UTS REGRESI MODEL

RAISYA ATHAYA KAMILAH

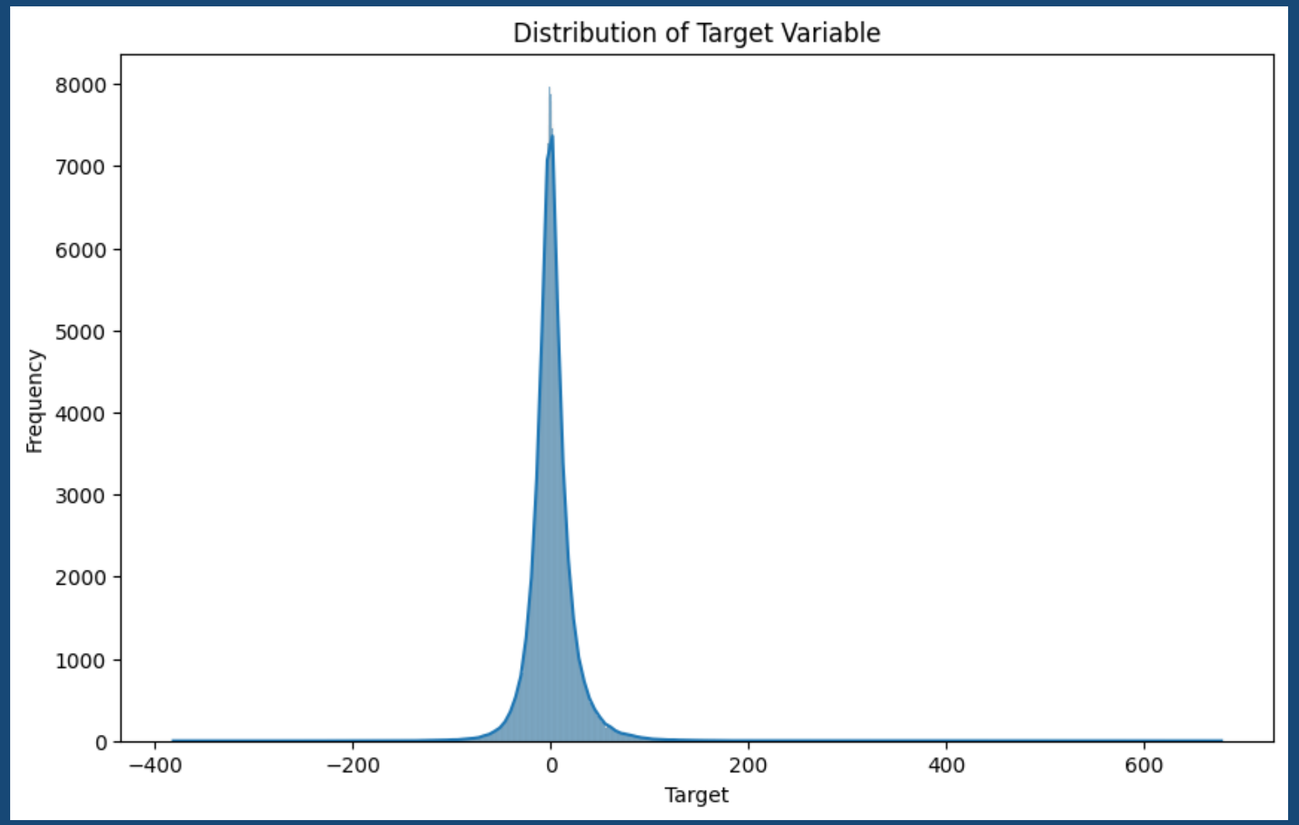
101032380253

TKX-47-01

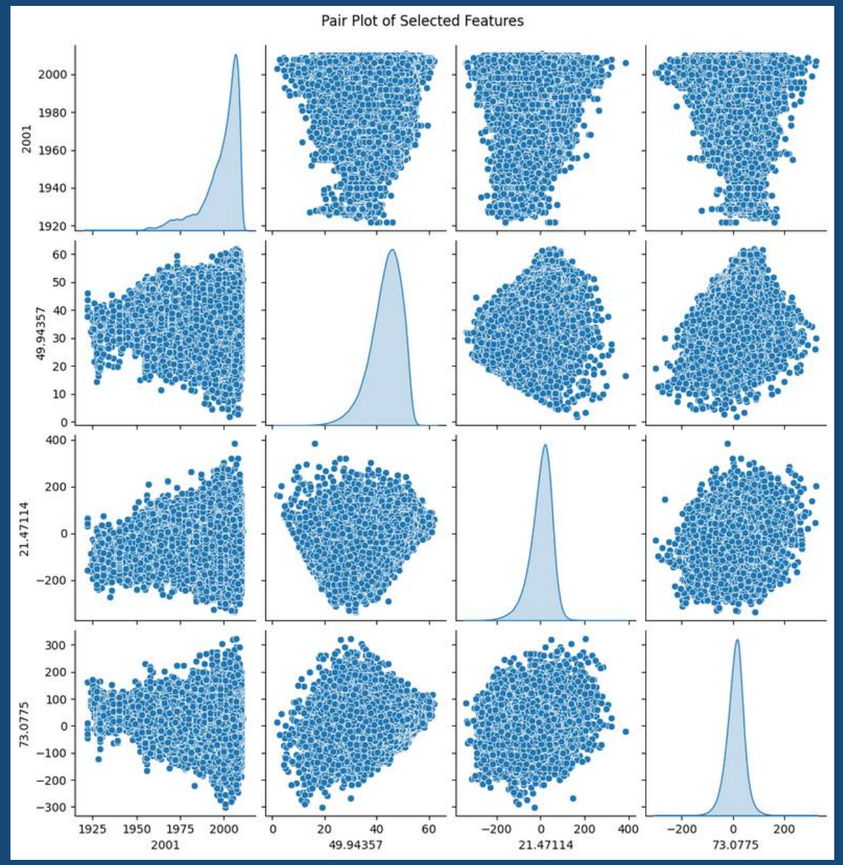
EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) adalah proses awal dalam analisis data yang bertujuan untuk mengeksplorasi dan memahami struktur, pola, hubungan, dan distribusi data sebelum melakukan analisis lebih lanjut atau pembangunan model prediktif. Tujuannya untuk memahami Karakteristik Data,Menemukan Pola dan Hubungan,Pengecekan Missing Data

VISUALIZATION DATA

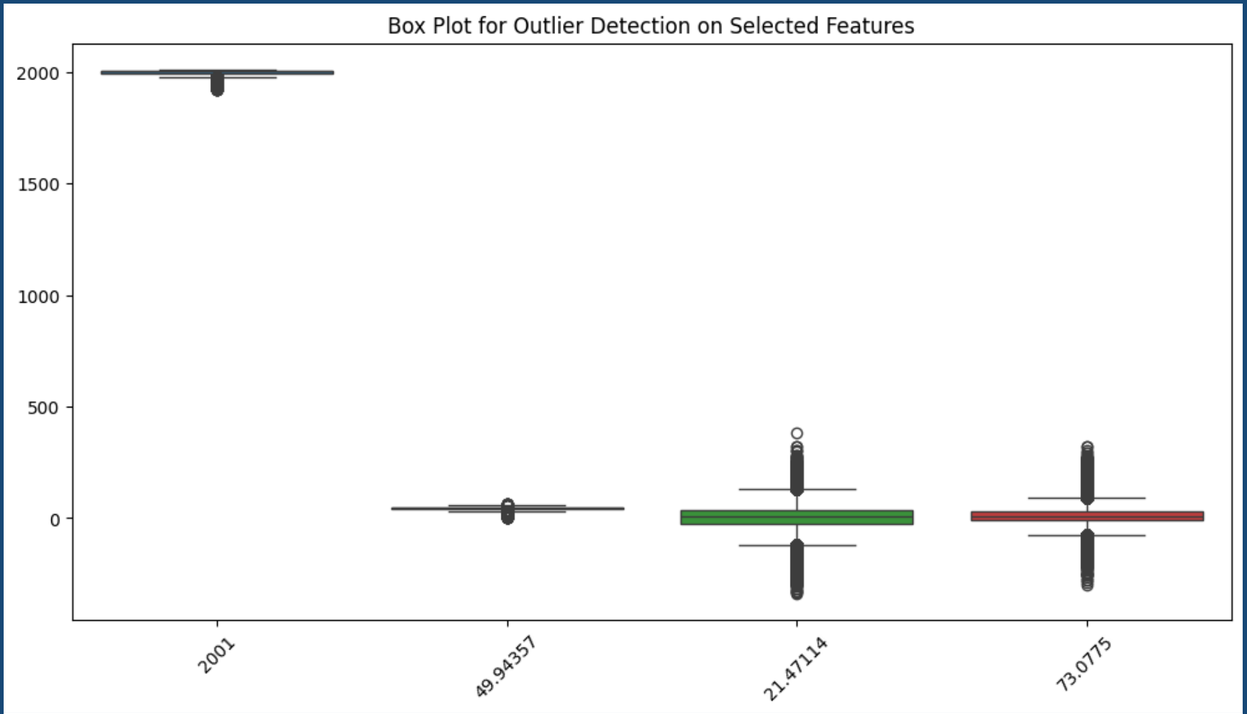


Pada Distribusi Data kebanyakan terpusat di Nol yang berarti kemungkinan besar memiliki distribusi skewed (miring), dengan banyak nilai yang cenderung lebih kecil atau lebih dekat ke nol. Dan Distribusi Tidak Normal karena distribusinya terkonsentrasi di satu sisi dan tidak berbentuk lonceng (normal)



Pada fitur 2001 menunjukkan distribusi miring dengan sebagian besar data terkonsentrasi pada nilai yang lebih rendah dan beberapa nilai yang sangat tinggi di sisi kanan. Fitur 49.94357 menunjukkan distribusi yang sangat miring ke kanan, dengan banyak nilai yang sangat rendah dan beberapa nilai ekstrem yang sangat tinggi.
Fitur '21.47114' dan '73.0775' memiliki distribusi yang lebih normal (kurang miring), menunjukkan lebih banyak data terdistribusi di sekitar nilai tengah.

Namun pada Fitur 2001 memiliki korelasi yang cukup kuat dengan '49.94357' dan '73.0775'.
Fitur '49.94357' dan '73.0775' juga memiliki korelasi positif yang cukup tinggi.
Fitur '49.94357' dan '21.47114' tidak memiliki korelasi sangat kuat karena lebih tersebar.

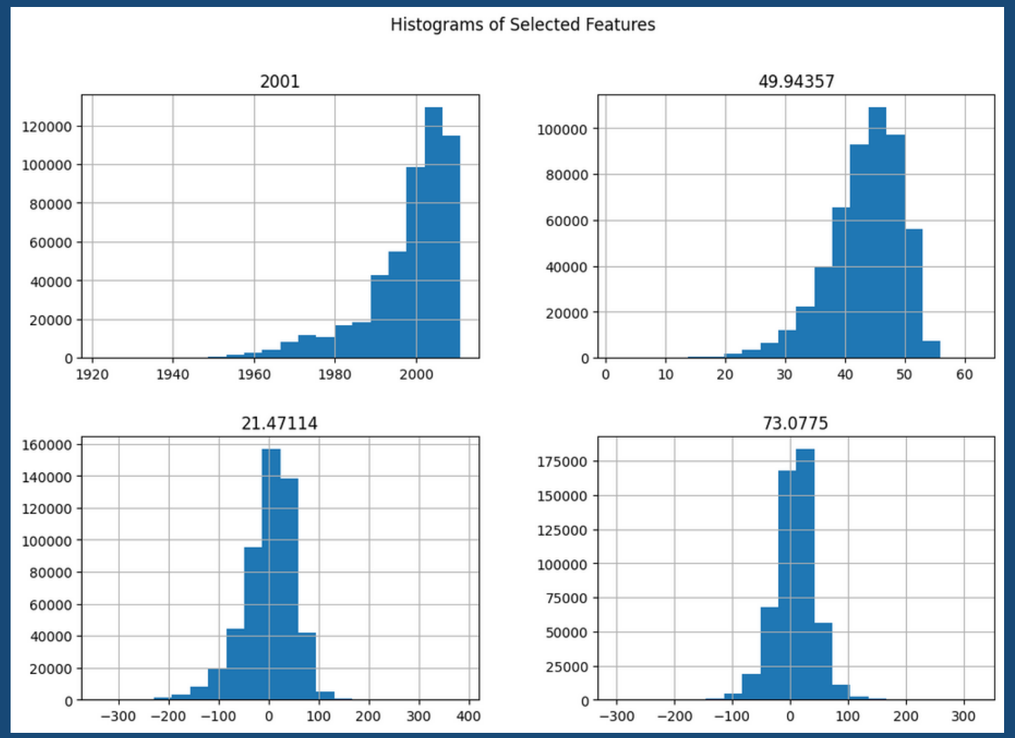


Pada fitur 2001 memiliki banyak outliers yang berarti adanya nilai ekstrem yang lebih tinggi daripada sebagian besar data lainnya.

Fitur '49.94357' Memiliki beberapa outliers di kedua sisi (kiri dan kanan), menandakan adanya nilai yang sangat rendah atau tinggi.

Fitur '21.47114' Tidak ada outliers besar pada fitur ini, yang menunjukkan data lebih terdistribusi normal.

Fitur '73.0775' Ada outliers pada sisi kanan, menunjukkan nilai yang sangat besar



Fitur '2001' memiliki distribusi sangat miring ke kanan Sebagian besar data terkonsentrasi pada nilai tahun yang lebih rendah, dengan lonjakan yang tajam setelah sekitar tahun 2000.

Fitur '49.94357' juga Distribusi miring ke kanan (right skewed), dengan puncak distribusi pada nilai sekitar 35-40. Namun Sebagian besar data terfokus pada nilai yang lebih rendah, dan ada beberapa nilai yang lebih tinggi yang lebih jarang (outliers).

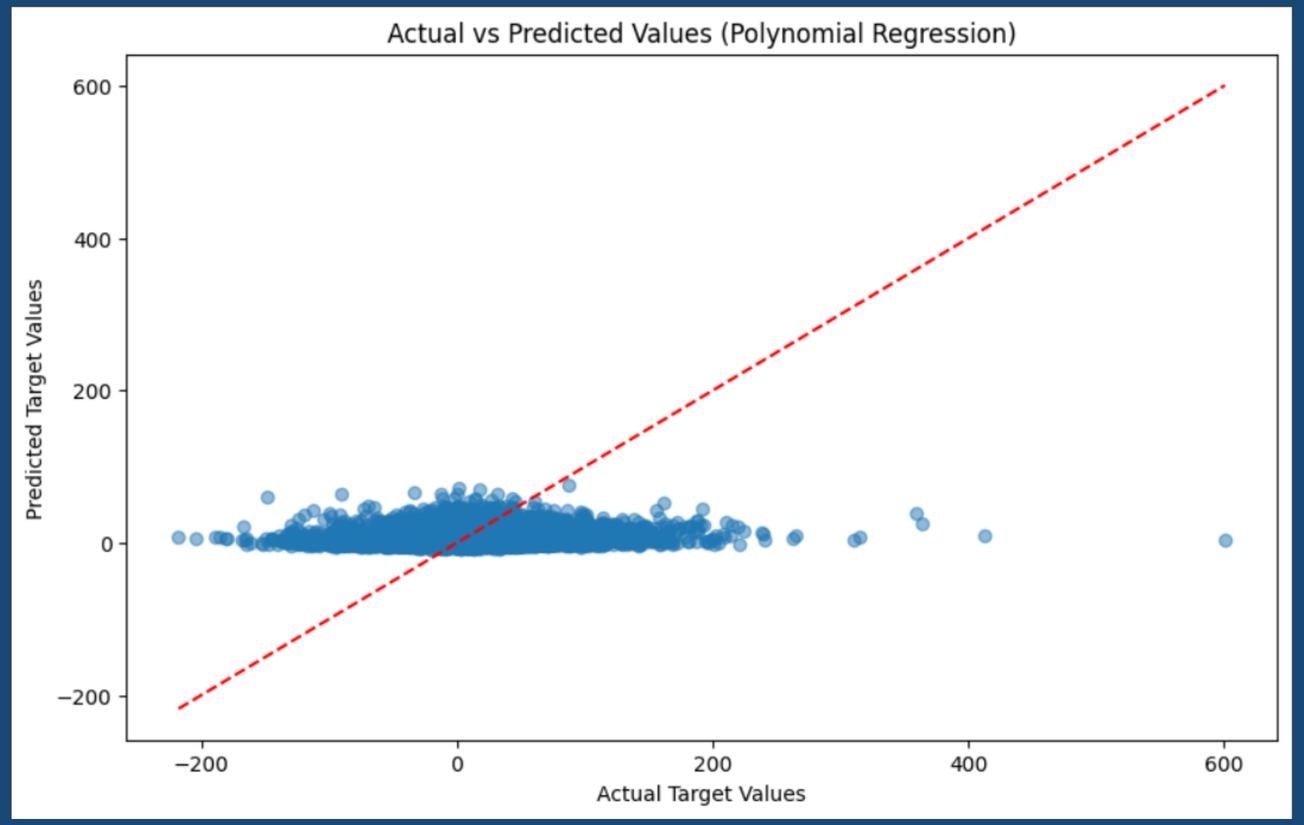
Fitur '21.47114' Distribusi dengan pola mirip distribusi normal namun Ada sedikit nilai ekstrem di kedua sisi, tetapi mayoritas data terdistribusi di sekitar nilai 0

Fitur '73.0775' Distribusi miring ke kanan (right skewed). Sebagian besar data terkonsentrasi pada nilai yang lebih rendah (sekitar 0-100), dengan sedikit data yang lebih tinggi (outliers di sisi kanan)

POLYNOMIAL REGRESSION

(BASIS FUNCTION)

Polynomial Regression adalah jenis regresi yang digunakan untuk memodelkan hubungan non-linier antara fitur (input) dan target (output). Polynomial Regression menambahkan pangkat fitur, seperti x^2 atau x^3, untuk menangkap pola yang lebih rumit.

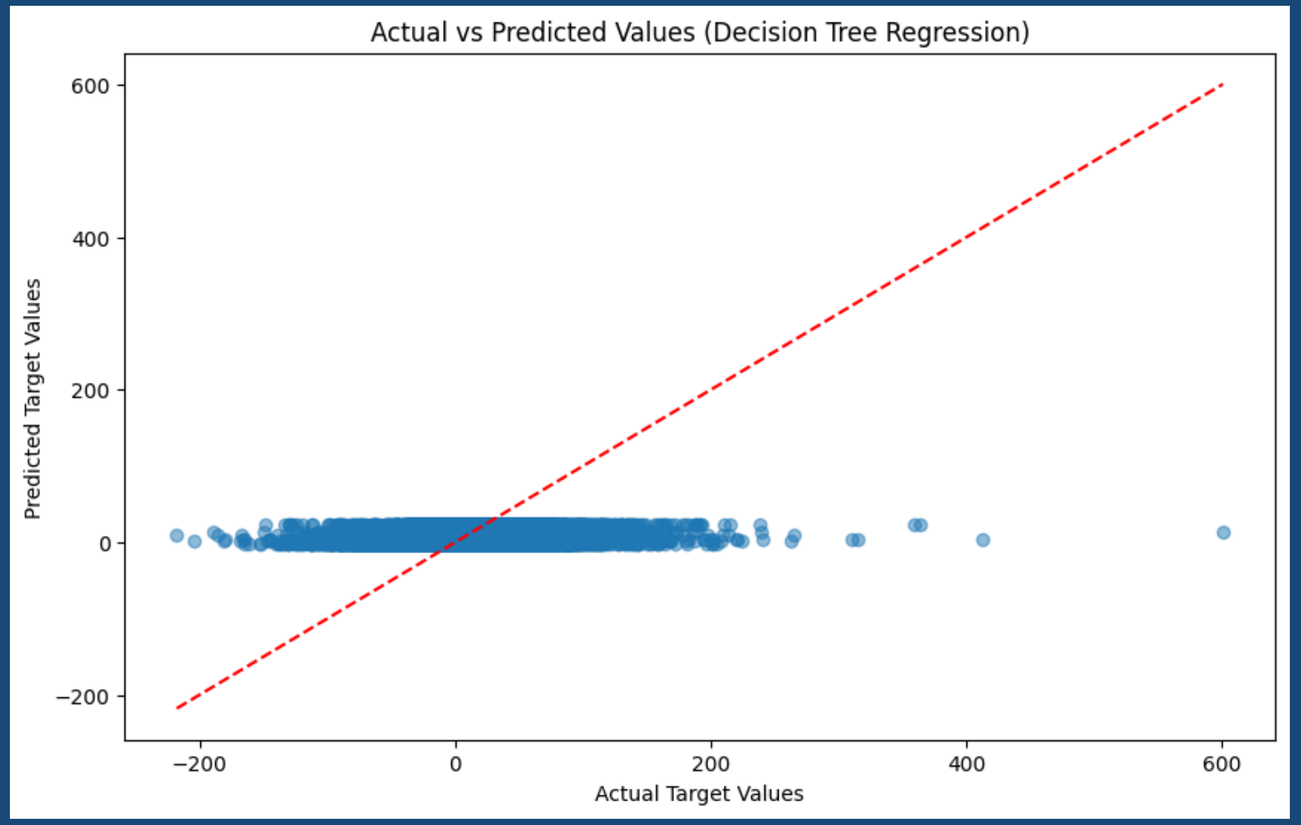


Polynomial Regression Mean Squared Error (MSE): 452.44972549756955
Polynomial Regression Root Mean Squared Error (RMSE): 21.270865649934642
Polynomial Regression R^2 Score: 0.07095244874873963

Model kesulitan dalam memprediksi nilai-nilai yang lebih tinggi dari target. Titik data cenderung terkonsentrasi pada bagian bawah grafik (nilai prediksi mendekati nol).
Nilai MSE Besar yang berarti bahwa model tidak sangat akurat dalam memprediksi data. RMSE juga menunjukkan kesulitan membedakan prediksi dan actual. Dan memiliki nilai R^2 hanya dapat menjelaskan sekitar 7% dari variabilitas data. Dengan kata lain, model ini tidak cukup baik dalam menjelaskan hubungan antara fitur dan target.
Model underfitting karena terlalu sederhana.

DECISION TREE

Decision Tree Regression adalah metode regresi yang menggunakan struktur pohon keputusan untuk memodelkan hubungan antara fitur (input) dan target (output).

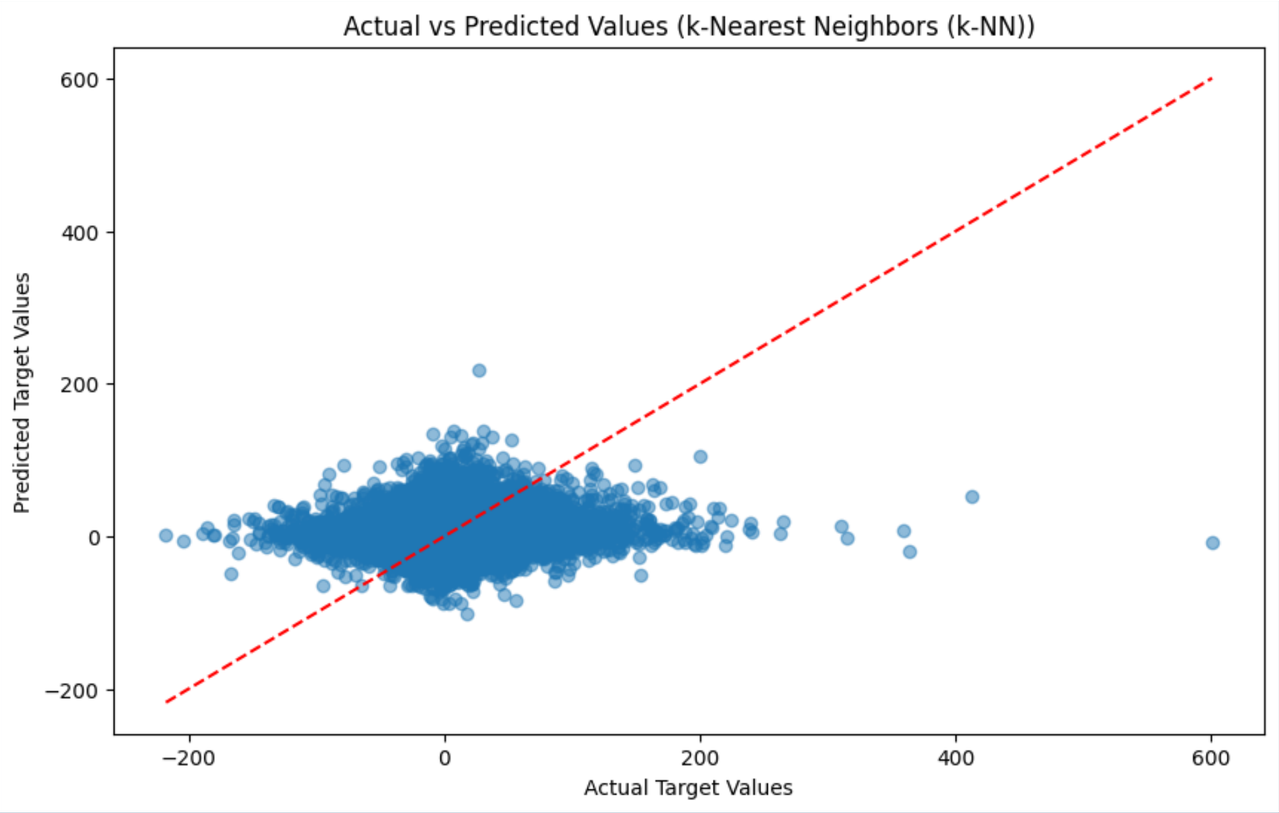


Decision Tree Regression Mean Squared Error (MSE): 460.421577197265
Decision Tree Regression Root Mean Squared Error (RMSE): 21.457436407857884
Decision Tree Regression R^2 Score: 0.05458327249960948

Plot persebaran menunjukkan bahwa prediksi model sangat terpusat di sekitar nilai 0 . Model ini gagal menangkap variabilitas yang lebih tinggi dari nilai target. Nilai MSE sangat tinggi yang berarti kurang akurat dalam prediksi. Pada RMSE pun masih kurang bisa membedakan prediksi dan aktual.
R^2 yang sekitar 5.5% menunjukkan bahwa model ini hanya mampu menjelaskan sekitar 5.5% dari variasi dalam dataset. Model juga underfitting

K-NN

k-NN Regression (k-Nearest Neighbors Regression) adalah metode regresi yang memprediksi nilai target berdasarkan rata-rata nilai target dari k tetangga terdekat di ruang fitur. K-NN mencari data yang mirip dengan data yang ingin diprediksi, kemudian menggunakannya untuk membuat prediksi



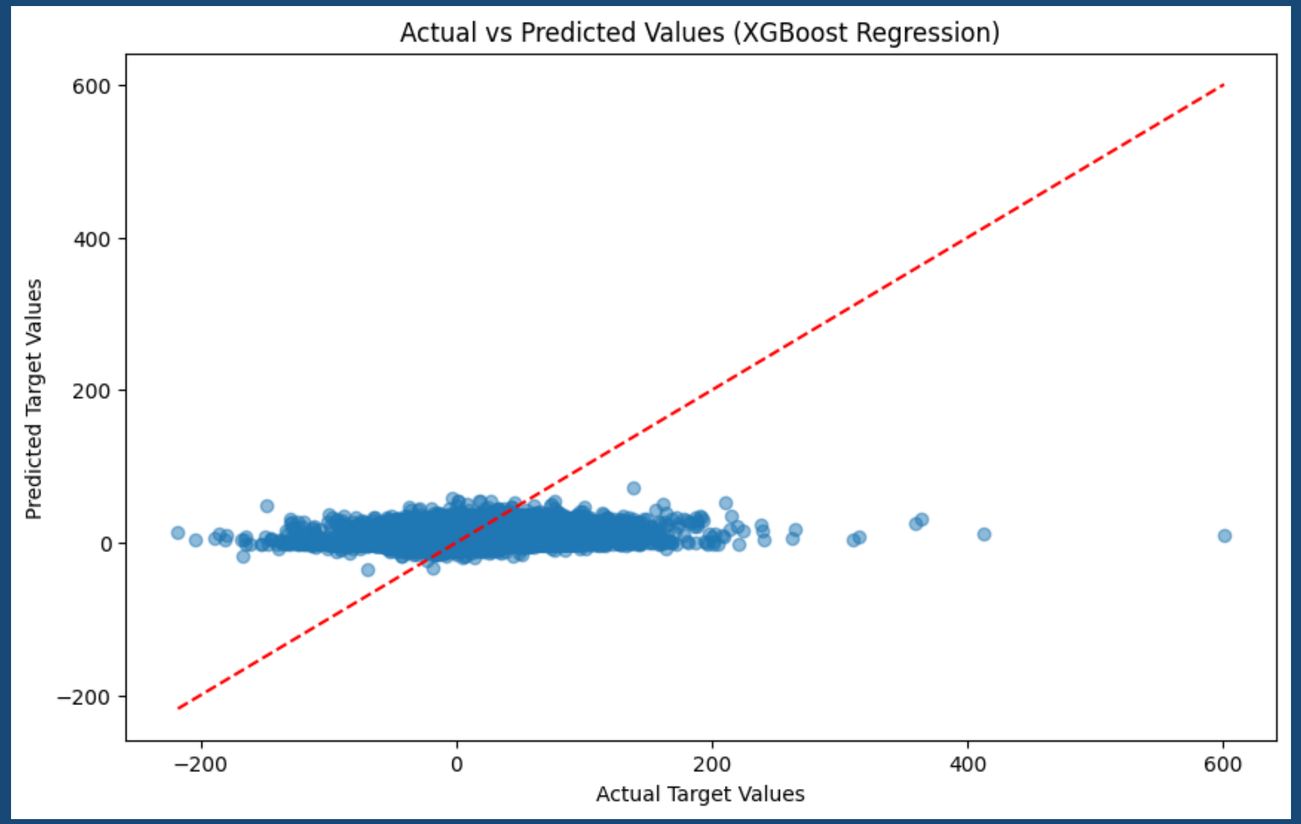
k-Nearest Neighbors (k-NN) Mean Squared Error (MSE): 602.5315239108189
k-Nearest Neighbors (k-NN) Root Mean Squared Error (RMSE): 24.54651755159617
k-Nearest Neighbors (k-NN) R^2 Score: -0.23722129840046402

k-Nearest Neighbors (k-NN) Mean Squared Error (MSE): 602.5315239108189
k-Nearest Neighbors (k-NN) Root Mean Squared Error (RMSE): 24.54651755159617
k-Nearest Neighbors (k-NN) R^2 Score: -0.23722129840046402

- *Karena K-NN nya di regresi bukan di klasifikasi maka yang perlukan hanya menggunakan metrik seperti MSE (Mean Squared Error), RMSE (Root Mean Squared Error), dan R^2 akan lebih berguna daripada visualisasi.*

XGBOOST REGRESSION

XGBoost Regression adalah teknik regresi yang menggunakan algoritma gradient boosting untuk membuat prediksi pada data. XGBoost, yang merupakan singkatan dari Extreme Gradient Boosting, adalah versi yang lebih canggih dan efisien dari metode gradient boosting yang populer, dengan penekanan pada kecepatan dan kinerja.



XGBoost Regression Mean Squared Error (MSE): 449.7842162550794
XGBoost Regression Root Mean Squared Error (RMSE): 21.208116754089207
XGBoost Regression R^2 Score: 0.07642573051910595

Model ini pun menandakan bahwa model tidak dapat menangkap variabilitas yang besar pada target yang lebih tinggi. Nilai MSE masih errornya besar namun dibanding dengan regresi lain XGBOOST yang memiliki error terkecil. RMSE masih pun masih kurang bisa membedakan prediksi dan aktual. Nilai R^2 di sekitar 7.4% yang berarti bahwa model ini hanya mampu menjelaskan sekitar 7.4% dari variasi dalam dataset tapi memiliki nilai R^2 paling tinggi dibanding dengan regressi lainnya.

KESIMPULAN

Dapat disimpulkan bahwa pada ke empat pipeline regressi untuk kasus dataset ini kebanyakan kurang cocok dan kurang mampu memperoleh nilai metrik yang lebih baik. Namun berdasarkan nilai metrik seperti MSE dan R^2 yang memiliki hasil terbaik untuk kasus dataset ini adalah pipeline XGBOOST Regression karena menggunakan teknik boosting untuk meningkatkan performa. Boosting bekerja dengan membangun model secara bertahap, memperbaiki kesalahan model sebelumnya pada setiap iterasi atau lebih efektif dalam menangani data yang kompleks. Dan perlu setting hyperparameter tuning lebih lanjut agar dapat memperoleh hasil nilai metrik yang lebih baik.

Link Youtube:
https://youtu.be/4cywSOxxrBc

THANK YOU

THANK YOU
THANK YOU