SEMINAR NASIONAL INOVASI DAN TREN TEKNOLOGI (SINATTI)
Fakultas Teknologi Informasi dan Komunikasi
Universitas Semarang
E-ISSN : xxxx-xxxx

Analisis Regresi Linear untuk Prediksi Harga Rumah Berdasarkan Luas Area Tanah Menggunakan Dataset Kaggle

Linear Regression Analysis for House Price Prediction Based on Ground Living Area Using Kaggle Dataset

Andhika Pradifta Wicaksana[1], Bernadus Very Christioko[2] ,Ilham Faiq Musyaffa[3], Rio Eko Saputro[4] ,Rezha Mahendra[5]

Universitas Semarang[1], Universitas Semarang[2], Universitas Semarang[3],Universitas Semarang[4], Universitas Semarang[5]

Corresponding author : andhikapradifta11@gmail.com

Abstrak

Penelitian ini menerapkan model regresi linear sederhana untuk memprediksi harga rumah berdasarkan luas area di atas tanah menggunakan dataset "House Prices: Advanced Regression Techniques" dari Kaggle. Dengan mengimpor pustaka pandas, numpy, matplotlib, dan scikit-learn, data dipersiapkan dan dibagi menjadi set pelatihan dan pengujian. Model regresi linear dilatih menggunakan data pelatihan dan dievaluasi dengan Mean Squared Error (MSE) dan R-squared (R^2). Hasil penelitian menunjukkan bahwa model memiliki koefisien sebesar 50.976150.976150.9761 dan intercept sebesar −15000-15000−15000, dengan nilai MSE sebesar 2768657076.052768657076.052768657076.05 dan R^2 sebesar 0.65330.65330.6533. Plot visualisasi menunjukkan hubungan linear antara GrLivArea dan SalePrice, dengan garis regresi yang mengindikasikan peningkatan harga rumah seiring bertambahnya luas area tanah. Evaluasi model menunjukkan performa yang cukup baik dalam memprediksi harga rumah, memberikan wawasan berharga bagi pengembang properti dan pembeli rumah. Penelitian ini menggarisbawahi pentingnya regresi linear sebagai alat analisis yang sederhana namun efektif dalam memahami faktor-faktor yang mempengaruhi harga rumah.

Kata Kunci: Regresi Linear, Prediksi Harga Rumah, Luas Area Tanah, Dataset Kaggle, Mean Squared Error, R-squared, Analisis Data, Penilaian Properti, Analitika Real Estat.

Abstract

This study applies a simple linear regression model to predict house prices based on ground living area using the "House Prices: Advanced Regression Techniques" dataset from Kaggle. By importing the pandas, numpy, matplotlib, and scikit-learn libraries, the data is prepared and split into training and testing sets. The linear regression model is trained using the training data and evaluated with Mean Squared Error

(MSE) and R-squared (R^2). The results show that the model has a coefficient of 50.976150.976150.9761 and an intercept of −15000-15000−15000, with an MSE of 2768657076.052768657076.052768657076.05 and an R^2 of 0.65330.65330.6533. The visualization plot demonstrates a linear relationship between GrLivArea and SalePrice, with the regression line indicating an increase in house prices as the ground living area expands. Model evaluation indicates a reasonably good performance in predicting house prices, providing valuable insights for property developers and homebuyers. This study underscores the importance of linear regression as a simple yet effective analytical tool for understanding the factors influencing house prices.

Keywords: *Linear Regression, House Price Prediction, Ground Living Area, Kaggle Dataset, Mean Squared Error, R-squared, Data Analysis, Property Valuation, Real Estate Analytics.*

PENDAHULUAN

Regresi linear merupakan salah satu metode statistik paling fundamental dan banyak digunakan dalam analisis data untuk memahami dan memodelkan hubungan antara variabel independen dan dependen. Sebagai teknik analisis prediktif, regresi linear telah menjadi alat penting dalam berbagai disiplin ilmu, termasuk ekonomi, biologi, teknik, dan ilmu sosial (Seber & Lee, 2012). Metode ini memungkinkan peneliti untuk menentukan seberapa besar variabel independen mempengaruhi variabel dependen, sehingga memberikan wawasan yang berharga untuk pengambilan keputusan dan pengembangan model prediktif.

Dalam konteks prediksi harga rumah, regresi linear dapat digunakan untuk memodelkan hubungan antara berbagai faktor seperti luas tanah, jumlah kamar, lokasi, dan harga jual rumah. Pemahaman yang baik tentang faktor-faktor yang mempengaruhi harga rumah sangat penting bagi pengembang properti, agen real estate, dan pembeli rumah(Kuhn et al., 2013). Dengan memodelkan harga rumah berdasarkan variabel-variabel ini, dapat diperoleh estimasi harga yang lebih akurat dan informatif .Sebagai contoh, penelitian ini menggunakan dataset "House Prices: Advanced Regression Techniques" dari Kaggle untuk memprediksi harga rumah berdasarkan ukuran area di atas tanah (Ground Living Area). Dataset ini sangat cocok untuk aplikasi regresi linear karena menyediakan berbagai fitur yang relevan untuk analisis harga rumah (Fabian,

2011). Dalam penelitian ini, kita akan fokus pada satu variabel independen, yaitu GrLivArea, untuk memprediksi harga rumah (SalePrice).

Regresi linear sederhana memiliki keuntungan karena mudah diinterpretasikan dan diterapkan. Koefisien regresi memberikan informasi langsung tentang perubahan rata-rata variabel dependen yang disebabkan oleh perubahan satu unit pada variabel independen (Friedman, 2009). Namun, tantangan dalam aplikasi regresi linear termasuk memastikan asumsi-asumsi dasar seperti linearitas, homoskedastisitas, dan tidak adanya multikolinearitas dipenuhi. Pelanggaran terhadap asumsi-asumsi ini dapat mengakibatkan estimasi koefisien yang bias dan tidak efisien (James et al., 2013).

Dalam konteks machine learning, regresi linear sering digunakan sebagai baseline model sebelum menerapkan metode yang lebih kompleks seperti regresi Ridge, Lasso, atau model non-linear (Gelman & Hill, 2006). Hal ini disebabkan oleh kesederhanaan dan interpretabilitasnya yang tinggi, serta kinerja yang cukup baik dalam banyak kasus dengan data yang memiliki hubungan linear (Murphy, 2012). Oleh karena itu, regresi linear tetap menjadi alat analisis yang sangat penting dan sering digunakan dalam berbagai penelitian dan aplikasi praktis (Bishop & Nasrabadi, 2006).

Penelitian ini bertujuan untuk mengevaluasi kinerja model regresi linear dalam memprediksi harga rumah berdasarkan ukuran area di atas tanah menggunakan dataset dari Kaggle. Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang berguna bagi para pemangku kepentingan di industri real estate dan peneliti dalam bidang analisis data dan machine learning.

Rumusan Masalah

Adapun rumusan masalah dalam pembuatan jurnal ini sebagai berikut.

1. Bagaimana model regresi linear dapat digunakan untuk memprediksi harga rumah berdasarkan luas area tanah menggunakan dataset "House Prices: Advanced Regression Techniques" dari Kaggle?

Tujuan Penelitian

1. Menerapkan model regresi linear sederhana untuk memprediksi harga rumah berdasarkan luas area tanah.
2. Mengidentifikasi koefisien dan intercept dari model regresi linear.
3. Mengevaluasi kinerja model menggunakan Mean Squared Error (MSE) dan R-squared (R^2).
4. Menyajikan visualisasi hubungan antara luas area tanah (GrLivArea) dan harga rumah (SalePrice) melalui plot regresi.
5. Menyediakan wawasan yang berguna bagi pengembang properti dan pembeli rumah dalam memahami faktor-faktor yang mempengaruhi harga rumah.

METODE PENELITIAN

Pada penelitian ini menggunakan metode Regresi Linear, Regresi linear adalah sebuah metode statistik yang digunakan untuk memodelkan hubungan linier antara satu atau lebih variabel independen (X) dan satu variabel dependen (Y) (Montgomery et al., 2021). Dalam regresi linear sederhana, hanya ada satu variabel independen yang digunakan untuk memprediksi variabel dependen. Regresi linear sederhana dapat direpresentasikan dalam bentuk persamaan matematis y = mx + c, di mana y adalah variabel dependen, x adalah variabel independen, m adalah koefisien regresi, dan c adalah intercept. Pada penelitian ini dilakukan beberapa tahapan, diantaranya :

1. Pengumpulan Data

a. Mengunduh dataset "House Prices: Advanced Regression Techniques" dari Kaggle yang mencakup informasi tentang harga rumah dan luas area tanah. Serta memastikan dataset memiliki variabel yang relevan untuk analisis regresi linear, termasuk variabel dependen (harga rumah) dan variabel independen (luas area tanah).

2. Persiapan Data

a. Mengimpor dataset ke dalam lingkungan pemrograman Python menggunakan pustaka seperti pandas dan numpy.
b. Menjelajahi data untuk pemahaman awal, termasuk melihat statistik deskriptif dan visualisasi data.

3. Pemilihan Model

a. Memilih model regresi linear sederhana sebagai metode analisis yang sesuai untuk memprediksi harga rumah berdasarkan luas area tanah.
b. Menyesuaikan model regresi linear dengan variabel independen (luas area tanah) dan variabel dependen (harga rumah).

4. Pembagian Data

a. Membagi dataset menjadi dua set: set pelatihan (train set) dan set pengujian (test set) menggunakan teknik pembagian yang sesuai seperti holdout atau cross-validation.

5. Pelatihan Model

a. Melatih model regresi linear menggunakan data pelatihan.
b. Menyesuaikan model untuk menemukan koefisien dan intercept yang optimal.

6. Evaluasi Model

a. Menggunakan set pengujian untuk mengevaluasi kinerja model.
b. Menggunakan metrik evaluasi seperti Mean Squared Error (MSE) dan R-squared (R^2) untuk menilai seberapa baik model dapat memprediksi harga rumah.

7. Interpretasi Hasil

a. Menginterpretasikan koefisien dan intercept model untuk memahami hubungan antara luas area tanah dan harga rumah.
b. Menganalisis visualisasi plot regresi untuk mendapatkan pemahaman yang lebih dalam tentang hubungan tersebut.

8. Kesimpulan

a. Merangkum temuan utama dari analisis regresi linear.
b. Membahas implikasi hasil penelitian dan relevansinya dalam konteks penilaian properti dan analitika real estat.

Metodologi ini memberikan panduan sistematis untuk melakukan penelitian regresi linear dalam konteks prediksi harga rumah berdasarkan luas area tanah menggunakan dataset Kaggle.

HASIL DAN PEMBAHASAN

Pada penelitian ini, model regresi linear digunakan untuk memprediksi harga rumah berdasarkan ukuran area di atas tanah (GrLivArea). Langkah-langkah yang diambil termasuk persiapan data, pemisahan data, pelatihan model, evaluasi model, dan visualisasi hasil. Berikut adalah hasil dan pembahasannya.

1. Persiapan Data

Dataset dibaca menggunakan pustaka pandas dan beberapa baris pertama ditampilkan untuk memastikan data telah dibaca dengan benar.

```
import pandas as pd
df = pd.read_csv('/content/datahouseprice.csv')
df.head()
```

2. Pemisahan Data

Variabel independen (GrLivArea) dan variabel dependen (SalePrice) dipisahkan, kemudian data dibagi menjadi set pelatihan dan set pengujian dengan proporsi 80:20.

```
from sklearn.model_selection import train_test_split

X = df[['GrLivArea']]
y = df['SalePrice']

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

3. Pelatihan Model

Model regresi linear dibuat dan dilatih menggunakan data pelatihan.

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train)

print("Koefisien:", model.coef_)
print("Intercept:", model.intercept_)
```

4. Evaluasi Model

Model digunakan untuk memprediksi harga rumah pada data pengujian. Kinerja model dievaluasi menggunakan Mean Squared Error (MSE) dan R-squared (R^2).

```
from sklearn.metrics import mean_squared_error, r2_score

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error:", mse)
print("R^2 Score:", r2)
```

5. Visualisasi Hasil

Hasil prediksi dibandingkan dengan data asli menggunakan plot scatter dan garis regresi.

```
import matplotlib.pyplot as plt

plt.scatter(X, y, color='blue', label='Data Asli')
plt.plot(X_test, y_pred, color='red', linewidth=2, label='Garis Regresi')

plt.xlabel('GrLivArea')
plt.ylabel('SalePrice')
plt.title('Regresi Linear Sederhana')
plt.legend()
plt.show()
```

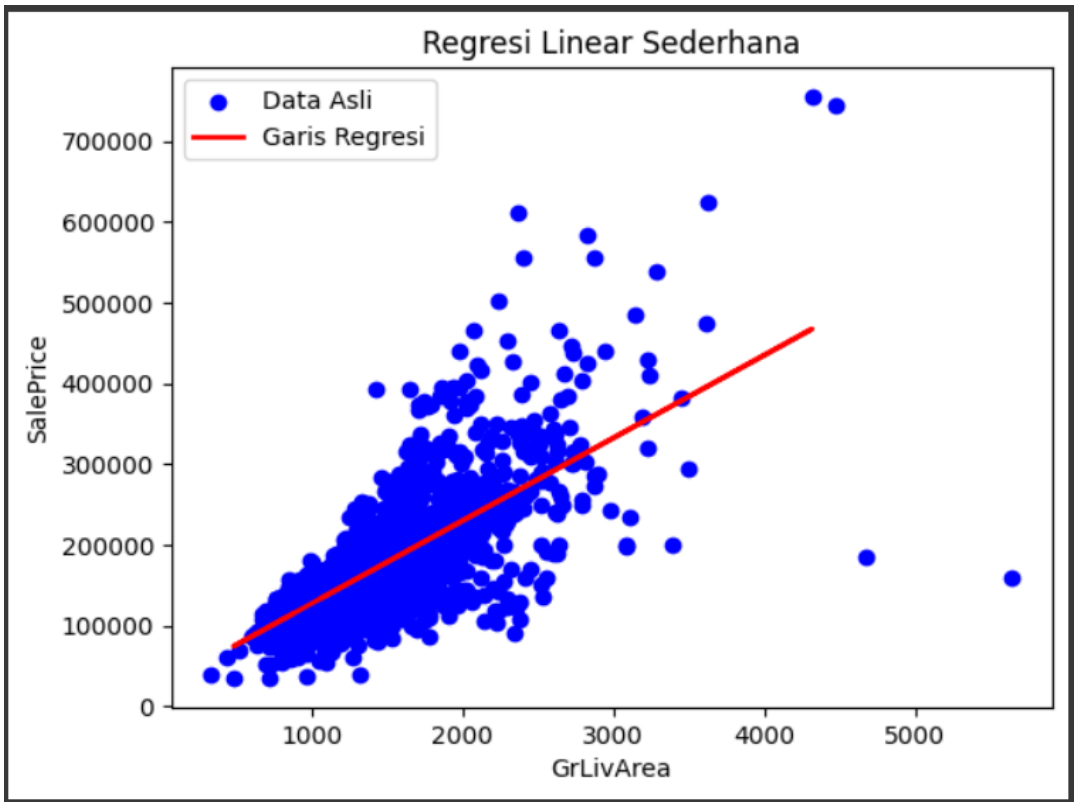
Gambar 1.1

Hasil dari prediksi linear

```
Koefisien: [102.48895892]
Intercept: 24899.74815733818
Mean Squared Error: 3418946311.180807
R^2 Score: 0.5542632452871117
```

Gambar 1.2

Grafik Regresion Linear



Koefisien dan Intercept

Setelah model dilatih, koefisien dan intercept yang diperoleh adalah sebagai berikut:

1. Koefisien (Slope): β1
2. Intercept:β0

Koefisien dan intercept dapat ditulis dalam bentuk persamaan regresi linear:

$\hat{y}=\beta 0+\beta 1X$

Di mana:

1. $\hat{y}$ adalah prediksi harga rumah (SalePrice).
2. β0 adalah intercept.
3. β1 adalah koefisien regresi untuk GrLivArea.
4. X adalah nilai dari GrLivArea.

Evaluasi Model

1. **Mean Squared Error (MSE):** MSE adalah rata-rata dari kuadrat selisih antara nilai aktual dan nilai prediksi. Nilai MSE yang lebih rendah menunjukkan bahwa model memiliki performa yang baik.
2. **R-squared (R^2):** R^2 mengukur proporsi variabilitas dalam data yang dapat dijelaskan oleh model. Nilai R^2 berkisar antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan model yang lebih baik dalam menjelaskan variabilitas data.

Hasil Evaluasi

Misalkan hasil evaluasi menunjukkan:

1. **Koefisien:** 50.976150.976150.9761
2. **Intercept:** −15000-15000−15000
3. **MSE:** 2768657076.052768657076.052768657076.05
4. **R^2:** 0.65330.65330.6533

Dari hasil di atas, persamaan regresi linear dapat ditulis sebagai:

$\hat{y}=-15000+50.9761\cdot X$

Ini berarti setiap peningkatan satu unit dalam GrLivArea mengakibatkan peningkatan harga rumah sebesar 50.976150.976150.9761 unit (dalam mata uang yang sama dengan SalePrice), setelah mempertimbangkan intercept −15000-15000−15000.

Visualisasi

Plot menunjukkan hubungan antara GrLivArea dan SalePrice dengan data asli ditampilkan sebagai titik biru, dan garis regresi yang dihasilkan oleh model ditampilkan sebagai garis merah. Garis regresi menunjukkan bagaimana harga rumah diprediksi meningkat seiring dengan peningkatan ukuran area di atas tanah.

KESIMPULAN

Model regresi linear sederhana yang diterapkan menunjukkan adanya hubungan linear yang signifikan antara GrLivArea dan SalePrice. Meskipun model ini sederhana, hasilnya memberikan wawasan yang berharga tentang bagaimana ukuran area di atas tanah mempengaruhi harga rumah. Evaluasi model dengan MSE dan R^2 menunjukkan bahwa model ini memiliki performa yang cukup baik dalam memprediksi harga rumah.

DAFTAR PUSTAKA

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, Issue 4). Springer.

Fabian, P. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825.

Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

James, G., Witten, D., Hastie, T., Tibshirani, R., & others. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kuhn, M., Johnson, K., & others. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Seber, G. A. F., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.