

Prediksi Harga Rumah menggunakan Algoritma *Random Forest Regression*

House Price Prediction using the Random Forest Regression Algorithm

¹Fika Halimah Balqis*, ²Qurrotul Aini

^{1,2}Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Syarif Hidayatullah Jakarta

^{1,2}Jl. Ir H. Juanda No.95, Ciputat, Kec. Ciputat Tim., Kota Tangerang Selatan, Banten 15412

*e-mail: qurrotul.aini@uinjkt.ac.id

(received: 28 August 2025, revised: 8 December 2025, accepted: 10 January 2026)

Abstrak

Prediksi harga rumah merupakan masalah kompleks karena dipengaruhi berbagai faktor seperti kualitas bangunan, lokasi, dan luas area tinggal sehingga metode konvensional sering kali kurang akurat dalam mengestimasi. Penelitian ini bertujuan menerapkan algoritma *Random Forest Regression* (RFR) untuk memprediksi harga rumah pada *dataset House Prices – Advanced Regression Techniques* dari Kaggle yang berisi 1.460 data properti. Metode SEMMA (*Sample, Explore, Modify, Model, Assess*) digunakan karena memiliki alur kerja yang sistematis dan memiliki fokus yang terstruktur sehingga model lebih dapat diandalkan. Pada tahap pemodelan, RFR diterapkan karena mampu menangani pola non-linier dan performanya tetap stabil pada jumlah fitur yang besar. Berdasarkan pengujian, model menghasilkan nilai *Root Mean Squared Error* (RMSE) sebesar 28.452,75 dan koefisien determinasi 89%, diikuti uji *robustness* sebesar 30.665,40 yang mengindikasikan stabilitas model. Analisis *feature importance* juga menunjukkan bahwa *OverallQual* memiliki pengaruh terbesar dalam penentuan harga. Temuan ini menegaskan bahwa RFR diandalkan untuk memprediksi harga rumah serta berpotensi untuk dikembangkan dalam sistem rekomendasi harga, penilaian properti otomatis, dan integrasi ke platform digital industri *real estate*.

Kata kunci: harga rumah, *random forest regression*, R^2 , RMSE

Abstract

House price prediction is a complex problem because it is influenced by various factors such as building quality, location, and living area size. As a result, conventional methods often lack accuracy in estimating housing prices. This study aims to apply the *Random Forest Regression* (RFR) algorithm to predict house prices using the *House Prices – Advanced Regression Techniques* dataset from Kaggle, which contains 1,460 property records. The SEMMA (*Sample, Explore, Modify, Model, Assess*) methodology was adopted due to its systematic workflow and structured focus, which improves the reliability of the developed model. In the modeling stage, RFR was implemented because it is capable of handling non-linear patterns and maintains stable performance even with a large number of features. Based on the evaluation results, the model achieved a *Root Mean Squared Error* (RMSE) of 28,452.75 and a coefficient of determination (R^2) of 89%. This was followed by a *robustness* test with an RMSE of 30,665.40, indicating the stability of the model. *Feature importance* analysis also revealed that *OverallQual* had the greatest influence on house price prediction. These findings confirm that *Random Forest Regression* is a reliable method for predicting house prices and has strong potential to be further developed for price recommendation systems, automated property valuation, and integration into digital platforms within the real estate industry.

Keywords: house prices, *random forest regression*, R^2 , RMSE

1 Pendahuluan

Prediksi merupakan proses membangun model yang dilatih dari data historis untuk memperkirakan nilai pada data baru yang belum diketahui [1]. Pendekatan ini menjadi inti dari

machine learning, yang secara khusus dikembangkan untuk menghasilkan model prediktif berdasarkan data historis, seperti harga, risiko, maupun probabilitas suatu peristiwa [1], [2]. Dalam ilmu data modern, prediksi berperan penting dalam pengambilan keputusan yang objektif dan berbasis data [3].

Sektor properti merupakan salah satu domain yang sangat membutuhkan kemampuan prediktif. Rumah tidak hanya berfungsi sebagai tempat tinggal, tetapi juga sebagai instrumen investasi strategis [4], [5], [6]. Nilai rumah dipengaruhi oleh banyak faktor, mulai dari kondisi ekonomi makro, kebijakan pemerintah, lokasi, hingga luas bangunan dan jumlah kamar [7], [8]. Kompleksitas ini menjadikan prediksi harga rumah sebagai tantangan multidimensi yang memerlukan pendekatan analitis yang tepat.

Secara umum, terdapat dua pendekatan dalam menentukan harga rumah, yaitu konvensional dan digital [9], [10]. Pendekatan konvensional mengandalkan penilai profesional yang menggunakan pengalaman subjektif, sedangkan pendekatan digital memanfaatkan algoritma *machine learning* untuk mengolah data besar secara lebih sistematis dan objektif [11], [12], [13]. Meskipun demikian, penggunaan algoritma *machine learning* juga menghadapi tantangan tersendiri, salah satunya risiko *overfitting*, yaitu kondisi model terlalu menyesuaikan diri dengan data *training* sehingga performanya menurun pada data baru. Hal ini sering terjadi pada *dataset* dengan banyak variabel, pola yang non-linier, atau sebaran fitur yang tidak merata dimana itu merupakan karakteristik umum pada data properti. Karena itu, memilih algoritma yang mampu mengurangi risiko *overfitting* menjadi langkah krusial agar prediksi harga rumah tetap akurat dan dapat diandalkan.

Urgensi akan metode prediksi digital yang akurat semakin meningkat seiring dengan dinamika pasar perumahan global. Berdasarkan data jangka panjang dari Bank for International Settlements (BIS), harga rumah terus mengalami peningkatan global. Harga rumah riil global berada 21 % di atas level pasca-Krisis Keuangan 2007-2009 [14] seperti pada Gambar 1.



Gambar 1 Tren global harga rumah riil

Fakta ini mengindikasikan bahwa harga rumah secara umum mengalami peningkatan di banyak negara. Oleh karena itu, prediksi harga dengan algoritma yang akurat sangat diperlukan untuk mendukung berbagai keputusan strategis. Algoritma *Random Forest Regression* (RFR) banyak digunakan karena terbukti stabil, lebih tahan terhadap *overfitting* dibandingkan model berbasis pohon tunggal, serta mampu menangani relasi non-linear dan data dengan noise [15], [16], [17], [18].

Penelitian ini bertujuan untuk menerapkan dan memperoleh kinerja RFR dalam memprediksi harga rumah menggunakan *dataset House Prices – Advanced Regression Techniques* dari Kaggle [19]. Penelitian ini juga diharapkan memberikan beberapa kontribusi praktis seperti menerapkan model RFR untuk memprediksi harga rumah, memanfaatkan Google Colab dan Python dalam penerapan model, serta performa model pada data properti terkini.

2 Tinjauan Literatur

Penelitian mengenai prediksi harga rumah telah banyak memanfaatkan berbagai metode *machine learning*, termasuk *Random Forest Regression* (RFR). Dengan menggunakan kerangka SEMMA

(*Sample, Explore, Modify, Model, Assess*), literatur terkait dapat dianalisis secara sistematis mulai dari pemilihan sampel hingga penilaian performa model.

Penelitian sebelumnya banyak menggunakan data properti yang memuat informasi seperti ukuran bangunan, lokasi, dan kondisi fisik rumah karena faktor-faktor tersebut sangat memengaruhi nilai suatu hunian [10], [11]. Di sisi lain, tren kenaikan harga rumah secara global menegaskan pentingnya memilih data yang benar-benar mewakili kondisi pasar [14]. Hal ini juga terlihat pada penelitian sebelumnya yang memiliki 676 data rumah di Nigde [6] dan 28.550 transaksi rumah di Beijing yang dilengkapi informasi lingkungan sekitar [12]. Dari temuan tersebut, dapat dipahami bahwa *dataset* dengan variasi fitur yang beragam cocok untuk membangun model yang prediktif.

Eksplorasi fitur juga dilakukan untuk memahami faktor yang memengaruhi harga rumah, seperti kualitas bangunan, fasilitas lingkungan, dan konteks geografis [11], [14], [17]. Selain faktor mikro tersebut, studi tingkat nasional menyoroti pengaruh ekonomi makro terhadap harga properti di Indonesia [7]. Diketahui juga dari sisi pasar pembiayaan, minat terhadap Kredit Pemilikan Rumah (KPR) turut mendorong permintaan perumahan sehingga berpengaruh pada nilai pasar properti [8]. Dengan cakupan faktor yang luas tersebut, Naz et al. menegaskan perlunya pemilihan fitur yang sesuai dengan karakteristik data properti agar prediksi tetap akurat [9]. Karena itu, eksplorasi awal dalam penelitian ini difokuskan untuk menemukan fitur yang paling berpengaruh terhadap harga rumah sebelum masuk ke tahap selanjutnya.

Tahap modifikasi data dilakukan untuk memastikan kualitas input model, antara lain melalui pembersihan data, imputasi, *encoding*, dan seleksi fitur [10], [17], [20]. Dalam penelitian lainnya, Zhao et al. menerapkan *geocoding* dan *feature engineering* untuk meningkatkan representasi fitur pada data properti [12] sedangkan Ulucan et al. menambahkan variabel temporal agar model mampu menangkap perubahan harga selama pandemi [6]. Hal-hal tersebut dilakukan untuk memastikan kinerja model lebih meningkat saat diterapkan.

Berbagai studi menunjukkan bahwa RFR, *Gradient Boosting*, dan *Artificial Neural Network* (ANN) merupakan model yang kuat dalam memprediksi harga rumah karena mampu menangani pola data yang kompleks [10], [11], [17]. Penelitian sebelumnya juga membuktikan bahwa RFR memiliki performa yang stabil dan akurat pada data properti [6] serta bekerja sangat baik ketika fitur berasal dari berbagai sumber data [12]. Selain itu, Naz et al. menegaskan bahwa RFR unggul karena kemampuannya mengurangi risiko *overfitting* [9].

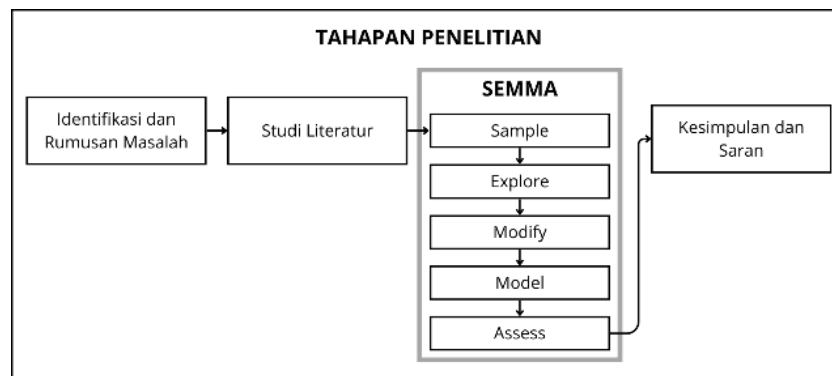
Evaluasi performa model pada penelitian terdahulu umumnya dilakukan menggunakan metrik *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan koefisien determinasi (R^2) untuk menilai tingkat akurasi prediksi [10], [11], [17]. Dalam penelitian sebelumnya, Ulucan et al. menggunakan R^2 dan *Price Related Differential* (PRD) untuk mengukur kemampuan model dalam menjelaskan variasi harga [6], sementara Zhao et al. menekankan perlunya perbandingan antar-model untuk memastikan prediksi yang stabil pada berbagai kondisi data [12].

Secara keseluruhan, kerangka SEMMA memiliki keunggulan dalam analisis literatur maupun penelitian prediktif karena menyediakan alur kerja yang jelas dan terstruktur. Tahap eksplorasi dan modifikasi data memastikan bahwa model dibangun dari input yang berkualitas. Selain itu, pemisahan tahap pemodelan dan penilaian memudahkan pengukuran performa secara objektif.

Berdasarkan tinjauan ini, temuan menunjukkan bahwa kualitas bangunan, luas, dan lokasi merupakan faktor utama penentu harga rumah [11], [17]. Peningkatan kualitas prediksi juga dipengaruhi oleh *preprocessing* data yang tepat [10], [12], [20]. Selain itu, RFR dinilai efektif untuk data properti yang kompleks [6], [9] dengan RMSE dan R^2 sebagai evaluasi model yang diterapkan [10], [17]. Dengan demikian, penggunaan SEMMA dan RFR dalam penelitian ini konsisten dengan temuan dan pendekatan dalam penelitian sebelumnya.

3 Metode Penelitian

Penelitian ini mengadopsi pendekatan SEMMA (*Sample, Explore, Modify, Model, Assess*) sebagai metodologi dalam tahapan penambangan data. Pendekatan ini digunakan untuk memandu proses ekstraksi pengetahuan dari kumpulan data dalam skala besar secara terstruktur. Seluruh tahapan penelitian dilaksanakan secara sistematis sebagaimana ditunjukkan pada Gambar 2.



Gambar 2 Diagram tahapan penelitian

Tahapan sistematis [9], [20] tersebut dijabarkan sebagai berikut:

a. Identifikasi dan Rumusan Masalah

Tahap awal penelitian ini berfokus pada identifikasi isu terkait kebutuhan prediksi harga rumah yang akurat dan efisien. Rumusan ini menjadi dasar dalam penyusunan tujuan dan arah metodologi penelitian secara sistematis.

b. Studi Literatur

Menelaah penelitian terdahulu yang berkaitan dengan prediksi harga rumah menggunakan RFR, guna memperoleh landasan teoretis dan metodologis.

c. Penambahan Data

Penelitian ini mengadopsi pendekatan SEMMA (*Sample, Explore, Modify, Model, Assess*) yang mencakup lima tahapan utama [21], sebagai berikut:

1) *Sample*

Pada tahap *Sample*, penelitian ini menggunakan *dataset* sekunder *House Prices – Advanced Regression Techniques* yang tersedia untuk umum di platform Kaggle [19]. *Dataset* ini berisi 1.460 data properti residensial dengan berbagai informasi seperti luas bangunan, jumlah ruangan, lokasi, tahun renovasi, dan kualitas bangunan. Semua fitur tersebut digunakan untuk memprediksi variabel target yaitu harga rumah (*SalePrice*) yang dinyatakan dalam dolar AS.

2) *Explore*

Memasuki tahap *Explore*, peneliti mulai mengenali karakteristik *dataset* secara lebih mendalam. Analisis dilakukan untuk melihat bagaimana data terdistribusi, mengidentifikasi keberadaan nilai ekstrim (*outlier*), dan mengetahui seberapa kuat hubungan setiap fitur dengan harga rumah. Berbagai visualisasi, seperti *heatmap*, *scatterplot*, dan *histogram*, membantu menunjukkan pola yang mungkin tidak langsung terlihat dari angka. Hasil eksplorasi ini kemudian menjadi acuan penting dalam menentukan fitur-fitur yang layak dipertahankan pada tahap selanjutnya.

3) *Modify*

Pada tahap *Modify*, data dipersiapkan agar dapat digunakan secara maksimal oleh model. Proses ini mencakup penanganan nilai yang hilang melalui imputasi, mengubah fitur kategorikal menjadi numerik, dan memisahkan *dataset* menjadi *training set* dan *test set*. Dengan begitu, data yang digunakan dalam pemodelan menjadi lebih bersih, konsisten, dan siap diolah.

4) *Model*

Pada tahap *Model*, penelitian ini menggunakan algoritma *Random Forest Regression* (RFR) untuk melakukan pemodelan. Algoritma ini bekerja dengan membangun banyak pohon keputusan dari data yang dipilih secara acak (*bagging*) [15], [16]. Setiap pohon memberikan prediksi yang mungkin berbeda-beda, kemudian semua prediksi dirata-ratakan untuk menghasilkan nilai akhir yang lebih stabil dan akurat. Mekanisme ini ditunjukkan pada Persamaan (1).

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k T_i(x) \quad (1)$$

Keterangan:

$\hat{y}$: nilai prediksi akhir
 $T_i(x)$: prediksi dari pohon ke- i
 k : jumlah total pohon dalam hutan

Proses pelatihan model dilakukan menggunakan pustaka *scikit-learn* di Google Colab. Selain itu, beberapa *hyperparameter* telah disesuaikan untuk meningkatkan performa model, seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), serta parameter lain seperti *min_samples_split*, *min_samples_leaf*, dan *max_features*.

5) Assess

Setelah model dilatih, dilakukan tahap *Assess* untuk mengevaluasi kinerjanya. Dua metrik utama yang digunakan yaitu *Root Mean Squared Error* (RMSE) dan koefisien determinasi (R^2). RMSE menunjukkan seberapa jauh hasil prediksi menyimpang dari nilai sebenarnya [1]. Semakin kecil nilai RMSE maka performa model semakin baik. Nilai RMSE di bawah 10-15% dari rata-rata target umumnya dianggap cukup baik [22]. Rumus perhitungan RMSE dapat dilihat pada Persamaan (2).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

Keterangan:

y_i : nilai aktual
 $\hat{y}_i$: nilai prediksi
 n : jumlah observasi

Selain itu model juga diukur menggunakan R^2 untuk melihat seberapa besar variasi harga rumah yang dapat dijelaskan oleh model tersebut. Nilai R^2 berada pada rentang 0 sampai 1. Semakin mendekati 1 maka kemampuan model dalam menjelaskan data semakin baik [11]. Rumus R^2 disajikan pada Persamaan (3).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

Keterangan:

$\bar{y}$: rata-rata nilai aktual

Dengan tahap ini, performa model dapat diukur secara kuantitatif dan kemampuan generalisasi terhadap data baru dapat dinilai. Selain itu, dilakukan pula analisis *feature importance* untuk mengidentifikasi fitur paling berpengaruh serta evaluasi *robustness* untuk menilai kestabilan performa model pada berbagai *subset* data.

d. Kesimpulan dan Saran

Bagian ini berisi rangkuman temuan utama penelitian berdasarkan hasil pemodelan dan evaluasi kinerja model. Selain itu, bagian ini juga memberikan rekomendasi untuk pengembangan penelitian selanjutnya.

4 Hasil dan Pembahasan

Penelitian ini memiliki hasil sebagai berikut:

4.1 Sample

Dataset ini berisi informasi lebih dari 1.400 rumah, mencakup fitur-fitur seperti ukuran rumah, jumlah kamar mandi, kondisi bangunan, hingga lokasi geografis. Sebelum proses pembacaan dan pengolahan data dilakukan, sejumlah pustaka (*library*) penting terlebih dahulu diimpor untuk mendukung tahapan *Explore*, *Modify*, *Model*, dan *Assess*. Pustaka-pustaka tersebut meliputi *pandas*, *numpy*, *matplotlib*, *seaborn*, serta beberapa modul dari *scikit-learn*. Setelah proses unggah *dataset* ke lingkungan kerja Google Colab selesai, *file train.csv* secara otomatis tersimpan di direktori utama dan siap untuk dibaca menggunakan pustaka *pandas* seperti ditunjukkan pada Gambar 3.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import OneHotEncoder
```

Gambar 3 Import library

Untuk mengunggah *dataset* ke lingkungan kerja Google Colab, ditunjukkan potongan kode seperti pada Gambar 4.

```
from google.colab import files
uploaded = files.upload()
```

Choose Files train.csv

train.csv(text/csv) - 460676 bytes, last modified: 6/23/2025 - 100% done
Saving train.csv to train.csv

Gambar 4 Unggah dataset

Setelah proses unggah selesai, *file train.csv* secara otomatis tersimpan di direktori utama dan siap dibaca menggunakan pustaka *pandas* seperti ditunjukkan pada Gambar 5.

```
import pandas as pd
df = pd.read_csv('train.csv')
```

Gambar 5 Membaca dataset

Tahapan ini memastikan bahwa *dataset* telah tersedia dalam bentuk *dataframe* sehingga dapat digunakan pada proses berikutnya.

4.2 Explore

Langkah ini bertujuan untuk mengidentifikasi Fitur Penting. Langkah awal dalam proses eksplorasi data adalah memahami struktur *dataset* yang digunakan. Hal ini dilakukan dengan menampilkan dimensi data serta informasi mengenai tipe dan jumlah atribut yang tersedia, kemudian dilakukan analisis korelasi antara fitur numerik dan variabel target *SalePrice* menggunakan *heatmap* sebagaimana ditunjukkan pada Gambar 6.

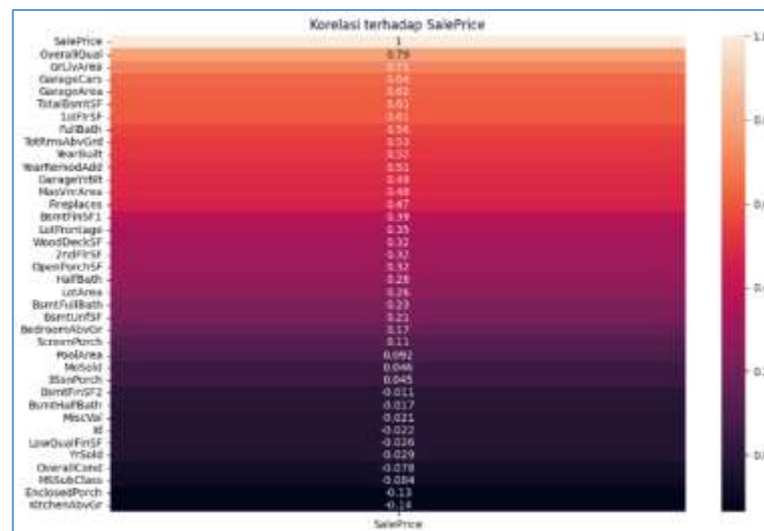
```
# Cek struktur data
print(df.shape)
print(df.info())

# Cek korelasi terhadap target
corr = df.corr(numeric_only=True)
plt.figure(figsize=(12, 8))
sns.heatmap(corr[['SalePrice']].sort_values(by='SalePrice', ascending=False), annot=True)
plt.title('Korelasi terhadap SalePrice')
plt.show()

# Pilih beberapa fitur yang berkorelasi tinggi
features = ['OverallQual', 'GrLivArea', 'GarageCars', 'TotalsBsmtSF', 'FullBath', 'YearBuilt']
target = 'SalePrice'
```

Gambar 6 Cek struktur data dan korelasi

Hasil *heatmap* tersebut dapat dilihat pada Gambar 7.



Gambar 7 Heatmap korelasi terhadap saleprice

Berdasarkan hasil visualisasi, fitur-fitur seperti *OverallQual*, *GrLivArea*, *GarageCars*, dan *TotalBsmtSF* menunjukkan korelasi signifikan terhadap variabel target *SalePrice*. Oleh karena itu, fitur-fitur tersebut dipilih untuk dijadikan variabel prediktor dalam proses pemodelan.

4.3 Modify

Langkah ini bertujuan untuk membersihkan dan mempersiapkan data sebelum masuk ke tahap pemodelan. Setelah mengeksplorasi data, beberapa fitur numerik yang digunakan dalam pemodelan diketahui memiliki nilai kosong (*missing values*). Untuk memastikan hal ini, dilakukan pengecekan jumlah *missing value*. Setelah itu, nilai kosong pada fitur numerik diimputasi menggunakan nilai tengah (*median*) agar tidak bias terhadap *outlier* seperti pada Gambar 8.

```
# Cek missing value
print(df[features].isnull().sum())

# Imputasi numerik (median)
df[features] = df[features].fillna(df[features].median())
```

Gambar 8 Cek missing value dan imputasi numerik

Untuk fitur kategorikal yang memiliki pengaruh terhadap harga rumah dilakukan transformasi ke bentuk numerik dengan teknik *one-hot encoding*. Sebelumnya, fitur ini juga digabungkan ke dalam *subset* data yang kemudian ditransformasi. Setelah semua data bersih dan sudah ditransformasi, *dataset* dibagi menjadi dua bagian dengan *training set* sebanyak 80% dan *test set* sebanyak 20%. Hal tersebut dilakukan seperti pada Gambar 9.

```
# Tambahkan fitur kategorikal untuk encoding
categorical = ['Neighborhood']
df = df[features + categorical + [target]]

# One-hot encoding
df_encoded = pd.get_dummies(df, columns=categorical, drop_first=True)

# Pembagian dataset
X = df_encoded.drop('SalePrice', axis=1)
y = df_encoded['SalePrice']

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

Gambar 9 Encoding fitur kategorikal dan split dataset

Pembagian ini bertujuan untuk mengukur kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya.

4.4 Model

Pada tahap ini, model RFR dibangun menggunakan pustaka *scikit-learn*. Model awal diinisialisasi dengan parameter bawaan (*default parameters*), kemudian dilatih pada data pelatihan (*training set*) dan digunakan untuk melakukan prediksi terhadap data uji (*test set*). Untuk mengukur kinerja awal model digunakan dua metrik utama, yaitu RMSE dan R^2 . Hal tersebut dapat dilakukan seperti pada Gambar 10.

```
# Inisialisasi model dasar
model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

# Prediksi awal
y_pred = model.predict(X_test)

# Pengukuran Performa
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print(f"RMSE: {rmse:.2f}")
print(f"R²: {r2:.2f}")

RMSE: 28452.75
R²: 0.89
```

Gambar 10 Inisialisasi model, prediksi, dan pengukuran performa

Output memperlihatkan bahwa model yang digunakan menghasilkan nilai RMSE sebesar 28.452,75, sedangkan R^2 yang diperoleh mencapai 0,89. Selanjutnya, dilakukan penyetelan *hyperparameter* untuk meningkatkan akurasi model. Parameter yang diuji meliputi:

- n_estimators* (jumlah pohon)
- max_depth* (kedalaman maksimum pohon)
- max_features* (jumlah fitur yang dipertimbangkan pada tiap pemisahan)

Proses ini dilakukan dengan menggunakan *GridSearchCV* dan validasi silang (*cross-validation*) 3-fold. Potongan kode dan hasil *grid search* dapat dilihat pada Gambar 11.

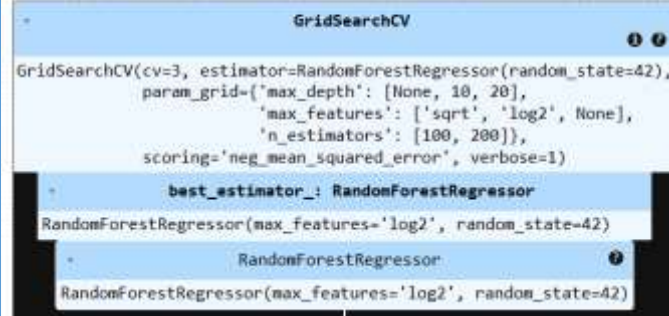
```
from sklearn.model_selection import GridSearchCV

param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'max_features': ['sqrt', 'log2', None]}

grid_search = GridSearchCV(
    RandomForestRegressor(random_state=42),
    param_grid,
    cv=3,
    scoring='neg_mean_squared_error',
    verbose=1)

grid_search.fit(X_train, y_train)

Fitting 3 folds for each of 18 candidates, totalling 54 fits
```



```
GridSearchCV(cv=3, estimator=RandomForestRegressor(random_state=42),
  param_grid={'max_depth': [None, 10, 20],
    'max_features': ['sqrt', 'log2', None],
    'n_estimators': [100, 200]},
  scoring='neg_mean_squared_error', verbose=1)

Best estimator: RandomForestRegressor
RandomForestRegressor(max_features='log2', random_state=42)
```

Gambar 11 Validasi silang dan hasil *grid search*

Dengan demikian, parameter siap digunakan untuk pengujian ulang model pada tahap berikutnya.

4.5 Assess

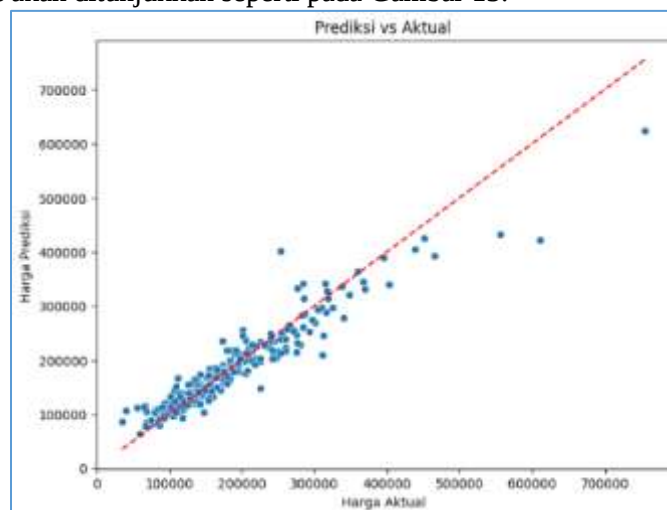
Model hasil penyetelan diukur kembali untuk melihat peningkatan performa. Prediksi dilakukan terhadap data uji menggunakan model terbaik. Perbandingan antara nilai aktual dan prediksi dapat divisualisasikan dalam bentuk *scatter plot* dengan perintah seperti pada Gambar 12.

```
# Prediksi data uji
y_best_pred = grid_search.best_estimator_.predict(X_test)

# Scatter plot prediksi vs aktual
plt.figure(figsize=(8, 6))
sns.scatterplot(x=y_test, y=y_best_pred)
plt.xlabel("Harga Aktual")
plt.ylabel("Harga Prediksi")
plt.title("Prediksi vs Aktual")
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], '--r')
plt.show()
```

Gambar 12 Prediksi data uji dan pembentukan *scatter plot*

Hasil *scatter plot* akan ditunjukkan seperti pada Gambar 13.



Gambar 13 Hasil *scatter plot*

Selain divisualisasikan dalam bentuk *scatter plot*, perbandingan tersebut juga dapat ditampilkan dalam bentuk tabel berisi sejumlah data uji menggunakan perintah seperti pada Gambar 14.

```
hasil_prediksi = pd.DataFrame({
    'Harga Aktual': y_test.values,
    'Harga Prediksi': y_best_pred
})

print(hasil_prediksi.head(10))
```

Gambar 14 Pembentukan tabel hasil

Hasil yang akan dikembalikan adalah 10 data pertama yang memperlihatkan nilai harga aktual beserta hasil prediksi seperti pada Tabel 1.

Tabel 1 Hasil prediksi dan harga aktual

Harga Aktual	Harga Prediksi
154500	140635.00
325000	297334.00
115000	110407.75

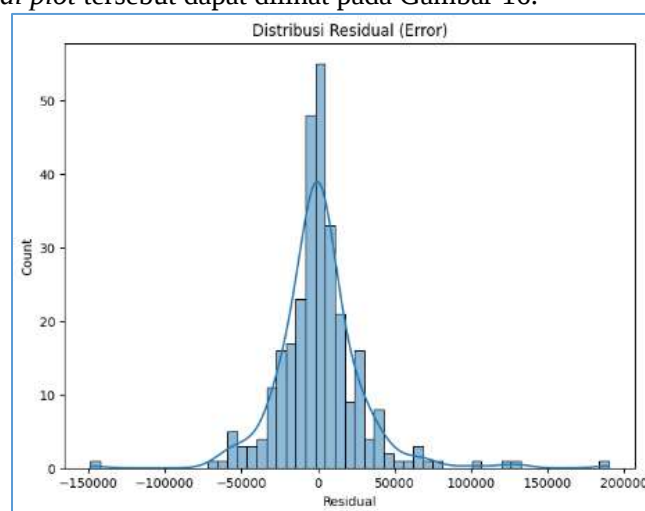
159000	165225.00
315500	288870.65
75500	83408.00
311500	210402.00
146000	149427.00
84500	83843.00
135500	119245.00

Selanjutnya dilakukan distribusi kesalahan prediksi (*residual*) untuk memperoleh pola *error* dengan perintah seperti pada Gambar 15.

```
residuals = y_test - y_best_pred
plt.figure(figsize=(8, 6))
sns.histplot(residuals, kde=True)
plt.title("Distribusi Residual (Error)")
plt.xlabel("Residual")
plt.show()
```

Gambar 15 Distribusi *residual*

Visualisasi *residual plot* tersebut dapat dilihat pada Gambar 16.



Gambar 16 Hasil *residual plot*

Mayoritas *residual* terkonsentrasi di sekitar nilai nol, yang mengindikasikan bahwa sebagian besar prediksi model cukup mendekati nilai aktual terutama untuk rumah dengan harga menengah. Sebaran *residual* yang relatif simetris ke arah positif dan negatif juga menunjukkan bahwa model tidak cenderung bias ke salah satu arah tertentu.

Temuan tersebut sejalan dengan hasil kinerja algoritma, di mana nilai RMSE sebesar 28.452,75, yang hanya sekitar 15,7% dari rata-rata harga rumah dan jauh lebih kecil dibandingkan standar deviasi target (79.442,50). Hal ini menunjukkan bahwa model memiliki tingkat presisi prediksi yang cukup baik [22].

Sementara itu, nilai R^2 sebesar 0,89 menunjukkan bahwa model mampu menjelaskan 89% atau sebagian besar variabilitas harga rumah, sejalan dengan penelitian terdahulu yang juga menggunakan R^2 sebagai metrik utama dalam evaluasi model [11]. Temuan tersebut juga diperkuat dengan membandingkan hasil penelitian lain. Pada studi Yağmur et al., model ANN memperoleh R^2 sebesar 0,827, SVR sebesar 0,749, dan MLR sebesar 0,721, sehingga kemampuan model berada pada kisaran 72-83% [11]. Angka tersebut masih lebih rendah dibandingkan hasil penelitian ini, yang menunjukkan bahwa RFR memiliki kemampuan prediksi yang lebih akurat.

Selain mengukur metrik *error*, penelitian ini juga menganalisis *feature importance* untuk mengetahui fitur yang paling berpengaruh dalam penentuan harga rumah. Analisis dilakukan melalui atribut *feature_importances* pada model terbaik hasil *GridSearchCV* seperti pada Gambar 17.

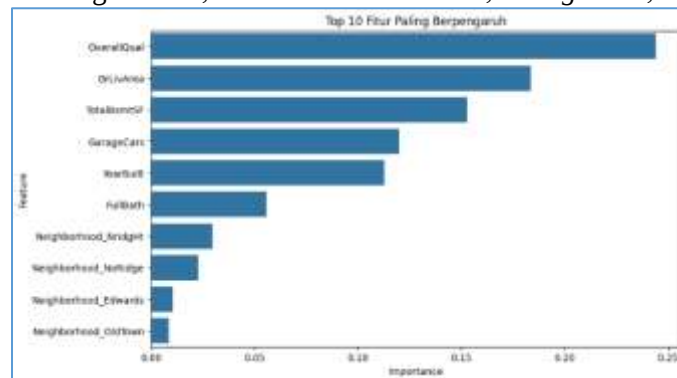
```
importances = grid_search.best_estimator_.feature_importances_
feature_names = X_train.columns

feat_imp = pd.DataFrame({
    'Feature': feature_names,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)

plt.figure(figsize=(10, 6))
sns.barplot(data=feat_imp.head(10), x='Importance', y='Feature')
plt.title('Top 10 Fitur Paling Berpengaruh')
plt.show()
```

Gambar 17 Analisis *feature importance*

Visualisasi pada Gambar 18 menunjukkan bahwa *OverallQual* merupakan faktor yang paling kuat dalam memengaruhi harga rumah, diikuti oleh *GrLivArea*, *GarageCars*, dan *TotalBsmtSF*.



Gambar 18 Hasil analisis *feature importance*

Untuk melengkapi hasil tersebut, dilakukan evaluasi *robustness* melalui validasi silang dari *GridSearchCV* dengan mengonversi skor *neg_mean_squared_error* menjadi *Mean Squared Error* (MSE) dan RMSE. Kode pada Gambar 19 digunakan untuk memperoleh rerata performa dari seluruh lipatan validasi.

```
cv_results = grid_search.cv_results_

# mean_test_score adalah negatif MSE
mean_mse = -cv_results['mean_test_score'].mean()
std_mse = cv_results['std_test_score'].mean()

rmse_cv_mean = mean_mse ** 0.5
rmse_cv_std = std_mse ** 0.5

print("MSE rata-rata CV      :", mean_mse)
print("Standar deviasi MSE   :", std_mse)
print("RMSE rata-rata CV     :", rmse_cv_mean)
print("Standar deviasi RMSE  :", rmse_cv_std)

MSE rata-rata CV      : 940366787.7905501
Standar deviasi MSE   : 189226716.15930164
RMSE rata-rata CV     : 30665.40049943177
Standar deviasi RMSE  : 13755.97020058206
```

Gambar 19 Evaluasi *robustness* dan hasil

Hasil validasi silang menunjukkan rata-rata MSE sebesar 940.366.787,79 atau sama dengan nilai RMSE sekitar 30.665,40. Hasil ini hanya sedikit lebih tinggi dibandingkan RMSE data uji yaitu 28.452,75. Standar deviasi RMSE yang menunjukkan nilai 13.755,97 juga menjelaskan bahwa variasi performa antar lipatan masih wajar. Meskipun temuan ini menunjukkan bahwa RFR stabil dan tidak terdapat tanda-tanda *overfitting*, masih ada beberapa keterbatasan yang perlu diperhatikan.

Keterbatasan penelitian ini mencakup penggunaan *dataset* yang berlatar di Amerika Serikat sehingga kurang sesuai dengan konteks Indonesia. Fitur yang digunakan juga hanya mencakup aspek fisik rumah tanpa faktor eksternal serta performa yang diukur terbatas pada satu algoritma sehingga tidak memungkinkan perbandingan metode. Namun, penelitian ini tetap memberikan kontribusi-kontribusi yang berarti.

Kontribusi ini terlihat dari penerapan kerangka SEMMA secara sistematis dan bukti bahwa RFR mampu memberikan performa prediksi yang cukup baik berdasarkan RMSE dan R^2 . Selain itu, kualitas bangunan dan luas ruang juga ditegaskan dalam analisis *feature importance* sebagai faktor utama, sehingga model ini berpotensi digunakan dalam sistem penilaian harga otomatis.

Untuk pengembangan ke depan, penelitian selanjutnya dapat menggunakan *dataset* lokal, menambah fitur spasial dan ekonomi, membandingkan algoritma lain seperti *XGBoost* atau *LightGBM*, serta menerapkan pendekatan *explainable AI* untuk meningkatkan transparansi model dalam pengambilan keputusan.

5 Kesimpulan

Penelitian ini menunjukkan bahwa RFR dapat memprediksi harga rumah dengan cukup baik menggunakan *dataset House Prices* dari Kaggle. Dengan alur kerja SEMMA, proses dilakukan secara runtut mulai dari pengambilan sampel hingga evaluasi akhir. Model menghasilkan performa yang kuat, dengan RMSE 28.452,75 dan R^2 0,89, serta validasi silang yang stabil tanpa tanda *overfitting*. Hasil *feature importance* menegaskan bahwa kualitas bangunan (*OverallQual*) dan luas ruang tinggal (*GrLivArea*, *GarageCars*, *TotalBsmntSF*) adalah faktor yang paling berpengaruh terhadap harga. Meskipun model bekerja dengan baik, penelitian ini terbatas pada data Amerika Serikat dan fitur fisik, serta belum membandingkan algoritma lain. Secara keseluruhan, RFR terbukti sebagai metode yang andal dan berpotensi dikembangkan untuk sistem estimasi harga rumah yang lebih praktis dan berbasis data lokal.

Referensi

- [1] T. M. Mitchell, *Machine Learning*. in McGraw-Hill international editions - computer science series. McGraw-Hill Education, 1997. [Online]. Available: <https://books.google.co.id/books?id=xOGAngEACAAJ>
- [2] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press, 2015.
- [3] E. Siegel, "Praise for Predictive Analytics," Wiley, uSA, 2015, [Online]. Available: https://www.predictiveanalyticsworld.com/book/pdf/Predictive_Analytics_by_Eric_Siegel_Excerpts.pdf
- [4] C.-H. Hung and S.-W. Tzang, "Consumption and Investment Values in Housing Price: A Real Options Approach," *Int. J. Strateg. Prop. Manag.*, Vol. 25, No. 4, pp. 278–290, 2021, DOI: 10.3846/ijspm.2021.14914.
- [5] W. Zhang and T. A. Masron, "Impact of Financial Market Development on Housing Prices: Evidence from China," *Int. J. Strateg. Prop. Manag.*, Vol. 29, No. 2, pp. 114–127, 2025, DOI: 10.3846/ijspm.2025.24036.
- [6] A. G. Ulucan, A. Bozdağ, M. Karakoyun, and T. Alkan, "Forecasting Pandemic-Induced Changes in Real Estate Market Values Through Machine Learning Approaches," *Int. J. Strateg. Prop. Manag.*, Vol. 29, No. 3, pp. 196–214, 2025, DOI: 10.3846/ijspm.2025.24063.
- [7] F. Pratami, T. Marwa, S. Andaiyani, and A. Abukosim, "What Factors Can Affect Indonesian Property Price?," *TRIKONOMIKA*, Vol. 23, No. 1 SE-Articles, pp. 49–54, Jun. 2023, DOI: 10.23969/trikononika.v23i1.7730.
- [8] M. M. Mutiara, Y. S. J. Nasution, and A. Syakir, "Factors Affecting Customer Interest in KPR" <http://sistemasi.ftik.unisi.ac.id>

- iB Griya Housing Financing at PT Bank Sumut, Sibolga Syariah Branch,” *Amkop Manag. Account. Rev.*, Vol. 5, No. 1 SE-Articles, pp. 575–588, Jun. 2025, DOI: 10.37531/amar.v5i1.2669.
- [9] R. Naz, B. Jamil, and H. Ijaz, “Machine Learning, Deep Learning, and Hybrid Approaches in Real Estate Price Prediction: A Comprehensive Systematic Literature Review,” *Proc. Pakistan Acad. Sci. A. Phys. Comput. Sci.*, Vol. 61, No. 2 SE-Review Articles, pp. 129–144, Jun. 2024, DOI: 10.53560/PPASA(61-2)863.
- [10] P. Abhishek and R. B. Sankar, “House Price Prediction App,” *Int. SCI. J. Eng. Manag.*, Vol. 04, No. 07, pp. 1–9, 2025, DOI: 10.55041/isjem04918.
- [11] A. Yağmur, M. Kayakuş, and M. Terzioğlu, “House Price Prediction Modeling using Machine Learning Techniques: A Comparative Study,” *Aestimum*, Vol. 81, pp. 39–51, 2022, DOI: 10.36253/aestim-13703.
- [12] Y. Zhao, J. Zhao, and E. Y. Lam, “House Price Prediction: A Multi-Source Data Fusion Perspective,” *Big Data Min. Anal.*, Vol. 7, No. 3, pp. 603–620, 2024, DOI: 10.26599/BDMA.2024.9020019.
- [13] C. Zhan, Y. Liu, Z. Wu, M. Zhao, and T. W. S. Chow, “A Hybrid Machine Learning Framework for Forecasting House Price,” *Expert Syst. Appl.*, Vol. 233, p. 120981, 2023, DOI: <https://doi.org/10.1016/j.eswa.2023.120981>.
- [14] Bank for International Settlement, “Statistical release : BIS residential property price statistics in Q1 2021,” no. August, pp. 1–8, 2022.
- [15] L. Breiman, “Random Forests,” *Mach. Learn.*, Vol. 45, No. 1, pp. 5–32, 2001, DOI: 10.1023/A:1010933404324.
- [16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. in Springer Series in Statistics. Springer New York, 2009. [Online]. Available: <https://books.google.co.id/books?id=a3NazwEACAAJ>
- [17] I. Moreno-Foronda, M.-T. Sánchez-Martínez, and M. Pareja-Eastaway, “Comparative Analysis of Advanced Models for Predicting Housing Prices: A Review,” 2025. DOI: 10.3390/urbansci9020032.
- [18] T. E. A. F. Elmuna, T. Chamidy, & F. Nugroho, “Optimization of the Random Forest Method Using Principal Component Analysis to Predict House Prices: A Case Study of House Prices in Malang City.” doi: 10.25008/ijadis.v4i2.1290.
- [19] A. Montoya and DataCanary, “House Prices - Advanced Regression Techniques,” 2016.
- [20] A. Wisnuwardhana, A. Nizar Hidayanto, N. Fitriah Ayuning Budi, I. Chandra Hapsari, Denny, and A. Haidaroh, “Systematic Literature Review: Critical Success Factor in the Application of Data Mining,” *J. Phys. Conf. Ser.*, Vol. 1444, No. 1, p. 12023, 2020, DOI: 10.1088/1742-6596/1444/1/012023.
- [21] SAS Institute Inc, *Data mining using SAS® Enterprise Miner: A Case Study Approach (Fourth Edition)*. 2018.
- [22] C. J. Willmott and K. Matsuura, “Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance,” *Clim. Res.*, Vol. 30, pp. 79–82, 2005, [Online]. Available: <https://www.int-res.com/abstracts/cr/v30/cr030079>