Nama : Rizki Ramadhan

NIM : 1103213091

Regression Model

Dataset :

Nama

RegresiUTSTelkom.csv

1. Importing Library

Import Library

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import PolynomialFeatures, StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import xgboost as xgb
```

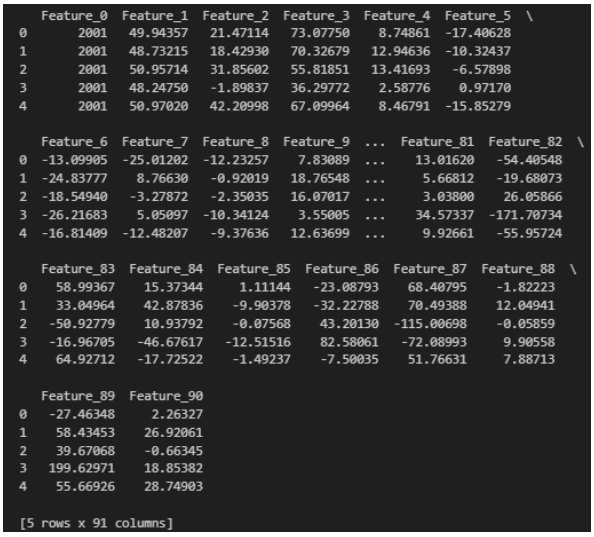
✓ 0.0s

2. Membaca File CSV ke dalam dataframe

```
     0        1        2        3        4        5        6        7   \
0  2001  49.94357  21.47114  73.07750   8.74861 -17.40628 -13.09905 -25.01202
1  2001  48.73215  18.42930  70.32679  12.94636 -10.32437 -24.83777   8.76630
2  2001  50.95714  31.85602  55.81851  13.41693  -6.57898 -18.54940  -3.27872
3  2001  48.24750  -1.89837  36.29772   2.58776   0.97170 -26.21683   5.05097
4  2001  50.97020  42.20998  67.09964   8.46791 -15.85279 -16.81409 -12.48207

         8         9   ...        81         82        83        84        85  \
0 -12.23257   7.83089  ...  13.01620  -54.40548  58.99367  15.37344   1.11144
1  -0.92019  18.76548  ...   5.66812  -19.68073  33.04964  42.87836  -9.90378
2  -2.35035  16.07017  ...   3.03800   26.05866 -50.92779  10.93792  -0.07568
3 -10.34124   3.55005  ...  34.57337 -171.70734 -16.96705 -46.67617 -12.51516
4  -9.37636  12.63699  ...   9.92661  -55.95724  64.92712 -17.72522  -1.49237

         86         87        88         89        90
0 -23.08793   68.40795  -1.82223  -27.46348   2.26327
1 -32.22788   70.49388  12.04941   58.43453  26.92061
2  43.20130 -115.00698  -0.05859   39.67068  -0.66345
3  82.58061  -72.08993   9.90558  199.62971  18.85382
4  -7.50035   51.76631   7.88713   55.66926  28.74903

[5 rows x 91 columns]
```

Dataset ini terdiri dari 91 kolom, termasuk satu kolom penanda waktu atau kategori di kolom pertama, dan sisanya merupakan fitur numerik. Nilai-nilai pada kolom menunjukkan variasi yang signifikan, mencakup rentang angka positif dan negatif dengan skala yang cukup besar (misalnya, kolom ke-81 berkisar dari 13.01620 hingga -171.70734). Ini mengindikasikan adanya fitur dengan pengaruh berbeda, baik positif maupun negatif, yang dapat memengaruhi target atau respon dari analisis lebih lanjut.

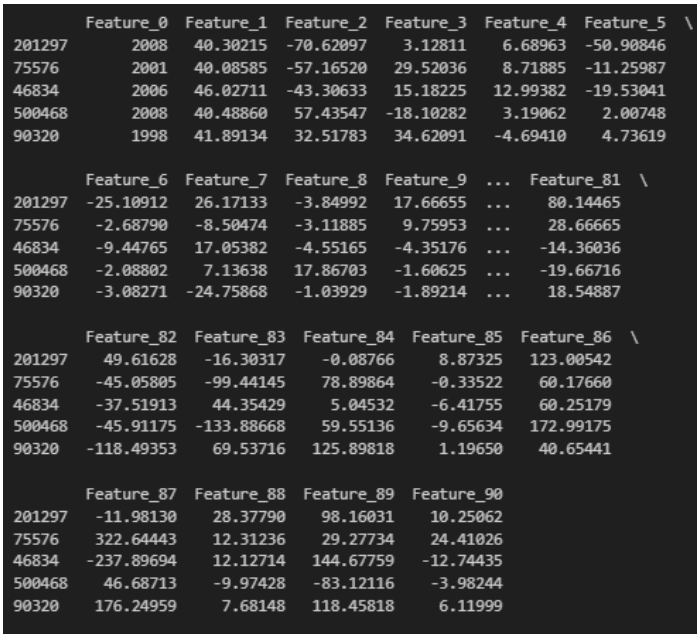
3. Penambahan header sementara



```
   Feature_0  Feature_1  Feature_2  Feature_3  Feature_4  Feature_5  \
0       2001   49.94357   21.47114   73.07750    8.74861  -17.40628
1       2001   48.73215   18.42930   70.32679   12.94636  -10.32437
2       2001   50.95714   31.85602   55.81851   13.41693   -6.57898
3       2001   48.24750   -1.89837   36.29772    2.58776    0.97170
4       2001   50.97020   42.20998   67.09964    8.46791  -15.85279

   Feature_6  Feature_7  Feature_8  Feature_9  ...  Feature_81  Feature_82  \
0  -13.09905  -25.01202  -12.23257    7.83089  ...    13.01620   -54.40548
1  -24.83777    8.76630   -0.92019   18.76548  ...     5.66812   -19.68073
2  -18.54940   -3.27872   -2.35035   16.07017  ...     3.03800    26.05866
3  -26.21683    5.05097  -10.34124    3.55005  ...    34.57337  -171.70734
4  -16.81409  -12.48207   -9.37636   12.63699  ...     9.92661   -55.95724

   Feature_83  Feature_84  Feature_85  Feature_86  Feature_87  Feature_88  \
0    58.99367    15.37344     1.11144   -23.08793    68.40795    -1.82223
1    33.04964    42.87836    -9.90378   -32.22788    70.49388    12.04941
2   -50.92779    10.93792    -0.07568    43.20130  -115.00698    -0.05859
3   -16.96705   -46.67617   -12.51516    82.58061   -72.08993     9.90558
4    64.92712   -17.72522    -1.49237    -7.50035    51.76631     7.88713

   Feature_89  Feature_90
0   -27.46348     2.26327
1    58.43453    26.92061
2    39.67068    -0.66345
3   199.62971    18.85382
4    55.66926    28.74903

[5 rows x 91 columns]
```

Dikarenakan dataset hanya berisis angka saja tanpa adanya header kita bisa menambahkan header sementara pada DataFrame df untuk mempermudah akses ke kolom-kolom dalam dataset, ini mempermudah manipulasi data selanjutnya, misalnya ketika mengakses kolom dengan nama yang lebih mudah dipahami

4. Data sampling



```
        Feature_0  Feature_1  Feature_2  Feature_3  Feature_4  Feature_5  \
201297       2008   40.30215  -70.62097    3.12811    6.68963  -50.90846
75576        2001   40.08585  -57.16520   29.52036    8.71885  -11.25987
46834        2006   46.02711  -43.30633   15.18225   12.99382  -19.53041
500468       2008   40.48860   57.43547  -18.10282    3.19062    2.00748
90320        1998   41.89134   32.51783   34.62091   -4.69410    4.73619

        Feature_6  Feature_7  Feature_8  Feature_9  ...  Feature_81  \
201297  -25.10912   26.17133   -3.84992   17.66655  ...    80.14465
75576    -2.68790   -8.50474   -3.11885    9.75953  ...    28.66665
46834    -9.44765   17.05382   -4.55165   -4.35176  ...   -14.36036
500468   -2.08802    7.13638   17.86703   -1.60625  ...   -19.66716
90320    -3.08271  -24.75868   -1.03929   -1.89214  ...    18.54887

        Feature_82  Feature_83  Feature_84  Feature_85  Feature_86  \
201297    49.61628   -16.30317    -0.08766     8.87325   123.00542
75576    -45.05805   -99.44145    78.89864    -0.33522    60.17660
46834    -37.51913    44.35429     5.04532    -6.41755    60.25179
500468   -45.91175  -133.88668    59.55136    -9.65634   172.99175
90320   -118.49353    69.53716   125.89818     1.19650    40.65441

        Feature_87  Feature_88  Feature_89  Feature_90
201297  -11.98130    28.37790    98.16031    10.25062
75576   322.64443    12.31236    29.27734    24.41026
46834  -237.89694    12.12714   144.67759   -12.74435
500468   46.68713    -9.97428   -83.12116    -3.98244
90320   176.24959     7.68148   118.45818     6.11999
```

Melakukan data sampling sebanyak 8000 baris untuk mempercepat proses pelatihan atau training data pipeline dan juga hyperparameter tuning.

5. Explanatory data

```
Informasi dasar dataset:
<class 'pandas.core.frame.DataFrame'>
Index: 8000 entries, 201297 to 407941
Data columns (total 91 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Feature_0   8000 non-null   int64
 1   Feature_1   8000 non-null   float64
 2   Feature_2   8000 non-null   float64
 3   Feature_3   8000 non-null   float64
 4   Feature_4   8000 non-null   float64
 5   Feature_5   8000 non-null   float64
 6   Feature_6   8000 non-null   float64
 7   Feature_7   8000 non-null   float64
 8   Feature_8   8000 non-null   float64
 9   Feature_9   8000 non-null   float64
 10  Feature_10  8000 non-null   float64
 11  Feature_11  8000 non-null   float64
 12  Feature_12  8000 non-null   float64
 13  Feature_13  8000 non-null   float64
 14  Feature_14  8000 non-null   float64
 15  Feature_15  8000 non-null   float64
 16  Feature_16  8000 non-null   float64
 17  Feature_17  8000 non-null   float64
 18  Feature_18  8000 non-null   float64
 19  Feature_19  8000 non-null   float64
 20  Feature_20  8000 non-null   float64
 21  Feature_21  8000 non-null   float64
 22  Feature_22  8000 non-null   float64
 23  Feature_23  8000 non-null   float64
 24  Feature_24  8000 non-null   float64
 25  Feature_25  8000 non-null   float64
 26  Feature_26  8000 non-null   float64
 27  Feature_27  8000 non-null   float64
 28  Feature_28  8000 non-null   float64
 29  Feature_29  8000 non-null   float64
 30  Feature_30  8000 non-null   float64
 31  Feature_31  8000 non-null   float64
 32  Feature_32  8000 non-null   float64
 33  Feature_33  8000 non-null   float64
 34  Feature_34  8000 non-null   float64
 35  Feature_35  8000 non-null   float64
 36  Feature_36  8000 non-null   float64
 37  Feature_37  8000 non-null   float64
 38  Feature_38  8000 non-null   float64
...
 90  Feature_90  8000 non-null   float64
dtypes: float64(90), int64(1)
```

Dataset ini terdiri dari 8000 entri dan 91 kolom, yang semuanya merupakan tipe data numerik (float64) kecuali satu kolom pertama. Tidak adanya nilai null menunjukkan bahwa dataset sudah bersih dan siap untuk dianalisis lebih lanjut. Dengan jumlah fitur yang cukup besar, penting untuk memahami bagaimana setiap fitur berkontribusi terhadap target atau tujuan analisis. Proses ini dapat melibatkan analisis korelasi untuk mengidentifikasi fitur-fitur dengan hubungan kuat, sekaligus mengeliminasi fitur dengan variansi rendah yang mungkin tidak relevan.

6. Statistik deskriptif

```
Statistik deskriptif:
        Feature_0    Feature_1    Feature_2    Feature_3    Feature_4  \
count  8000.000000  8000.000000  8000.000000  8000.000000  8000.000000
mean   1998.399625    43.502842     1.691354     8.926562     0.886095
std      10.938699     6.019570    51.921426    34.700056    15.865213
min    1929.000000     1.749000  -287.287360  -164.563050  -100.944550
25%    1994.000000    40.055082   -25.514950   -11.154170    -8.727373
50%    2002.000000    44.387890     8.742010    10.618645    -0.785780
75%    2006.000000    47.804140    36.869750    29.305752     8.567410
max    2010.000000    58.483620   182.620710   281.153890   118.036030

         Feature_5    Feature_6    Feature_7    Feature_8    Feature_9  ...  \
count  8000.000000  8000.000000  8000.000000  8000.000000  8000.000000  ...
mean     -6.635673    -9.656860    -2.261605    -1.860409     3.870715  ...
std      22.688105    12.787412    14.471103     7.717207    10.650746  ...
min    -112.436140   -59.999320   -91.558200   -39.386550   -74.049010  ...
25%     -20.332662   -18.433715   -10.679735    -6.516920    -2.151652  ...
50%      -6.133170   -11.448420    -1.890630    -1.821125     3.873930  ...
75%       7.462668    -2.525457     6.634370     2.709293    10.283115  ...
max     191.615990    68.300090   140.558400    45.010470    61.108430  ...

        Feature_81   Feature_82   Feature_83   Feature_84   Feature_85  \
count  8000.000000  8000.000000  8000.000000  8000.000000  8000.000000
mean     15.893349   -71.293261    39.785701    38.871935     0.365204
std      32.519537   171.549967   120.816625    92.907016    16.303763
min    -224.101380 -2349.595000  -804.746320  -918.373760  -146.386200
25%      -1.569180  -135.935135   -21.484977    -2.421575    -6.571833
50%       8.959315   -50.646880    28.058230    34.436075     0.759950
75%      25.860568    13.284695    87.983603    77.437195     8.425840
max     587.732680  1498.541460  3210.701700  1097.919030   189.061420

        Feature_86   Feature_87   Feature_88   Feature_89   Feature_90
count  8000.000000  8000.000000  8000.000000  8000.000000  8000.000000
mean     16.429972   -25.160682     4.308544    21.317853     1.108753
std     110.501122   169.321320    13.434088   198.918644    22.229509
min   -1666.355050 -2179.274330   -85.488880 -1985.137200  -258.215020
25%     -31.840355   -99.365930    -2.559852   -57.131825    -9.027648
50%      16.420775   -20.071060     3.163455     7.889145    -0.232080
75%      67.095935    53.953180     9.932110    81.720902     9.412835
max     900.187190  1251.217210   275.353660  7393.398440   251.842650

[8 rows x 91 columns]
```

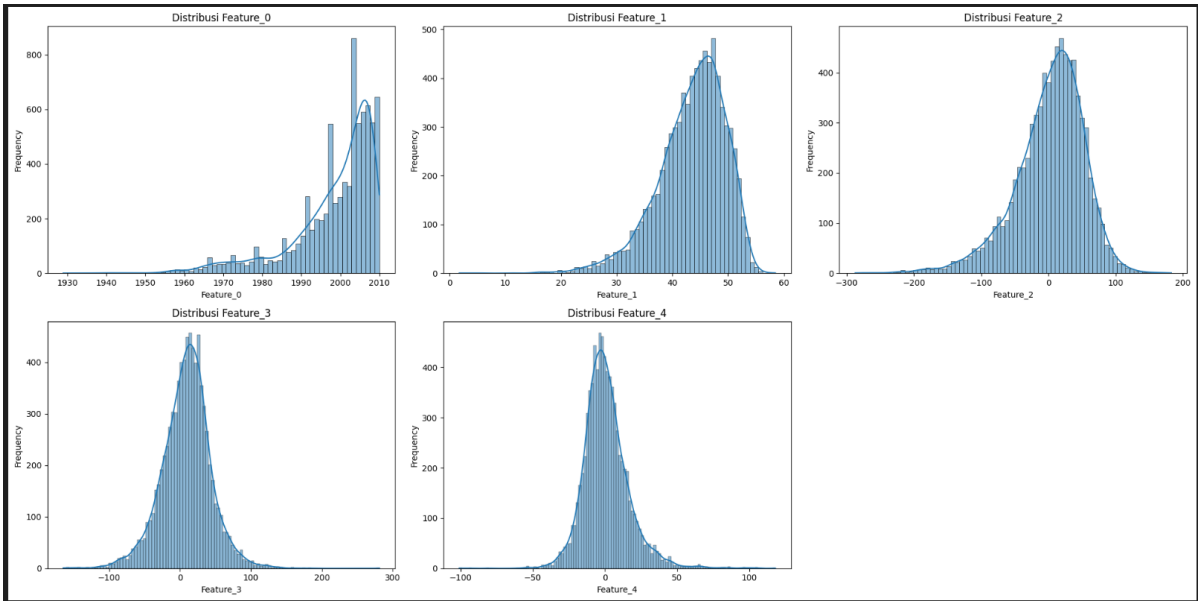
Dataset ini memiliki 8000 entri dengan statistik deskriptif yang memberikan gambaran tentang distribusi nilai untuk setiap fitur. Nilai rata-rata dan standar deviasi menunjukkan adanya variasi yang cukup besar di antara fitur, seperti pada *Feature_81* yang memiliki rata-rata sekitar 80, namun dengan standar deviasi yang sangat besar, yaitu 171.549967. Hal ini menunjukkan bahwa fitur tersebut memiliki penyebaran nilai yang luas, kemungkinan dengan outlier atau nilai ekstrem. Sebaliknya, fitur-fitur seperti *Feature_4* memiliki penyebaran nilai yang lebih terkonsentrasi dengan standar deviasi yang relatif kecil.

7. Missing Values

```
Missing values:
Feature_0     0
Feature_1     0
Feature_2     0
Feature_3     0
Feature_4     0
             ..
Feature_86    0
Feature_87    0
Feature_88    0
Feature_89    0
Feature_90    0
Length: 91, dtype: int64
```

Dataset ini tidak memiliki nilai yang hilang pada semua 91 kolom, yang memastikan data bersih dan siap untuk analisis lebih lanjut tanpa perlu langkah imputasi atau penanganan nilai kosong. Hal ini sangat menguntungkan untuk efisiensi dalam proses eksplorasi data, pelatihan model, dan tuning parameter, karena tidak ada risiko bias atau gangguan akibat data yang tidak lengkap.

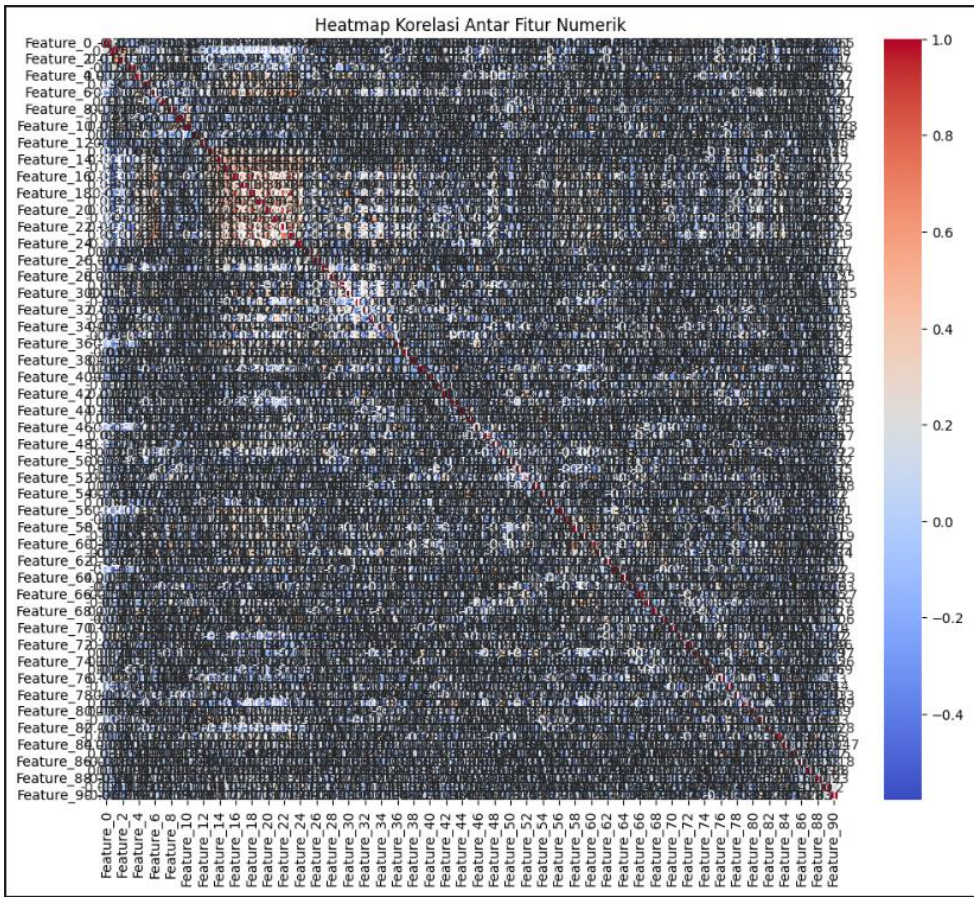
8. Visulisasi dasar distribusi data



Dari visualisasi distribusi, *Feature_0* menunjukkan pola distribusi yang meningkat secara bertahap dan berfokus pada nilai yang lebih baru, yang mungkin mencerminkan elemen temporal. Sementara itu, *Feature_1* hingga *Feature_4* memiliki pola distribusi yang menyerupai distribusi

normal dengan berbagai rata-rata dan penyebaran. *Feature_2*, misalnya, memiliki distribusi yang lebih simetris dibandingkan dengan *Feature_3* dan *Feature_4*, yang menunjukkan variasi data yang lebih terkonsentrasi di sekitar nilai rata-rata.

9. Visualisasi Heatmap



Dari heatmap korelasi antar fitur, terlihat bahwa beberapa fitur memiliki korelasi tinggi (ditandai dengan warna merah yang mendekati 1), sementara yang lain menunjukkan korelasi rendah atau bahkan negatif (berwarna biru). Fitur-fitur dengan korelasi tinggi dapat mengindikasikan redundansi data, sehingga dapat dipertimbangkan untuk dilakukan pengurangan dimensi guna menyederhanakan analisis atau pemodelan.

10. Membagi dataset menjadi data pelatihan dan pengujian

```
Jumlah data pelatihan: 6400
Jumlah data pengujian: 1600
```

Dataset telah dibagi menjadi 6400 data untuk pelatihan dan 1600 data untuk pengujian, yang sesuai dengan rasio umum 80:20. Proporsi ini memastikan model memiliki cukup data untuk belajar (training) sekaligus menyediakan data yang memadai untuk mengevaluasi performa (testing). Dengan pembagian ini, model yang dilatih diharapkan dapat menggeneralisasi dengan baik dan memberikan hasil yang akurat saat diaplikasikan pada data baru yang belum pernah dilihat sebelumnya

11. Pipeline Model Polynomial Regression

```
Polynomial Regression - MSE: 2156.8145542277266
Polynomial Regression - R^2: -3.699270086618289
```

Hasil evaluasi Polynomial Regression menunjukkan nilai **MSE (Mean Squared Error)** sebesar 2156.81 dan **R^2 (R-squared)** sebesar -3.69. Nilai MSE yang tinggi mengindikasikan bahwa model memiliki kesalahan prediksi yang cukup besar pada data pengujian. Selain itu, nilai R^2 negatif menunjukkan bahwa model tidak mampu menjelaskan variansi data target dengan baik, bahkan performanya lebih buruk dibandingkan jika menggunakan rata-rata target sebagai prediksi. Hal ini bisa disebabkan oleh overfitting, pemilihan derajat polinomial yang tidak sesuai, atau fitur yang kurang relevan.

12. Pipeline Model Decision Tree Regression

```
Decision Tree Regression - MSE: 501.01313789677283
Decision Tree Regression - R^2: -0.09160801391387507
```

Hasil evaluasi Decision Tree Regression menunjukkan nilai **MSE (Mean Squared Error)** sebesar 501.01 dan **R^2 (R-squared)** sebesar -0.091. Nilai MSE yang lebih rendah dibandingkan Polynomial Regression menunjukkan bahwa model ini memiliki kesalahan prediksi yang lebih kecil. Namun, nilai R^2 yang negatif tetap mengindikasikan bahwa model tidak mampu menjelaskan variansi data dengan baik, dan performanya lebih buruk dibandingkan prediksi menggunakan rata-rata target. Hal ini mungkin disebabkan oleh overfitting pada Decision Tree.

13. Pipeline Model k-NN Regression

```
k-NN Regression - MSE: 328.0115614035456
k-NN Regression - R^2: 0.2853280243555468
```

Hasil evaluasi k-NN Regression menunjukkan nilai **MSE (Mean Squared Error)** sebesar 328.01 dan **R^2 (R-squared)** sebesar 0.285. Nilai MSE yang lebih rendah dibandingkan Polynomial Regression dan Decision Tree Regression menunjukkan bahwa model ini memberikan prediksi dengan kesalahan yang lebih kecil. Selain itu, nilai R^2 yang positif (0.285) mengindikasikan bahwa model dapat menjelaskan sekitar 28.5% variansi data target, meskipun belum optimal.

14. Pipeline Model XGBoost Regression

```
XGBoost Regression - MSE: 209.6474879756609
XGBoost Regression - R^2: 0.5432198067063536
```

Hasil evaluasi XGBoost Regression menunjukkan nilai **MSE (Mean Squared Error)** sebesar 209.65 dan **R^2 (R-squared)** sebesar 0.543. Nilai MSE yang paling rendah di antara model sebelumnya mengindikasikan bahwa XGBoost menghasilkan prediksi dengan kesalahan terkecil. Selain itu, nilai R^2 sebesar 0.543 menunjukkan bahwa model dapat menjelaskan sekitar 54.3% variansi dalam data target, menjadikannya model dengan performa terbaik dalam eksperimen ini.

15. Hyperparameter Tuning Polynomial Regression

```
Best parameters for Polynomial Regression: {'poly_features__degree': 1}
Best R^2 score for Polynomial Regression: 0.3172688538431502
```

Setelah dilakukan tuning hyperparameter, hasil terbaik untuk Polynomial Regression didapatkan dengan derajat polinomial (**degree**) sebesar 1, yang secara efektif sama dengan regresi linear. **R^2** terbaik yang dicapai adalah 0.317, menunjukkan bahwa model mampu menjelaskan sekitar 31.7% variansi data target. Hasil ini menandakan bahwa derajat polinomial yang lebih tinggi justru menyebabkan overfitting, dan model sederhana seperti regresi linear lebih efektif untuk dataset ini.

16. Hyperparameter Tuning Decision Tree Regression

```
Best parameters for Decision Tree Regression: {'regressor__max_depth': 5, 'regressor__min_samples_split': 5}
Best R^2 score for Decision Tree Regression: 0.24964402236571726
```

Setelah dilakukan tuning hyperparameter, model **Decision Tree Regression** mencapai performa terbaik dengan parameter **max_depth = 5** dan **min_samples_split = 5**. Nilai **R^2** terbaik adalah 0.249, menunjukkan bahwa model ini mampu menjelaskan sekitar 24.9% variansi data target. Meskipun performanya meningkat dibandingkan sebelumnya, model ini masih kurang optimal dibandingkan dengan model lain seperti XGBoost. Parameter yang digunakan tampaknya membantu mengurangi overfitting dengan membatasi kedalaman pohon (*max_depth*) dan jumlah sampel minimum untuk split (*min_samples_split*).

17. Hyperparameter Tuning Decision k-NN Regression

```
Best parameters for k-NN Regression: {'regressor__n_neighbors': 7}
Best R^2 score for k-NN Regression: 0.31401369267170043
```

Setelah tuning hyperparameter, model **k-NN Regression** mencapai performa terbaik dengan parameter **n_neighbors = 7**, menghasilkan **R²** sebesar 0.314. Ini menunjukkan bahwa model mampu menjelaskan sekitar 31.4% variansi data target, yang lebih baik dibandingkan Decision Tree tetapi masih kalah dibandingkan XGBoost. Parameter ini menunjukkan bahwa mempertimbangkan 7 tetangga terdekat memberikan keseimbangan optimal antara underfitting dan overfitting.

18. Hyperparameter Tuning Decision k-NN Regression

```
Best parameters for XGBoost Regression: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}
Best R^2 score for XGBoost Regression: 0.6357064228285454
```

Setelah tuning hyperparameter, model **XGBoost Regression** mencapai performa terbaik dengan parameter **learning_rate = 0.1**, **max_depth = 3**, dan **n_estimators = 200**, menghasilkan nilai **R² = 0.636**. Ini menunjukkan bahwa XGBoost mampu menjelaskan sekitar 63.6% variansi dalam data target, menjadikannya model dengan performa terbaik dibandingkan model lainnya. Hyperparameter yang digunakan menunjukkan bahwa kombinasi kedalaman pohon yang moderat (*max_depth = 3*) dan jumlah estimator yang cukup banyak (*n_estimators = 200*) memberikan hasil optimal.

KESIMPULAN

Berdasarkan evaluasi dan tuning hyperparameter, model **XGBoost Regression** memberikan performa terbaik dengan nilai **R² sebesar 0.636** dan **MSE terendah** dibandingkan dengan Polynomial Regression, Decision Tree Regression, dan k-NN Regression. Hal ini menunjukkan bahwa XGBoost mampu menangkap pola kompleks dalam data dengan baik dan memiliki kemampuan generalisasi yang lebih baik dibandingkan model lainnya. Model ini mengungguli model lain berkat kombinasi hyperparameter optimal seperti **learning_rate = 0.1**, **max_depth = 3**, dan **n_estimators = 200**, yang menjaga keseimbangan antara bias dan variansi.

Meskipun Polynomial Regression dan k-NN Regression memiliki R^2 yang lebih rendah, tuning hyperparameter berhasil meningkatkan performa kedua model tersebut, dengan k-NN mencapai R^2 sebesar 0.314. Decision Tree Regression memiliki performa yang relatif rendah meskipun setelah tuning, dengan R^2 hanya 0.249. Berdasarkan analisis ini, **XGBoost adalah model yang paling direkomendasikan** untuk dataset ini.