Nama : Eko Putra Nugraha

NIM : 1103213212

Regresi Linear Model

Langkah pertama adalah menerapkan model regresi linier sederhana, yang kemudian dibandingkan dengan model regresi polinomial untuk menangkap hubungan non-linear antara X dan Y. Metrik Mean Squared Error (MSE) digunakan untuk menilai kinerja kedua model ini, yang menunjukkan bahwa regresi polinomial memberikan hasil sedikit lebih baik dengan MSE lebih rendah, meskipun dengan peringatan terhadap risiko overfitting jika model terlalu kompleks.

Selanjutnya, tiga model regresi yang lebih kompleks diimplementasikan: Decision Tree Regressor, k-Nearest Neighbors (k-NN), dan XGBoost Regressor. Visualisasi hasil prediksi dari ketiga model ini mengungkapkan perbedaan karakteristik masing-masing. Decision Tree menghasilkan prediksi dengan pola langkah-langkah yang tajam, yang menunjukkan ketidakstabilan saat menangani data dengan variasi halus. Model k-NN lebih sensitif terhadap fluktuasi data lokal, yang menyebabkan prediksi yang lebih bervariasi dan sering kali overfitting, terutama pada bagian data tertentu. Sebaliknya, model XGBoost memberikan prediksi yang lebih mulus dan stabil, mampu menangkap tren global tanpa overfitting, menjadikannya pilihan yang lebih baik untuk data dengan pola non-linear.

Kemudian, evaluasi model dilakukan dengan menghitung MSE untuk setiap model regresi, dan XGBoost keluar sebagai pemenang dengan MSE lebih rendah dibandingkan Decision Tree dan k-NN. Meskipun MSE antara Decision Tree dan XGBoost hampir sama, XGBoost dianggap lebih unggul dalam hal kestabilan dan kemampuan untuk menangani data yang lebih kompleks.

Langkah terakhir adalah tuning hyperparameter menggunakan RandomizedSearchCV untuk tiga model regresi tersebut, yang menunjukkan peningkatan stabilitas dan performa pada model Decision Tree dan k-NN. Meskipun ada perbaikan setelah tuning, XGBoost tetap menjadi pilihan terbaik karena prediksinya yang lebih halus dan kestabilan yang lebih baik, dengan MSE yang lebih rendah dibandingkan dengan model lainnya.

Secara keseluruhan, XGBoost menunjukkan performa terbaik, baik sebelum maupun setelah tuning, sehingga model ini disarankan untuk digunakan dalam analisis dataset ini.

Classification Model

1. Memuat Data dan Visualisasi Awal

Kode dimulai dengan memuat data dari file CSV yang disimpan di Google Drive, lalu data tersebut divisualisasikan menggunakan boxplot dan scatter plot. Visualisasi ini bertujuan untuk memberikan pemahaman awal tentang distribusi data dan hubungan antar variabel. Boxplot menunjukkan bahwa variabel X memiliki distribusi yang lebih lebar dan beberapa outliers, sedangkan variabel Y memiliki distribusi yang lebih simetris dengan sejumlah outliers yang tersebar di atas dan di bawah median. Scatter plot memperlihatkan hubungan yang lemah antara X dan Y, namun menunjukkan tren peningkatan Y terhadap X secara umum, meskipun terdapat variasi dan outliers yang cukup signifikan. Dari sini, kita bisa melihat bahwa hubungan antara X dan Y mungkin tidak sepenuhnya linier.

2. Regresi Linear vs. Regresi Polinomial

Langkah berikutnya adalah membandingkan dua model regresi: regresi linear dan regresi polinomial (dengan derajat 3). Model regresi polinomial menunjukkan hasil yang lebih baik dalam menangkap pola data dibandingkan regresi linear, yang lebih sederhana. Hal ini terlihat dari perbedaan nilai Mean Squared Error (MSE) antara kedua model tersebut, di mana regresi polinomial menghasilkan MSE yang sedikit lebih rendah. Namun, penting untuk dicatat bahwa regresi polinomial cenderung lebih kompleks dan berisiko overfitting, sehingga perlu kehati-hatian dalam memilih derajat polinomial yang digunakan. Overfitting dapat terjadi jika model terlalu rumit dan tidak mampu generalisasi dengan baik pada data baru.

3. Penggunaan Model Regresi Lebih Kompleks

Selanjutnya, tiga model regresi yang lebih kompleks diterapkan untuk menganalisis data: Decision Tree Regressor, k-Nearest Neighbors (k-NN) Regressor, dan XGBoost Regressor. Setiap model ini diuji untuk memprediksi Y berdasarkan X, dan visualisasinya menunjukkan hasil yang sangat bervariasi:

- Decision Tree Regressor menghasilkan prediksi berbentuk langkah-langkah (step-wise), yang berarti model ini cenderung "memotong" data pada titik-titik tertentu, mengikuti tren umum tetapi tidak halus dalam pergerakan data. Meskipun Decision Tree dapat menangkap tren data dengan baik, ia lebih cenderung overfitting, terutama pada data dengan noise atau variasi yang kecil.
- k-NN Regressor menunjukkan prediksi yang sangat tidak stabil, dengan fluktuasi tajam pada titik-titik tertentu, terutama sebelum tahun 1960. Ini mengindikasikan bahwa model k-NN sangat sensitif terhadap data lokal dan cenderung mengikuti pola data yang tidak signifikan, yang menyebabkan prediksi menjadi sangat bervariasi dan lebih rentan terhadap overfitting.
- XGBoost Regressor, di sisi lain, memberikan prediksi yang lebih mulus dan stabil, mengikuti tren data dengan lebih baik tanpa banyak fluktuasi. XGBoost, yang merupakan model boosting yang berbasis pohon keputusan, mampu menangkap pola global dan variasi data dengan baik tanpa terlalu sensitif terhadap noise. Oleh karena itu, XGBoost terbukti lebih efektif dalam menangani data dengan pola non-linear dan kompleks.

4. Evaluasi Model Menggunakan MSE

Setelah implementasi model, Mean Squared Error (MSE) digunakan untuk mengevaluasi kinerja setiap model pada data uji. Hasilnya menunjukkan bahwa XGBoost memiliki MSE terendah, diikuti oleh Decision Tree dan k-NN dengan MSE yang lebih tinggi. Ini menegaskan bahwa meskipun Decision Tree memberikan hasil yang mirip dengan XGBoost dalam hal MSE, XGBoost lebih unggul karena memberikan hasil yang lebih stabil dan konsisten, sedangkan Decision Tree lebih cenderung menghasilkan prediksi yang lebih tajam dan kurang mulus.

5. Hyperparameter Tuning Menggunakan RandomizedSearchCV

Untuk meningkatkan kinerja model, dilakukan hyperparameter tuning menggunakan RandomizedSearchCV. Proses ini mencari kombinasi hyperparameter yang optimal untuk masing-masing model dengan mengevaluasi berbagai kombinasi parameter dan memilih yang memberikan kinerja terbaik berdasarkan MSE. Tuning ini menghasilkan peningkatan dalam performa model Decision Tree dan k-NN, meskipun XGBoost tetap menunjukkan performa terbaik setelah tuning. XGBoost memiliki keunggulan dalam hal kestabilan dan ketepatan prediksi, sedangkan k-NN meskipun mengalami peningkatan, tetap memiliki MSE yang lebih tinggi dibandingkan dengan Decision Tree dan XGBoost.

6. Kesimpulan

Secara keseluruhan, setelah dilakukan hyperparameter tuning, XGBoost tetap menjadi model terbaik di antara ketiga model yang diuji. Meskipun Decision Tree dan XGBoost memiliki MSE yang hampir sama, XGBoost lebih unggul karena prediksinya yang lebih halus dan stabil, serta kemampuannya dalam menangkap pola global tanpa terlalu rentan terhadap overfitting. k-NN, meskipun mengalami peningkatan, tetap menunjukkan performa yang kurang optimal dengan MSE yang lebih tinggi, menandakan bahwa model ini kurang cocok untuk dataset ini.

XGBoost dengan tuning hyperparameter adalah pilihan terbaik untuk menangani data ini, karena mampu menangkap kompleksitas pola dan variasi data dengan lebih baik, serta memberikan prediksi yang lebih stabil dan akurat.