BAB III
METODE PENELITIAN

3.1 Objek Penelitian

Penelitian ini menggunakan dataset harga rumah di wilayah Bogor sebagai objek penelitian untuk menganalisis dan memprediksi harga rumah. Dataset tersebut diambil dari jurnal karya Ariq & Kusumastuti, (2024) yang berjudul *"Analisis Algoritma Backpropagation pada Jaringan Syaraf Tiruan dalam Memprediksi Harga Rumah di Jabodetabek"* dan tersedia secara *open-source* di platform *Kaggle*.

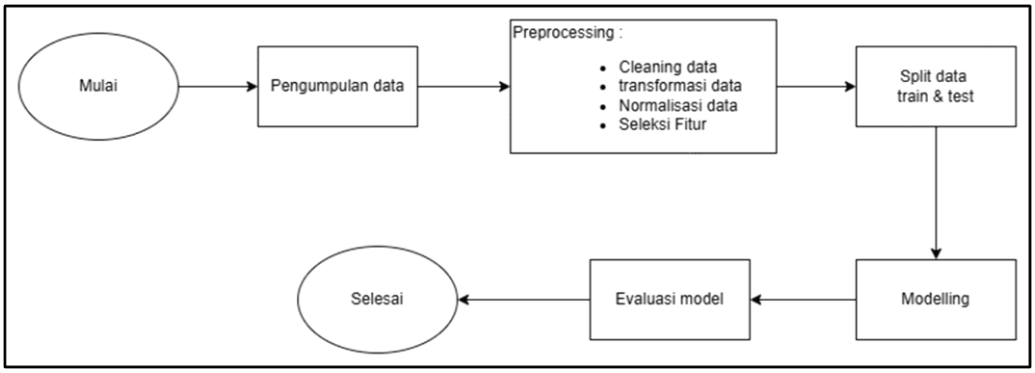
Dataset ini berisi daftar harga rumah di Jabodetabek yang diambil dari situs rumah123.com pada akhir 2022, lengkap dengan spesifikasi setiap properti. Tabel 3.1 menyajikan cuplikan data tersebut.

Tabel 3. 1 Spesifikasi Rumah dalam Dataset

No	Judul Properti	Kecamatan	...	Kondisi	Harga (Rp)
1	Rumah Strategis di Pasir Kuda Dekat Stasiun Bogor	Pasirmulya	...	Butuh renovasi	690,000,000
2	Rumah Asri Siap Huni di Taman Yasmin	Cilendek Timur	...	Bagus	995,000,000
3	Rumah Full Furnished Taman Yasmin	Cilendek Timur	...	Bagus	4,500,000,000
4	Rumah Mewah Cluster Premium Sentul City	Sentul City	...	Bagus	11,500,000,000
.	...	...	...	...	...
881	...	...	...	...	...

3.2 Prosedur Penelitian

Langkah-langkah penelitian ini meliputi beberapa tahap yang akan dijelaskan secara rinci pada Gambar 3.1



Gambar 3. 1 Prosedur Penelitian

Penelitian ini terdiri dari enam tahapan yang dilakukan secara berurutan. Tahap pertama adalah pengumpulan data, diikuti oleh *preprocessing* yang mencakup proses pembersihan data (*data cleaning*), transformasi data, normalisasi data, dan seleksi fitur. Tahap berikutnya adalah *Split* data. Selanjutnya, dilakukan penerapan algoritma dan evaluasi metrik untuk menilai kinerja model yang dihasilkan.

3.2.1 Pengumpulan Data

Pengumpulan data dalam penelitian ini tidak dilakukan secara langsung di lapangan, melainkan dengan mengakses dan mengunduhnya melalui situs Kaggle pada tautan: https://www.kaggle.com/datasets/nafisbarizki/daftar-harga-rumah-jabodetabek , yang diakses pada 15 Desember 2024 pukul 13:00 WIB. Data yang diperoleh adalah data daftar harga rumah dijual di bogor. Data berjumlah 881 dengan 27 kolom. Adapun kolom – kolom yang terdapat dalam dataset.

Tabel 3. 2 Kolom atribut

Feature	Penjelasan
url	Memuat data yang diambil dari link halaman rincian penjualan rumah pada situs rumah123.com
title	Berisikan judul rumah
address	Berisikan keterangan alamat dari rumah pada iklan.

district	Memuat informasi kecamatan rumah yang ada pada iklan.
city	memuat keterangan kota rumah yang akan dijual.
lat	Menampilkan informasi koordinat latitude sesuai dengan posisi properti.
long	Memuat data longitude berdasarkan letak geografis properti.
facilities	Memuat daftar fasilitas yang tersedia di dalam rumah.
property_type	Berisikan tipe properti (misalnya, rumah, apartemen).
ads_id	Berisikan ID unik untuk iklan properti.
bedrooms	Memuat berapa banyak kamar tidur pada rumah.
bathrooms	Berisikan jumlah kamar mandi di rumah.
land_size_m2	Berisikan luas tanah dalam meter persegi.
building_size_m2	Berisikan luas bangunan dalam meter persegi.
carports	Berisikan luas atau kapasitas tempat mobil.
certificate	Memuat sertifikat pada saat proses trasnsaksi jual beli hunian / rumah (misalnya, SHM, HGB).

electricity	Memuat informasi mengenai daya listrik yang tersedia di rumah.
maid_bedrooms	Berisikan jumlah kamar tidur untuk pembantu.
maid_bathrooms	Menunjukkan jumlah kamar tidur yang disediakan untuk asisten rumah tangga.
floors	Memuat berapa banyak lantai yang ada pada rumah tersebut.
building_age	Memuat usia hunian atau rumah dalam tahun.
year_built	Berisikan tahun pembuatan rumah.
property_condition	Berisikan keterangan kondisi bangunan (misalnya, baru, renovasi).
building_orientation	Berisikan orientasi bangunan (misalnya, menghadap timur, selatan).
garages	Berisikan kapasitas garasi pada rumah.
furnishing	Berisikan kondisi perabotan rumah (misalnya, furnished, unfurnished).
price_in_rp	Berisikan data harga rumah dalam Rupiah.

Untuk tujuan penelitian ini, akan diambil 10 atribut yang terdiri dari: *building_size_m2, floors, land_size_m2, bedrooms, maid_bedrooms, bathrooms, garages, maid_bathrooms, year_built, dan price_in_rp*. Berdasarkan penelitian

sebelumnya, seperti yang dilakukan oleh Maula et al., (2023), atribut-atribut ini dianggap relevan karena memiliki pengaruh signifikan terhadap harga rumah.

3.2.2 *Preprocessing*

Tahap preprocessing dilakukan sebelum proses utama pengolahan data. Tahapan ini mencakup tahapan seperti pembersihan data, transformasi data, normalisasi data dan seleksi fitur.

1. Pembersihan data (*cleaning data*)

   Langkah pembersihan data dilakukan di awal proses analisis untuk menjamin bahwa informasi yang dipakai telah memenuhi standar kualitas. Proses ini mencakup beberapa langkah penting, antara lain:

   a. Penghapusan Atribut yang Tidak Relevan

      Beberapa atribut yang dianggap tidak relevan terhadap proses analisis dihapus agar fokus penelitian dapat diarahkan pada data yang lebih signifikan. Proses penghapusan dilakukan menggunakan fungsi *drop()*. Atribut – atribut yang dihapus meliputi: *property_type*, url, *address, city, lat, long, district, facilities, title, ads_id, certificate, electricity, property_condition, building_orientation, furnishing, building_age,* dan *carports*. Dengan menghapus kolom-kolom ini, fokus analisis dapat diarahkan pada data yang lebih relevan.

   b. Pemeriksaan *Missing Value*

      Dataset diperiksa untuk mendeteksi adanya nilai yang hilang (*missing value*) pada setiap kolom. Kehadiran nilai yang hilang dapat memengaruhi akurasi analisis dan hasil prediksi, sehingga penting untuk mengidentifikasinya sejak awal. Pemeriksaan dilakukan dengan menggunakan fungsi *isnull()* yang dikombinasikan dengan *sum()* untuk menghitung jumlah nilai yang hilang pada setiap kolom. Hasil pemeriksaan ini ditampilkan dalam bentuk angka, sehingga memudahkan dalam menentukan langkah

penanganan selanjutnya, seperti melakukan imputasi atau menghapus baris yang mengandung missing value.

```
df.isnull().sum()
```

Gambar 3. 2 Fungsi *isnull().sum()*

c. Pemeriksaan Duplikasi Data

Pemeriksaan dilakukan untuk mendeteksi baris-baris data yang memiliki nilai identik di seluruh kolom, atau disebut sebagai data duplikat. Proses ini menggunakan fungsi *duplicated()* dari pustaka pandas. Fungsi ini akan menghasilkan nilai *True* untuk setiap baris yang terdeteksi sebagai duplikat dari baris sebelumnya. Untuk menghitung jumlah total baris duplikat dalam dataset dapa dilihat pada gambar 3.3

```
#menghitung jumlah duplikat
jumlah_duplikat = df.duplicated().sum()
```

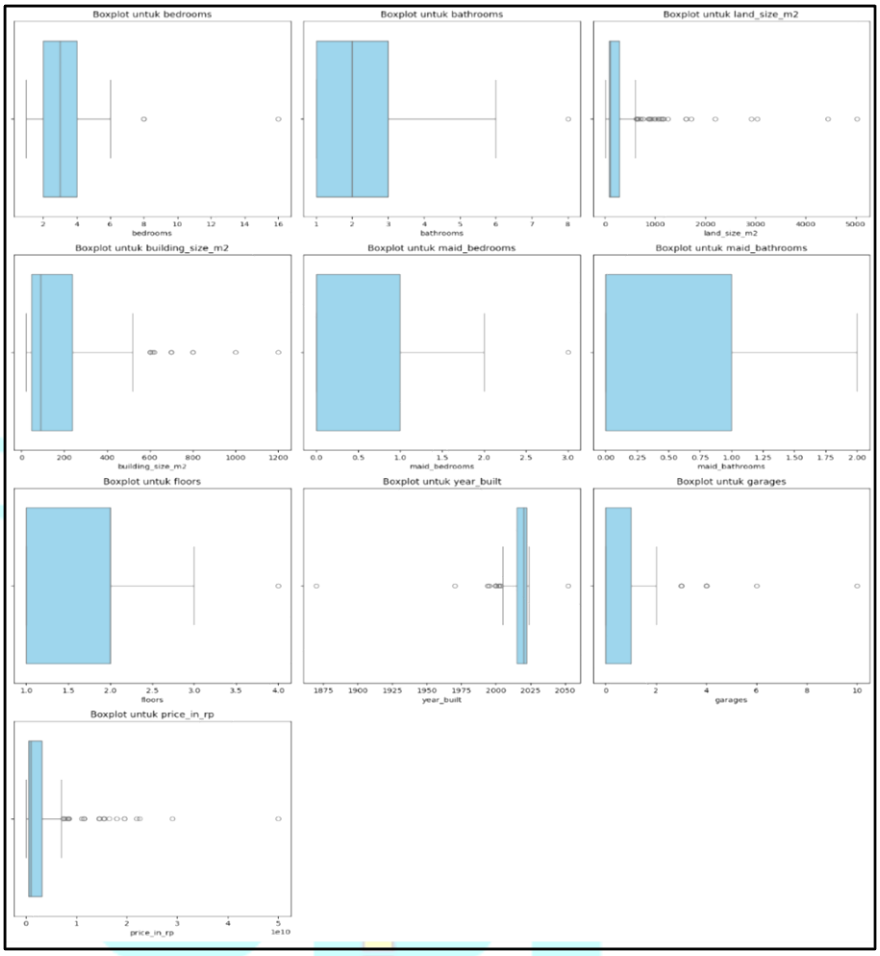
Gambar 3. 3 Fungsi *duplicated()*

d. Outlier

Outlier mengacu pada elemen data yang tidak mengikuti pola umum dari sebagian besar data. Keberadaan outlier dapat memengaruhi hasil analisis, sehingga penting untuk mendeteksinya. Salah satu metode yang digunakan untuk mendeteksi outlier adalah *Interquartile Range* (IQR). Rumus IQR adalah sebagai berikut:

$$IQR = Kuartil3 - Kuartil1 \quad (5)$$

Identifikasi *outlier* dilakukan untuk mendeteksi nilai-nilai yang tidak wajar, yang dapat memengaruhi hasil analisis. Berikut adalah hasil deteksi outlier pada beberapa atribut dalam dataset:



Gambar 3. 4 Hasil deteksi Outlier

- Jumlah Kamar Tidur (*bedrooms*): Outlier ada pada 8, 12, dan 16 kamar tidur.
- Jumlah Kamar Mandi (*bathrooms*): Outlier ada pada 7 dan 8 kamar mandi.
- Luas Tanah dalam m² (*land_size_m2*): Outlier ada pada berbagai titik di luar 1000 m², dengan outlier ekstrem sekitar 4000 dan 5000 m².
- Luas Bangunan dalam m² (*building_size_m2*): Outlier ada pada berbagai titik di luar 400 m², dengan outlier ekstrem sekitar 1000 dan 1200 m².
- Jumlah Kamar Pembantu (*maid_bedrooms*): Outlier ada pada 2 dan 3 kamar pembantu.

- Jumlah Kamar Mandi Pembantu (*maid_bathrooms*): Outlier ada pada 2 kamar mandi pembantu.
- Jumlah Lantai (*floors*): Outlier ada pada 3 dan 4 lantai.
- Tahun Dibangun (*year_built*): Outlier ada pada berbagai titik sebelum tahun 1950 dan setelah tahun 2025.
- Jumlah Garasi (*garages*): Outlier ada pada berbagai titik di luar 4 garasi, dengan outlier ekstrem sekitar 8 dan 10 garasi.
- Harga dalam Rupiah (*price_in_rp*): Outlier ada pada berbagai titik di luar 2 x 10^{10} Rp, dengan outlier ekstrem sekitar 4 x 10^{10} Rp

e. Noise

Noise adalah data yang tidak relevan, tidak valid, atau mengandung informasi yang menyimpang dari pola umum, sehingga dapat mengganggu proses analisis dan menurunkan kualitas model. Untuk mendeteksi data *noise*, salah satu metode yang digunakan adalah perhitungan *z-score*. Nilai *z-score* mengukur seberapa jauh suatu data menyimpang dari rata-rata dalam satuan standar deviasi. Nilai *z-score* yang terlalu tinggi atau terlalu rendah (misalnya, di luar rentang -3 hingga 3) dapat dianggap sebagai *noise.*

```
# Hitung z-score
z_scores = (df[numeric_columns] - df[numeric_columns].mean()) / df[numeric_columns].std()

# Temukan data noise (z-score di luar rentang -3 hingga 3)
noise = (z_scores < -3) | (z_scores > 3)
```

Gambar 3. 5 Deteksi *Noise*

2. Transformasi data

Pada tahap ini, dilakukan transformasi pada atribut harga dengan membagi setiap nilainya dengan satu juta. Transformasi ini bertujuan untuk mempermudah pembacaan data.

Tabel 3. 3 Transformasi Atribut Data

Sebelum	Sesudah
2300000000	2300

2300000000	2300
2750000000	2750
2400000000	2400

3. Normalisasi Data

Proses normalisasi data dilakukan menggunakan metode *MinMaxScaler*, yang berfungsi mengubah skala data ke dalam rentang tertentu, biasanya antara 0 dan 1. Tujuan dari langkah ini adalah untuk menyamakan skala antar fitur, sehingga model machine learning dapat bekerja secara lebih optimal dan tidak bias terhadap fitur dengan skala nilai yang lebih besar. Pada proses normalisasi akan menggunakan metode *MinMaxScaler* dengan rumus sebagai berikut:

$$x_{secaled}=\frac{x-x_{min}}{x_{max}-x_{min}} \quad (6)$$

Tabel 3. 4 Keterangan *Minmaxscaler*

Simbol	Keterangan
Xscaled	Nilai data yang telah dinormalisasi (skala antara 0 dan 1).
X	Nilai asli data.
Xmin	Data dengan nilai paling kecil.
Xmax	Data dengan nilai paling besar

4. Seleksi Fitur

Pada tahap ini, dilakukan analisis korelasi untuk mengukur sejauh mana hubungan antara atribut-atribut dalam dataset. Ketika nilai korelasi hampir mencapai 1, hal ini menandakan keterkaitan positif yang tinggi antara dua atribut. Sebaliknya, jika nilainya mendekati 0, maka kedua atribut tersebut cenderung tidak memiliki relasi yang berarti.

3.2.3 *Modelling*

Pada tahap modelling, dibangun dua model prediksi menggunakan algoritma *Random Forest* dan *Gradient Boosting*. Untuk memperoleh hasil prediksi yang optimal, dilakukan proses *hyperparameter tuning* dengan pendekatan yang berbeda: *GridSearchCV* digunakan untuk *Random Forest* karena mampu mengevaluasi semua kombinasi parameter secara menyeluruh, sedangkan *RandomizedSearchCV* digunakan untuk *Gradient Boosting* guna menghemat waktu komputasi dengan mengevaluasi kombinasi parameter secara acak. Kedua proses tuning ini menggunakan teknik *cross-validation* untuk memastikan model yang dihasilkan memiliki generalisasi yang baik terhadap data baru (Firdaus, 2022).

3.2.4 Evaluasi Model

Evaluasi dilakukan untuk menilai performa prediksi yang dihasilkan. Dua metrik utama yang digunakan dalam evaluasi ini adalah *RMSE* yang berkepanjangan *Root Mean Square Error* dan juga *R-squared* (R^2)

1. *Root Mean Square Error (RMSE)*

   RMSE dipakai guna mengukur besarnya rata-rata presiksi yang salaha atau kesalahan prediksi yang dilakukan oleh model dengan satuan yang sama seperti data asli. Rumus RMSE dapat ditulis sebagai berikut:

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \Sigma (y_i - y)^2} \quad (7)$$

Tabel 3. 5 Keterangan *RMSE*

Simbol	Keterangan
yi	Muatan aktual
y	Muatan yang diprediksi
n	Banyaknya data

2. *R²-squared*

R^2-squared atau Koefisien Determinasi merupakan indikator statistik yang berfungsi untuk mengukur kemampuan model regresi dalam menjelaskan keragaman pada nilai target. Skor R^2 berada dalam rentang antara nilai (nol) 0 hingga nilai (satu) 1, yang berarti juga nilai yang semakin mendekati (satu) 1 mengindikasikan bahwa model dapat merepresentasikan variasi data dengan tingkat akurasi yang tinggi. Sebaliknya, nilai R^2 yang mendekati 0 menunjukkan bahwa kemampuan model dalam menangkap hubungan antara variabel input dan output sangat terbatas. Secara sederhana, R^2 menggambarkan proporsi variasi data yang dapat dijelaskan oleh model dibandingkan dengan total variasi yang ada. Rumus perhitungan R^2 dapat ditulis sebagai berikut:

$$R^2 = 1 - \frac{Selisih\ Kuadrat\ Prediksi}{Selisih\ Kuadrat\ Total} \quad (8)$$