1. Jika model linear regression atau decision tree mengalami underfitting pada dataset ini, strategi apa yang akan digunakan untuk meningkatkan performanya? Bandingkan setidaknya dua pendekatan berbeda (misal: transformasi fitur, penambahan features, atau perubahan model ke algoritma yang lebih kompleks), dan jelaskan bagaimana setiap solusi memengaruhi bias-variance tradeoff!

   Jika model mengalami underfitting, ada beberapa pada preprocessing dan training yang dapat dilakukan.
   Pada preprocessing, dapat melakukan:
   a. Seleksi fitur, seperti multicollinearity, ANOVA, mutual information, dll. Teknik-teknik ini akan memilih fitur-fitur terbaik untuk digunakan berdasarkan teknik masing-masing.
   b. Transformasi data, seperti scaling (robust, standard, dll), Yeo-Johnson, Min-Max, dll. Transformasi data dilakukan ketika fitur-fitur memiliki data yang skew positif atau negatif, ataupun untuk menscaling data. Transformasi sangat berpengaruh bagi model selain tree (atau turunannya).

   Pada saat training model, dapat dilakukan tuning parameter, seperti mencari max_depth atau n_neighbors terbaik pada model. Dapat dilakukan dengan Grid Search.

2. Selain MSE, jelaskan dua alternatif loss function untuk masalah regresi (misal: MAE, Huber loss) dan bandingkan keunggulan serta kelemahannya. Dalam skenario apa setiap loss function lebih cocok digunakan? (Contoh: data dengan outlier, distribusi target non-Gaussian, atau kebutuhan interpretasi model).

   Mean Absolute Error (MAE) dapat memiliki keunggulan yaitu lebih robust terhadap outlier karena error tidak dikuadratkan. MAE merupakan rata-rata error absolut. Cocok digunakan ketika data mengandung outlier, atau ingin model yang robust terhadap nilai ekstrim.
   Huber Loss merupakan gabungan MSE dan MAE. Huber Loss sensitive terhadap error kecil, namun tetap robust terhadap outlier besar. Kelemahan metrik ini yaitu memerlukan parameter dan komputasinya lebih kompleks dibandingkan MSE atau MAE. Cocok digunakan ketika ada outlier namun ingin mempertahankan sensitivitas terhadap error kecil, atau ketika tidak yakin seberapa besar noise atau outlier dan ingin model yang adaptif.

3. Tanpa mengetahui nama fitur, metode apa yang dapat digunakan untuk mengukur pentingnya setiap fitur dalam model? Jelaskan prinsip teknikal di balik metode tersebut (misal: koefisien regresi, feature importance berdasarkan impurity reduction) serta keterbatasannya!

   Impurity-Based Feature Importance adalah mengukur berapa banyak penurunan impurity (gini, entropy, dll). Biasanya ini berada langsung dari model, terutama model

tree. Fitur ini bias terhadap fitur dengan banyak variabel numerik, dan tidak konsisten jika fitur saling berkorelasi.

4. Bagaimana mendesain eksperimen untuk memilih hyperparameter optimal (misal: learning rate untuk SGDRegressor, max_depth untuk Decision Tree) pada dataset ini? Sertakan analisis tradeoff antara komputasi, stabilitas pelatihan, dan generalisasi model!

   Pada dataset ini, dilakukan eksperimen untuk mencari parameter terbaik dengan menggunakan Grid Search. Parameter yang dicari yaitu max_depth untuk decision tree, n_neighbors untuk knn, kernel untuk SVR. Grid Search menguji satu per satu parameter yang ingin dicari. Kelemahan terbesar yaitu komputasi mahal, apalagi jika menggunakan dataset yang sangat besar (seperti dataset ini).

   Pengujian ada di notebook regression-training.ipynb

5. Jika menggunakan model linear regression dan residual plot menunjukkan pola non-linear serta heteroskedastisitas, langkah-langkah apa yang akan diambil? (contohnya: Transformasi data/ubah model yang akan dipakai/etc).

   Seperti yang dijelaskan pada nomor 1, kita dapat melakukan transformasi data.
   Selain itu, kita dapat menggunakan model non-linear seperti polynomial regression, atau tree-based.
   Model tree-based akan lebih unggul jika pola data non-linear. Pada pengujian, model yang terbaik pada dataset tersebut adalah model XGBRegression, dan satu-satunya yang melebihi R2 Score diatas 0.3.