64(7), object(2)
```

Gambar 4. Informasi dataset

### 4.2. *Preprocessing*

Pada tahap *preprocessing* terdiri dari beberapa langkah yaitu *cleaning* data, transformasi data, normalisasi data, dan seleksi fitur.

1) *Cleaning* data

Pada tahap *cleaning* data terdiri dari beberapa langkah yaitu drop variabel yang tidak relevan, pemeriksaan *missing value*, pemeriksaan duplikat data dan pemeriksaan *noise* dan *outlier*. Tahap pertama yang dilakukan yaitu drop variabel yang tidak relevan. Variabel yang akan di drop adalah variable *title dan address*. Hasil drop variabel ditunjukkan pada gambar 5.

| | years | surface_area | building_area | bedroom | bathroom | area_parking | price |
|---|---|---|---|---|---|---|---|
| **0** | 2023 | 60 | 30 | 2 | 1 | 1 | 695000000 |
| **1** | 2024 | 126 | 90 | 3 | 2 | 1 | 1600000000 |
| **2** | 2024 | 126 | 69 | 2 | 2 | 1 | 1100000000 |
| **3** | 2023 | 72 | 68 | 3 | 2 | 1 | 1240000000 |
| **4** | 2024 | 45 | 50 | 3 | 3 | 1 | 791000000 |

Gambar 5. Hasil drop variabel

Setelah proses drop variabel, tahap selanjutnya yaitu pengecekan *missing value* dengan menggunakan fungsi *isnull()*. Pengecekan *missing value* dilakukan untuk mencari nilai kosong diantara data yang memiliki nilai. Hasil dari penerapan *missing value* menunjukkan bahwa dataset yang digunakan tidak memiliki *missing value*. Hasil pengecekan *missing value* ditunjukkan pada tabel 2.

Tabel 2. Hasil pemeriksaan *missing value*

| | |
|---|---|
| years | 0 |
| Surface_area | 0 |
| Building_area | 0 |
| bedroom | 0 |
| bathroom | 0 |
| area_parking | 0 |
| price | 0 |

Setelah melakukan proses pengecekan *missing value*, tahap selanjutnya yaitu pengecekan duplikat data dengan menggunakan fungsi *duplicated()*. Tahap ini bertujuan untuk mengetahui jumlah data dengan nilai yang sama. Adapun *source code* dan hasil pengecekan duplikat data ditunjukkan pada gambar 6.

```
# jumlah duplikasi data
jumlah_duplikasi = predict.duplicated().sum()
print(f'Jumlah Duplikasi Data: {jumlah_duplikasi}')
```

```
Jumlah Duplikasi Data: 21
```

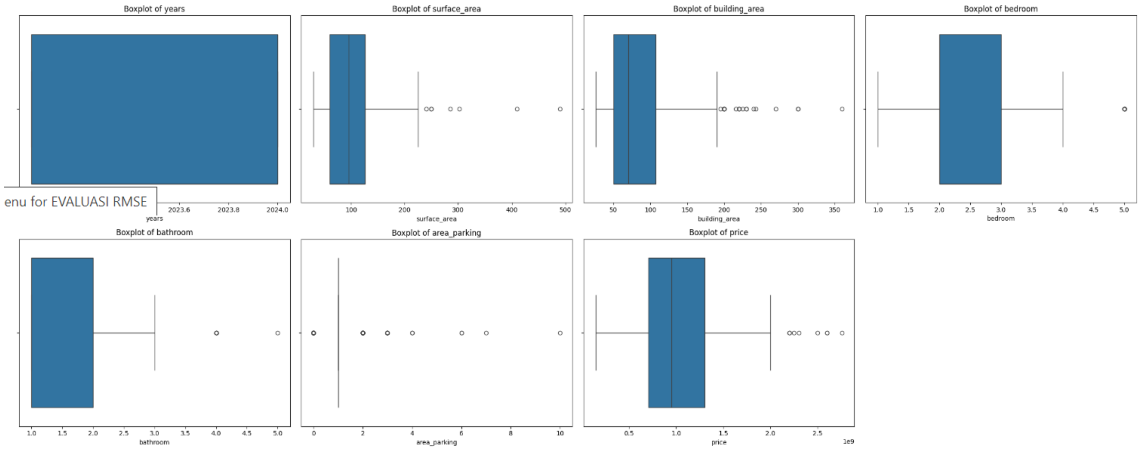
Gambar 6. *Source code* dan hasil duplikat data

Berdasarkan gambar 6. menunjukkan bahwa dataset yang digunakan memiliki duplikat data dengan jumlah sebanyak 21 data. Adapun langkah untuk menghapus duplikat data yaitu dengan menggunakan fungsi *drop.duplicated*(). Pada tabel 3 menunjukkan jumlah data sebelum dan sesudah dilakukan penghapusan duplikat data.

Tabel 3. Jumlah dataset sebelum dan sesudah pembersihan duplikat data

| | |
|---|---|
| Jumlah Data Awal: | 547 |
| Jumlah Data Bersih: | 526 |

Tahap selanjutnya yaitu proses pemeriksaan data *noise* dan *outlier*. Adapun visualisasi data *noise* dan *outlier* ditunjukkan pada gambar 7.




Gambar 7. Data noise dan outlier

Berdasarkan gambar 7. Data yang digunakan terdapat data *outlier*. Setelah mengetahui dataset yang digunakan memiliki *outlier,* langkah selanjutnya yaitu menghapus *outlier* pada baris dan kolom menggunakan teknik *interquartile range* (IQR). Adapun proses menghapus *outlier* dilakukan dengan cara menghitung nilai Q1, Q2 dan IQR. Selanjutnya menentukan nilai batas atas dan batas bawah. Jika nilai berada pada rentang batas atas dan bawah, maka data tersebut dianggap sebagai *outlier*. Hasil dari penghapusan data outlier, dataset yang digunakan menjadi 357. Adapun teknik yang digunakan untuk mendeteksi data *noise* yaitu dengan menggunakan teknik *z-score*. Proses deteksi *noise* dilakukan dengan cara mengidentifikasi nilai rata-rata dengan menggunakan standar deviasi. Setelah itu dilakukan perhitungan dengan menggunakan nilai ambang batas. Jika hasil perhitungan berada diluar rentang ambang batas, maka data tersebut dianggap sebagai *noise*. Berdasarkan penerapan teknik tersebut, terdapat 3 data yang dianggap sebagai data *noise*. Kemudian dilakukan pembersihan data *noise* pada dataset, sehingga dataset berubah menjadi 354 data.

2) Transformasi Data

Proses selanjutnya yaitu transformasi data, pada penelitian ini transformasi data digunakan untuk mengubah nilai yang terdapat pada variabel *price* dengan membagi data *price* dibagi dengan satu juta. Hasilnya akan mengurangi nilai 0 sebanyak 6 karakter dari belakang. Hal ini dilakukan agar data mudah untuk diolah dan hasilnya juga mudah dibaca. Berikut merupakan data sebelum dan sesudah transformasi pada pada variabel *price* yang ditunjukkan pada tabel 4.

Tabel 4. Data transfromasi variabel *price*

| **Sebelum** | **Sesudah** |
|---|---|
| 695000000 | 695 |
| 1600000000 | 1600 |

3) Normalisasi data

Tahap selanjutnya yaitu normalisasi data menggunakan teknik *MinMaxScaler*. Teknik ini dilakukan dengan cara menghitung nilai *minimum* dan *maximum* di setiap kolom pada dataset. Adapun hasil normalisasi akan menghasilkan nilai dengan batas range 0 sampai 1. Pada tabel 5 menunjukkan hasil dari normalisasi menggunakan teknik MinMaxScaller.

Tabel 5. Data hasil normalisasi

| **years** | **Surface_ area** | **...** | **bathroom** | **area_ parking** | **price** |
|---|---|---|---|---|---|
| 0,0 | 0,176471 | ... | 0,0 | 0,0 | 695 |
| 1,0 | 0,565706 | . .. | 0,5 | 0,0 | 1600 |
| 1,0 | 0,565706 | ... | 0,5 | 0,0 | 1100 |

| | | | | | |
|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... |
| 0,0 | 0,176471 | ... | 0,0 | 0,0 | 714 |
| 0,0 | 0,211765 | ... | 0,0 | 0,0 | 780 |
| 1,0 | 0,176471 | ... | 0,0 | 0,0 | 657 |

4) Seleksi Fitur

Tahap selanjutnya yaitu seleksi fitur. Teknik yang digunakan pada tahap ini yaitu *correlation based fitur*. Proses yang dilakukan adalah menghitung nilai korelasi antar atribut dengan target. Jika nilai korelasi mendekati angka 1 mengindikasikan adanya hubungan positif yang kuat antar atribut, sedangkan jika nilai korelasi mendekati 0, mengindikasikan bahwa tidak adanya hubungan antar atribut. Berikut merupakan visualisasi dari seleksi fitur yang ditunjukkan pada gambar 8.

| | years | surface_area | building_area | bedroom | bathroom | area_parking | price |
|---|---|---|---|---|---|---|---|
| years | 1.00 | -0.08 | -0.02 | 0.11 | 0.11 | nan | -0.03 |
| surface_area | -0.08 | 1.00 | 0.72 | 0.44 | 0.26 | nan | 0.62 |
| building_area | -0.02 | 0.72 | 1.00 | 0.56 | 0.38 | nan | 0.56 |
| bedroom | 0.11 | 0.44 | 0.56 | 1.00 | 0.61 | nan | 0.37 |
| bathroom | 0.11 | 0.26 | 0.38 | 0.61 | 1.00 | nan | 0.33 |
| area_parking | nan | nan | nan | nan | nan | nan | nan |
| price | -0.03 | 0.62 | 0.56 | 0.37 | 0.33 | nan | 1.00 |

Gambar 8. Visualisasi Korelasi antar atribut

Berdasarkan gambar 8. terdapat dua atribut yang memiliki korelasi yang tinggi dengan atribut *price*. Adapun atribut tersebut antara lain *surface_area* (0,62) dan *building_area* (0,56). Sedangkan *years* (-0,034), *bedroom* (0,37) dan *bathroom* (0,33 dan *area_parking* bernilai (nan) yang artinya dataset tersebut memiliki nilai yang sama, sehingga tidak ada kolelasi yang dapat dihitung. Oleh karena itu, atribut yang memiliki nilai korelasi dibawah 0,5 akan di drop. Adapun fungsi yang digunakan untuk menghapus atribut yaitu fungsi *drop().* Berikut merupakan atribut yang digunakan yaitu *surface_area*, *building_area* dan *price*.

### 4.3. Pembagian data

Tahap selanjutnya yaitu melakukan pembagian data latih dan data uji. Sebelum melakukan pembagian data, variabel pada dataset terlebih dahulu dibagi menjadi variabel X dan y. Berikut merupakan *source code* untuk membagi data ditunjukkan pada gambar 9.

```
from sklearn.model_selection import train_test_split

# Split Data (Training & Testing)
# Buat variabel independen X dan variabel dependen y
X = predict[['surface_area', 'building_area']]
y = predict['price']
```

Gambar 9. *Source Code* pembagian data

Berdasarkan gambar 9. menjelaskan bahwa variabel X terdiri dari variabel *surface_area*, dan *building_area*. Sedangkan, variabel y terdiri dari variabel *price*. Variabel y juga disebut sebagai variabel target untuk melakukan prediksi. Setelah dilakukan proses tersebut, tahap selanjutnya yaitu membagi dataset dengan menggunakan pembagian data 75:25. Adapun hasil pembagian dataset ditunjukkan pada tabel 6.

Tabel 6. Hasil pembagian data latih dan uji

| | |
|---|---|
| **Data latih:** | 265 |
| **Data uji:** | 89 |

### 4.4. Modeling

4.4.1. Algoritma Support Vector Regression

Sebelum melakukan implementasi model, terlebih dahulu menentukan parameter optimal agar model dapat memprediksi secara akurat. Berdasarkan penelitian [17] bahwa Algoritma *Support Vector Regression* dengan menentukan parameter optimal menggunakan *GridSearchCV* mampu untuk memprediksi dataset dengan cukup akurat. Sehingga pencarian parameter terbaik pada penelitian ini menggunakan *GridSearchCV* dengan parameter yang dicari C, gamma dan epsilon. Adapun tahap awal untuk menentukan mencari parameter terbaik dengan menggunakan *GridSearchCV* perlu ditentukan nilai pada setiap parameter. Berikut merupakan nilai dari setiap parameter untuk menemukan nilai parameter terbaik yang ditunjukkan pada tabel 7.

Tabel 7. Nilai masing-masing parameter [18]

| Parameter | Nilai |
|---|---|
| C | 0.1, 1, 10, 100 |
| Gamma | 0.01, 0.1, 1, 10 |
| Epsilon | 0.001, 0.01, 0.1, 1, 10 |

Berdasarkan tabel 7. Nilai parameter C, Gamma dan Epsilon diperoleh dari penelitian yang telah dilakukan sebelumnya [17], [18]. Berdasarkan nilai parameter tersebut, pada penelitian ini diperoleh hasil nilai parameter yang paling optimal yaitu C= 100, gamma = 10, epsilon = 10. Setelah mengetahui nilai terbaik dari masing-masing parameter, tahap selanjutnya yaitu implementasi Algoritma untuk memprediksi harga rumah yang berlokasi di Kabupaten Karawang menggunakan *Support Vector Regression* dengan menggunakan kernel RBF. Pada tabel 8. menunjukkan hasil prediksi dengan menampilkan perbandingan antara data aktual dan data prediksi.

Tabel 8. Nilai aktual dan prediksi Algoritma *Support Vector Regression*

| Index | Aktual | Prediksi | Selisih |
|---|---|---|---|
| 0 | 1200 | 1009 | 190 |
| 1 | 895 | 1033 | -138 |
| 2 | 670 | 930 | -260 |
| 3 | 1350 | 1196 | 153 |
| 4 | 1310 | 1468 | -158 |
| … | … | … | … |
| 85 | 650 | 719 | -69 |
| 86 | 750 | 985 | -235 |
| 87 | 810 | 1105 | -295 |
| 88 | 690 | 682 | 7 |
| 89 | 1330 | 1056 | 273 |

Berdasarkan tabel 8. Menunjukkan perbandingan antara nilai aktual dan nilai prediksi, terlihat selisih pada setiap dataset yang memiliki nilai yang beragam, seperti pada index 88 yang memiliki nilai selisih 7, kemudian terdapat selisih yang paling jauh yaitu pada index 87 yang memiliki nilai selisih mencapai (-295).

4.4.2. Linear Regression

Pada penelitian ini, jenis algoritma yang digunakan yaitu *Multiple Linear Regression*. Untuk implementasi algoritma, model di inisialisasi dengan '*lin_reg = LinearRegression()*' dan dilatih menggunakan data prediksi (*X_train*) dan data target (*y_train*) menggunakan '*lin_reg.fit(X_train, y_train)*'. Setelah melakukan pemodelan algoritma, langkah selanjutnya yaitu menentukan persamaan Algoritma *Multiple Linear Regression* yang menghasilkan nilai koefisien (b) dan intercept (a) yang ditunjukkan pada gambar 10.

```
Intercept: 625.9300739793332
Coef: [666.45007155 406.76113706]
```

Gambar 10. Persamaan Linear Regression

Berdasarkan gambar 10. terdapat nilai intercept dan *coefficient*. Nilai *intercept* merupakan nilai rata rata y, sedangkan nilai *coef* adalah bilangan yang melekat pada sebuah variabel, maka persamaannya dapat dituliskan sebagai berikut.

$$y = 625.9300739793332 + (666.45007155)\ x1 + (406.76113706)\ x2$$

Dari persamaan diatas, perhitungan Algoritma *Multiple Linear Regression* menghasilkan prediksi harga rumah di Kabupaten Karawang. Berikut merupakan hasil prediksi harga rumah di Kabupaten Karawang dengan menampilkan perbandingan harga aktual dengan harga prediksi yang ditunjukkan pada tabel 9.

Tabel 9. Hasil nilai aktual dan prediksi Algoritma *Linear Regression*

| Index | Aktual | Prediksi | Selisih |
|---|---|---|---|
| 0 | 1200 | 914 | 285 |
| 1 | 895 | 882 | 12 |
| 2 | 670 | 881 | -211 |
| 3 | 1350 | 1481 | -131 |
| 4 | 1310 | 1423 | -113 |
| … | … | … | … |
| 85 | 650 | 794 | -144 |
| 86 | 750 | 964 | -214 |
| 87 | 810 | 978 | -168 |
| 88 | 690 | 755 | -65 |
| 89 | 1330 | 1010 | 319 |

Berdasarkan tabel 9. menunjukkan perbandingan antara nilai aktual dan nilai prediksi, terlihat selisih pada setiap dataset yang memiliki nilai beragam, seperti pada index 1 yang memiliki nilai selisih 12, kemudian terdapat selisih yang paling jauh yaitu pada index 89 yang memiliki nilai selisih mencapai 319.

### 4.5. Evaluasi

Tahap selanjutnya yaitu evaluasi model dari kedua Algoritma menggunakan RMSE dan MAPE. Berikut merupakan hasil dari evaluasi menggunakan RMSE dan MAPE ditunjukkan pada tabel 10.

Tabel 10. Hasil Evaluasi RMSE dan MAPE

| Algoritma | RMSE | MAPE |
|---|---|---|
| Support Vector Regression | 234.257 | 21% |
| Linear Regression | 249.086 | 22% |

Berdasarkan tabel 10. hasil RMSE digunakan untuk menghitung seberapa besar hasil prediksi dan aktual yang diperoleh dari perhitungan akar yang dikuadratkan yang kemudian dirata-ratakan. Semakin mendekati angka 0 maka hasil prediksi sangat baik. Berdasarkan tabel 11. evaluasi pada kedua model algoritma menggunakan RMSE menghasilkan nilai 234.257 untuk Algoritma *Support Vector Regression* dan 249.086 untuk Algoritma Linear Regression, yang dimana hasil evaluasi algoritma menghasilkan selisih nilai prediksi dan nilai aktual yang cukup besar. Sedangkan pada evaluasi menggunakan MAPE, pada Algoritma *Support Vector Regression* menghasilkan nilai 21% dan untuk Algoritma *Linear Regression* menghasilkan nilai sebesar 22%. Dari hasil evaluasi RMSE dan MAPE dapat disimpulkan bahwa nilai evaluasi Algoritma *Support Vector Regression* lebih unggul dibandingkan Algoritma *Linear Regression*.

### 4.6. Pembahasan

Penelitian ini memperkuat dari penelitian sebelumnya tentang prediksi harga rumah menggunakan Algoritma *Support Vector Regression* dan *Linear Regression*. Berdasarkan penelitian sebelumnya menunjukan bahwa kedua Algoritma tersebut efektif dalam melakukan prediksi harga rumah dengan menghasilkan nilai prediksi yang bervariasi tergantung pada dataset dan tahapan yang digunakan. Pada penelitian [3] menunjukkan Algoritma SVR memiliki akurasi yang tinggi dibandingkan dengan Algoritma *Linear Regression*. Adapun pada penelitian [2] menunjukkan bahwa Algoritma *Linear Regression* menghasilkan nilai yang cukup baik untuk melakukan prediksi harga rumah.

Pada penelitian ini, pengujian dilakukan menggunakan Algoritma *Support Vector Regression* dan *Linear Regression*. Penelitian ini menggunakan dataset hasil *scraping* pada *website* jual beli rumah. Hasil dari penelitian ini menunjukkan bahwa Algoritma *Support Vector Regression* memiliki nilai error yang lebih rendah. Hal ini mendukung temuan dari penelitian sebelumnya [3] yang menunjukkan bahwa Algoritma SVR efektif dalam melakukan prediksi harga rumah. Dengan demikian, penelitian ini tidak hanya mendukung dari penelitian-penelitian sebelumnya, tetapi penelitian ini menerapkan teknik tambahan pada tahap preprocessing sehingga menghasilkan akurasi yang cukup baik.

## 5. Simpulan

Berdasarkan hasil penelitian dengan menggunakan Algoritma *Support Vector Regression* dan *Linear Regression*, dapat disimpulkan bahwa untuk memprediksi harga rumah dengan menggunakan Algoritma *Support Vector Regression* lebih baik dibandingkan Algoritma *Linear Regression*. Dapat dilihat hasil evaluasi dengan menggunakan RMSE dan MAPE menunjukkan Algoritma *Support Vector Regression* memiliki nilai *error* yang lebih kecil dibandingkan dengan Algoritma *Linear Regression*. Adapun nilai RMSE yang diperoleh algoritma *Support Vector Regression* sebesar 252.277, sedangkan pada algoritma *Linear Regression* diperoleh nilai RMSE sebesar 253.729. Sedangkan, pada evaluasi menggunakan MAPE pada Algoritma *Support Vector Regression* menghasilkan nilai 21%, sedangkan pada Algoritma *Linear Regression* menghasilkan nilai 22%. Berdasarkan hasil selisih nilai aktual dan nilai prediksi yang sangat beragam, diharapkan pada penelitian selanjutnya menggunakan data yang lebih banyak serta dengan menambahkan variabel-variabel independent yang relevan dan menggunakan Algoritma pembanding lainnya untuk mengetahui hasil yang lebih bagus serta pengujian lebih akurat.

## Daftar Referensi

[1] L. M. Muhammad, S. A. Damayanti, H. N. Zaki, T. Muhayat, and R. Wirawan, “Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression,” *Jurnal Informatik*, vol. 3, no. 17, pp. 238–245, 2021.

[2] C. Haryanto, N. Rahaningsih, and F. M. Basysyar, “Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah,” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 533–539, 2023.

[3] I. P. Putra and I. K. G. Suhartana, “Perbandingan Akurasi Algoritma Regresi Linier, Regresi Polinomial, dan Support Vector Regression Pada Model Sistem Prediksi Harga Rumah,” *Jurnal Nasional Teknologi Informasi dan Aplikasinya (JNATIA)*, vol. 1, no. 1, pp. 147–154, 2022.

[4] B. P. Statistik, “Penduduk, Laju Pertumbuhan Penduduk per Tahun dan Distribusi Persentase Penduduk Menurut Kecamatan di Kabupaten Karawang,” BADAN PUSAT STATISTIK. Accessed: Jan. 23, 2024. [Online]. Available: https://karawangkab.bps.go.id/statictable/2023/10/10/292/penduduk-laju-pertumbuhan-penduduk-per-tahun-dan-distribusi-persentase-penduduk-menurut-kecamatan-di-kabupaten-karawang-2022.html

[5] E. S. Lestari and I. Astuti, “Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat,” *Jurnal Ilmiah FIFO*, vol. 14, no. 2, p. 131, Nov. 2022, doi: 10.22441/fifo.2022.v14i2.003.

[6] L. Uswatun Hasanah, I. Maula, and A. Tholib, “Analisis Prediksi Harga Rumah di Jabodetabek Menggunakan Multiple Linear Regression,” *Jurnal Informatika Kaputama (JIK)*, vol. 7, no. 2, pp. 216–224, 2023.

[7] A. M. Siregar, S. Faisal, Y. Cahyana, and B. Priyatna, “Perbandingan Algoritme Klasifikasi Untuk Prediksi Cuaca,” *Jurnal Accounting Information System (AIMS)*, vol. 3, no. 1, pp. 15–24, 2020.

[8] W. Mulyana, Aryanto, and M. Aprilia, “Penerapan Metode Single Exponential Smoothing Untuk Prediksi Kasus Positif COVID 10 di Kabupaten Bengkalis,” *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 3, no. 3, pp. 415–421, Dec. 2022, doi: 10.37859/coscitech.v3i3.4363.

[9] S. Joya Arditna Br Bukit and R. R. Kurniawan, “Prediksi Harga Tandan Buah Segar dengan Algoritma K-Nearest Neighbor,” *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 5, no. 1, pp. 92–101, 2023, doi: 10.30865/json.v5i1.6818.

[10] L. M. Ginting, M. M. T. Sigiro, E. D. Manurung, and J. J. P. Sinurat, “Perbandingan Metode Algoritma Support Vector Regression dan Multiple Linear Regression Untuk Memprediksi Stok Obat,” *Journal of Applied Technology and Informatics Indonesia*, vol. 1, no. 2, pp. 29–34, 2021.

[11] V. A. P. Putri, A. B. Prasetijo, and D. Eridani, “Perbandingan Kinerja Algoritme Naïve Bayes Dan K-Nearest Neighbor (Knn) Untuk Prediksi Harga Rumah,” *Transmisi: Jurnal Ilmiah Teknik Elektro*, vol. 24, no. 4, pp. 162–171, 2022.

[12] A. N. Rais, W. Warjiyono, I. Alfarobi, S. W. Hadi, and W. Kurniawan, “Analisa Prediksi Harga Jual Rumah Menggunakan Algoritma Random Forest Machine Learning,” *Jurnal Riset Sistem Informasi dan Teknologi Informasi (JURSISTEKNI)*, vol. 6, no. 2, pp. 416–423, 2024.

[13] 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 (JACOST)*, vol. 4, no. 1, pp. 2723–1453, 2023, doi: 10.52158/jacost.491.

[14] D. Sepri and A. Fauzi, “Prediksi Harga Cabai Merah Menggunakan Support Vector Regression,” *Computer Based Information System Journal*, vol. 8, no. 2, pp. 1–5, 2020.

[15] D. Ismafillah, T. Rohana, and Y. Cahyana, “Implementasi Model Support Vector Machine dan Logistic Regression Untuk Memprediksi Penyakit Stroke,” *Jurnal Riset Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5478.

[16] A. Vermaysha and N. Nurmalitasari, “Prediksi Harga Rumah di Kabupaten Karanganyar Menggunakan Metode Regresi Linear,” in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 2023, pp. 6–11.

[17] M. L. Subiyanto, Y. Amanda, M. N. Fachrian, A. Y. B. Rohim, and N. Chamidah, “Peramalan Kasus Harian Monkeypox Dunia Berdasarkan Metode Support Vector Regression (SVR),” *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 15, no. 1, pp. 27–36, 2023.

[18] S. Balivada, G. Grant, X. Zhang, M. Ghosh, S. Guha, and R. Matamala, “A wireless underground sensor network field pilot for agriculture and ecology: Soil moisture mapping using signal attenuation,” *Sensors*, vol. 22, no. 10, p. 3913, 2022.