Nama/NIM: Rafli Limandijaya/1103210243

1. Feature Engineering, EDA, dan Data Visualization

-EDA (Exploratory Data Analysis):

- Visualisasi distribusi semua fitur numerik menggunakan histogram dan KDE plot.
- Visualisasi korelasi antar fitur numerik dengan heatmap.
- Visualisasi fitur kategorikal menggunakan countplot (baik yang punya sedikit maupun banyak kategori).

-Feature Engineering:

- Penanganan missing values:
- Fitur numerik diisi dengan median.
- Fitur kategorikal diisi dengan modus (nilai terbanya

- Encoding kategorikal: Label encoding menggunakan .astype('category').cat.codes.

-Penambahan fitur baru:

- Fitur TotalArea = TotalBsmtSF + 1stFlrSF + 2ndFlrSF.

-Transformasi log untuk fitur yang skewed:

- Dilakukan log1p (np.log1p) pada fitur dengan skewness > 0.75 untuk mengurangi efek outlier.

2. **Matriks Evaluasi**

Model dievaluasi dengan tiga metrik utama:

- **RMSE (Root Mean Squared Error)**:

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

Mengukur rata-rata kesalahan prediksi. Nilai lebih kecil = prediksi lebih akurat.

- **MSE (Mean Squared Error)**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Dipakai secara tidak langsung karena cross_val_score(..., scoring='neg_mean_squared_error')

- **R^2 Score (Koefisien Determinasi)**:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Menunjukkan seberapa baik model menjelaskan variasi data. Nilai 1 = sempurna.

3. Perbandingan model

- Saya menggunakan model linier regression dan random forest regression. Pengecekan overfitting dilakukan dengan membandingkan train RMSE dan validation RMSE. Didapatkan bahwa rasio Train/Val < 1 → artinya tidak overfit parah. Random Forest cenderung lebih akurat, tapi overfitting sedikit karena RMSE training-nya terlalu rendah.

- Training RMSE: 30504.67

-R^2: 0.8525

-Baseline RMSE (mean): 79415.29
Model saya jauh lebih baik daripada baseline (mean prediksi), artinya model ini sudah menangkap pola yang bermanfaat.

-Linear Regression:

- CV RMSE: 34968
- Train RMSE: 29065
- Val RMSE: 37446
- Overfitting Ratio: 0.776
  Cukup bagus, tapi masih bisa ditingkatkan.

-Random Forest:

- CV RMSE: 30084
- Train RMSE: 11019 (sangat kecil)
- Val RMSE: 28379
- Overfitting Ratio: 0.388
  Model Random Forest punya performa validasi yang jauh lebih baik, tapi sedikit overfitting karena train RMSE-nya terlalu rendah dibanding validasi.

  R^2 Score Random Forest: 0.8950 → Lebih tinggi dari Linear Regression, artinya lebih akurat secara umum.

4.