Laporan Analisis dan Pemodelan Regresi

Nama : Satya Athaya Daniswara

NIM : 1103213152

1. Pendahuluan

Dataset yang digunakan dalam analisis ini bertujuan untuk memprediksi hubungan antara variabel `Year` dan sejumlah fitur lainnya. Dataset terdiri dari fitur numerik yang diolah melalui tahapan eksplorasi data, visualisasi, preprocessing, pemodelan, dan evaluasi. Fokus utama adalah membangun model regresi yang dapat memprediksi dengan akurasi tinggi serta memahami hubungan antar variabel.

2. Eksplorasi Data

- Inisialisasi Dataset :Dataset dimuat menggunakan `pandas` dan diperiksa struktur awalnya. Kolom tanpa header dinamai ulang dengan format `x1`, `x2`, dll., untuk memberikan kejelasan.

-Pemeriksaan Nilai Kosong:

-Hasil: Beberapa kolom ditemukan memiliki nilai kosong (`NaN`).

- Implikasi:Penanganan data kosong akan dilakukan untuk menghindari error pada pemodelan.

- Statistik Deskriptif: Distribusi fitur menunjukkan variasi antar kolom, khususnya pada variabel `Year`.

3. Visualisasi Data

- Distribusi Tahun: Grafik batang horizontal menunjukkan bahwa beberapa tahun memiliki jumlah instance yang signifikan, mengindikasikan ketidakseimbangan distribusi temporal.

- Heatmap Korelasi:Korelasi antar fitur divisualisasikan untuk mengidentifikasi hubungan signifikan yang dapat digunakan dalam pemodelan.

4. Preprocessing Data

- Transformasi Data:

  - Nilai kosong diimputasi dan fitur dinormalisasi menggunakan `StandardScaler`.

  - Pemisahan data dilakukan dengan proporsi 80% data latih dan 20% data uji untuk validasi.

-Pemilihan Fitur: Fitur dengan korelasi kuat terhadap `Year` dipilih untuk meningkatkan efisiensi model.

5. Pemodelan

Empat model regresi diterapkan menggunakan pipeline preprocessing:

1. Polynomial Regression:

- Parameter terbaik: Derajat polinomial = 2.
- Keunggulan: Menangkap hubungan non-linear antar fitur.

2. Decision Tree

- Parameter terbaik: `max_depth=5`, `min_samples_split=2`.
- Keunggulan:Interpretabilitas tinggi, cocok untuk hubungan non-linear.

3. k-Nearest Neighbors (k-NN)

- Parameter terbaik: `n_neighbors=7`, `weights='distance'`.
- Keunggulan: Menggunakan tetangga terdekat untuk prediksi.

4. XGBoost:

- Parameter terbaik: `n_estimators=100`, `learning_rate=0.1`.
- Keunggulan Algoritma ensemble dengan kinerja tinggi dan kemampuan menangani data besar.

6. Evaluasi

-Hasil Akurasi:

- Semua model memberikan performa yang kompetitif dengan skor evaluasi yang serupa (~90% untuk `r2`).

-Matriks Korelasi: Variabel `Year` memiliki korelasi signifikan dengan beberapa fitur seperti `x15`, `x19`, dan `x23`.

- Fitur Penting:`x15` dan `x23` adalah fitur paling signifikan berdasarkan evaluasi korelasi dan hasil model.

- Korelasi Fitur: Hubungan antara `Year` dan fitur numerik tertentu menunjukkan pola signifikan yang relevan untuk analisis prediktif.

- Model Efektif: XGBoost memiliki keunggulan dalam menangani data besar, tetapi Polynomial Regression lebih sederhana untuk implementasi.

- Ketidakseimbangan Data Distribusi tahun yang tidak merata memengaruhi pemodelan dan membutuhkan perhatian lebih dalam pengolahan data.

8. Kesimpulan

- Polynomial Regression adalah model terbaik dalam studi ini karena dapat menangkap hubungan non-linear secara efektif.

- XGBoost memberikan alternatif dengan kinerja tinggi namun membutuhkan lebih banyak sumber daya komputasi.

- Analisis ini memberikan wawasan untuk pengembangan model prediktif serta pentingnya pemilihan fitur yang relevan dalam meningkatkan akurasi model.