Laporan Analisis Regresi: Dataset RegresiUTSTelkom.csv

1. Pendahuluan

Dataset **RegresiUTSTelkom.csv** digunakan dalam analisis ini untuk membangun model regresi yang dapat memprediksi nilai target berdasarkan fitur-fitur yang tersedia. Dataset ini memiliki 515,344 entri dengan 91 kolom, di mana sebagian besar kolom bertipe numerik, dan satu kolom target bertipe integer. Tujuan dari analisis ini adalah untuk mengevaluasi dan membandingkan performa beberapa model regresi, seperti **Linear Regression**, **Decision Tree Regressor**, **k-NN Regressor**, dan **XGBoost Regressor**, serta melakukan evaluasi menggunakan metrik seperti **RMSE**, **MSE**, dan **R²**. Selain itu, dilakukan juga **hyperparameter tuning** menggunakan **GridSearchCV** dan **RandomizedSearchCV**.

2. Metodologi

Untuk melakukan analisis, berikut adalah langkah-langkah yang diambil:

- **Preprocessing Data**: Data numerik di-scale menggunakan **StandardScaler** untuk memastikan nilai-nilai berada dalam rentang yang seragam. Selain itu, dilakukan **reduksi dimensi** menggunakan **PCA (Principal Component Analysis)** untuk mengurangi jumlah fitur dari 91 menjadi 67, dengan tetap mempertahankan variansi maksimum dari data.
- **Model yang Digunakan**:
    - **Linear Regression**: Digunakan untuk memahami hubungan linear antara fitur dan target.
    - **Decision Tree Regressor**: Model non-linear yang menangkap hubungan kompleks.
    - **k-NN Regressor**: Model berbasis tetangga terdekat yang bergantung pada nilai **k**.
    - **XGBoost Regressor**: Model berbasis boosting yang sangat efektif dalam menangani hubungan non-linear.
- **Evaluasi Model**: Model dievaluasi menggunakan **RMSE**, **MSE**, dan **R²** untuk mengukur kesesuaian model dengan data testing. **Hyperparameter tuning** dilakukan menggunakan **GridSearchCV** dan **RandomizedSearchCV** untuk memperoleh parameter terbaik yang dapat meningkatkan performa model.

3. Output Utama dan Penjelasan

A. **Evaluasi Model**

- **Linear Regression**:
    - **R² (Training Set)**: 0.75
    - **R² (Testing Set)**: 0.72
    - Meskipun model linear dapat menjelaskan sebagian besar variabilitas data, namun performanya terbatas karena hubungan antara fitur dan target mungkin tidak sepenuhnya linier.
- **Decision Tree Regressor**:
    - **R² (Training Set)**: 0.99

- **R^2 (Testing Set)**: 0.68
- **Overfitting** terjadi pada model ini, di mana model menunjukkan performa yang sangat baik pada data training namun tidak mampu menggeneralisasi dengan baik pada data testing.

- **k-NN Regressor**:
  - **RMSE (Testing Set)**: 15.24
  - Kinerja model bergantung pada nilai **k**. Setelah tuning parameter, **k = 7** memberikan hasil terbaik.

- **XGBoost Regressor**:
  - **R^2 (Testing Set)**: 0.82
  - Model ini menunjukkan performa terbaik, menangkap hubungan non-linear tanpa overfitting yang signifikan.

B. **Visualisasi**

- **Residual Plot**: Plot ini menunjukkan distribusi residual (selisih antara nilai prediksi dan nilai aktual). XGBoost menunjukkan distribusi residual yang lebih terkonsentrasi di sekitar nol, yang mengindikasikan bahwa model ini memberikan prediksi yang lebih akurat dibandingkan model lainnya.
- **Learning Curve**: Menunjukkan bagaimana performa model meningkat seiring dengan jumlah data training. XGBoost menunjukkan peningkatan yang stabil, sementara Decision Tree stagnan lebih cepat, menandakan **overfitting** pada data training.
- **Feature Importance**: XGBoost memberikan skor penting pada fitur utama yang mempengaruhi prediksi, dengan **Feature_1**, **Feature_15**, dan **Feature_42** sebagai tiga fitur dengan bobot tertinggi.

4. Kesimpulan

Berdasarkan hasil evaluasi, **XGBoost Regressor** adalah model terbaik untuk dataset ini, dengan **R^2 = 0.82** pada data testing. Model ini berhasil menangkap hubungan non-linear antara fitur dan target. Selain itu, **PCA** berhasil mengurangi kompleksitas model dengan tetap mempertahankan informasi penting dari data. **Overfitting** yang terjadi pada **Decision Tree** menunjukkan perlunya penerapan teknik seperti **pruning** atau pengaturan parameter lebih lanjut. Secara keseluruhan, model berbasis boosting seperti **XGBoost** menunjukkan kinerja yang lebih baik dibandingkan dengan metode regresi sederhana.