PROJEK 2
LINEAR DAN POLYNOMIAL REGRESSION

disusun untuk memenuhi tugas mata
kuliah *Pembelajaran Mesin*

oleh:

Arif Maulana	(2208107010067)
M. Nouval Rifqi	(2208107010075)
Azzariyat Azra	(2208107010079)
Tiara Agustin	(2208107010004)

JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA
2025

1) Pemahaman Dataset

a) Sumber Data

Dataset yang digunakan dalam analisis ini adalah *Crop Yield Prediction Dataset* yang tersedia di platform Kaggle. Dataset ini dirancang untuk membantu dalam memprediksi hasil panen berdasarkan berbagai faktor agronomis dan lingkungan.

Variabel yang Digunakan

Dataset "Crop Yield Prediction Dataset" yang disediakan oleh Mrigaank Jaswal mencakup beberapa variabel utama yang digunakan untuk menganalisis dan memprediksi hasil panen. Variabel-variabel tersebut meliputi:

1. **Area (Negara):** Menunjukkan negara atau wilayah tempat data pertanian dikumpulkan, memungkinkan analisis lintas wilayah dan perbandingan tren global.

2. **Item (Jenis Tanaman):** Menentukan jenis produk pertanian, termasuk berbagai macam tanaman seperti biji-bijian, buah-buahan, dan sayuran.ResearchGate

3. **Year (Tahun):** Mewakili tahun di mana data dicatat, mendukung analisis tren historis dan membantu dalam pemodelan efek jangka panjang dari perubahan iklim atau kebijakan terhadap hasil panen.

4. **hg/ha_yield (Hasil per Hektar):** Menunjukkan hasil panen yang diukur dalam hektogram per hektar, mencerminkan efisiensi produksi pertanian.

5. **average_rain_fall_mm_per_year (Curah Hujan Tahunan Rata-rata):** Melacak rata-rata curah hujan tahunan dalam milimeter untuk setiap negara, faktor penting untuk kesehatan tanaman, kondisi tanah, dan strategi pengelolaan air.

6. **pesticides_tonnes (Penggunaan Pestisida):** Menangkap jumlah pestisida yang diterapkan, diukur dalam ton, memungkinkan penilaian hubungan antara penggunaan pestisida dan hasil panen, serta dampak lingkungan dari praktik pertanian.

7. **avg_temp (Suhu Rata-rata):** Mencatat suhu rata-rata tahunan, faktor penting yang mempengaruhi pertumbuhan tanaman dan hasil panen.

b) Statistik Deskriptif dan Visualisasi Awal Data

```
print("\nStatistik deskriptif:")
print(df.describe())
```

```
Statistik deskriptif:
                Year    hg/ha_yield  average_rain_fall_mm_per_year
count  28242.000000   28242.000000                    28242.00000
mean    2001.544296   77053.332094                     1149.05598
std        7.051905   84956.612897                      709.81215
min     1990.000000      50.000000                       51.00000
25%     1995.000000   19919.250000                      593.00000
50%     2001.000000   38295.000000                     1083.00000
75%     2008.000000  104676.750000                     1668.00000
max     2013.000000  501412.000000                     3240.00000

       pesticides_tonnes      avg_temp
count       28242.000000  28242.000000
mean        37076.909344     20.542627
std         59958.784665      6.312051
min             0.040000      1.300000
25%          1702.000000     16.702500
50%         17529.440000     21.510000
75%         48687.880000     26.000000
max        367778.000000     30.650000
```

Gambar 1.1 Statistik Deskriptif

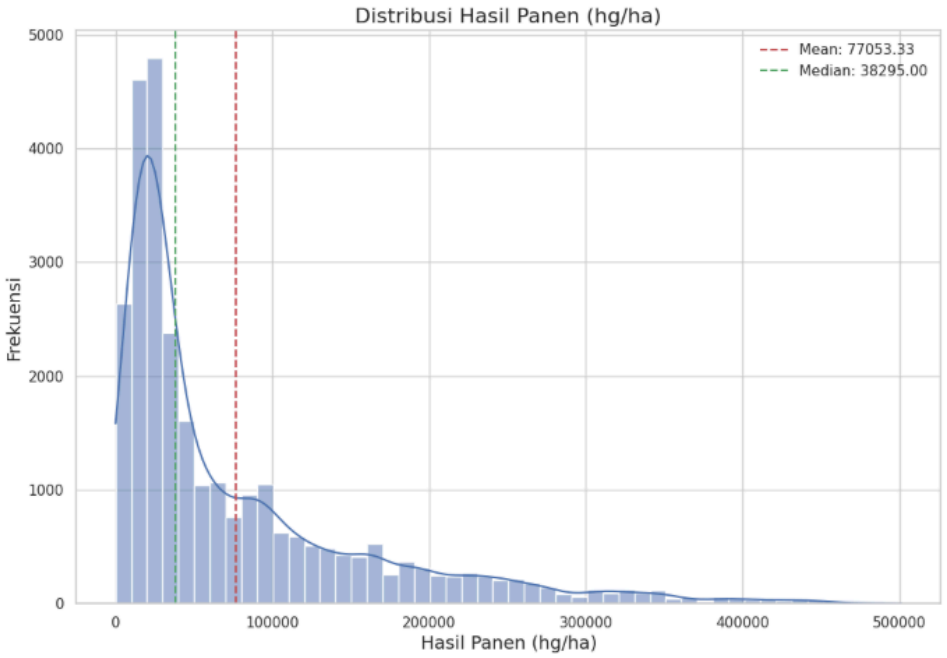
Analisis awal dilakukan dengan menghitung statistik deskriptif pada dataset untuk memahami distribusi data pada setiap variabel. Statistik deskriptif yang ditampilkan meliputi:

- **Jumlah data (count):** Menunjukkan jumlah observasi dalam dataset.
- **Mean (rata-rata):** Menunjukkan nilai rata-rata dari setiap variabel.
- **Standar deviasi (std):** Mengukur seberapa besar variasi data dari nilai rata-rata.
- **Min (minimum):** Nilai terkecil dalam dataset.
- **Max (maksimum):** Nilai terbesar dalam dataset.
- **Kuartil (25%, 50%, 75%):** Nilai-nilai yang menunjukkan distribusi data dalam persentil tertentu.

Dari hasil perhitungan statistik deskriptif, dapat diperoleh beberapa wawasan awal mengenai dataset:

- **Tahun (Year):** Data mencakup rentang tahun dari 1990 hingga 2013.
- **Hasil panen per hektar (hg/ha_yield):** Memiliki rata-rata sekitar 77.053 kg per hektar dengan minimum 50 kg dan maksimum mencapai 501.412 kg.
- **Curah hujan rata-rata per tahun (average_rain_fall_mm_per_year):** Berkisar antara 51 mm hingga 3.240 mm, dengan rata-rata 1.149 mm.
- **Penggunaan pestisida (pesticides_tonnes):** Memiliki variasi besar, dengan rata-rata sekitar 37.076 ton, tetapi nilai maksimumnya mencapai 367.778 ton.
- **Suhu rata-rata (avg_temp):** Suhu berkisar antara 1,3°C hingga 30,65°C, dengan rata-rata sekitar 20,54°C.

Selanjutnya, akan dilakukan visualisasi untuk memahami distribusi dan hubungan antara variabel dalam dataset.



Gambar 1.2 Histogram Hasil Panen

Histogram di atas menunjukkan distribusi data dari variabel target *hg/ha_yield*, yaitu hasil panen per hektar. Beberapa wawasan penting yang dapat diambil dari visualisasi ini:

- **Distribusi tidak normal:** Data menunjukkan kecenderungan *right-skewed* (condong ke kanan), yang berarti terdapat beberapa nilai hasil panen yang jauh lebih tinggi dibandingkan sebagian besar data lainnya.

- **Mayoritas hasil panen terpusat pada nilai rendah:** Sebagian besar sampel dalam dataset memiliki hasil panen kurang dari **100.000 hg/ha**, dengan puncak distribusi di sekitar **nilai yang lebih kecil dari median (38.295 hg/ha)**.
- **Outlier:** Nilai maksimum mencapai lebih dari **400.000 hg/ha**, yang cukup jauh dari median, menunjukkan adanya data ekstrem yang perlu dianalisis lebih lanjut.
- **Perbandingan Mean dan Median:**
    - **Mean (rata-rata):** 77.053,33 hg/ha (ditunjukkan dengan garis merah putus-putus).
    - **Median:** 38.295,00 hg/ha (ditunjukkan dengan garis hijau putus-putus). Perbedaan yang cukup signifikan antara mean dan median semakin menegaskan bahwa distribusi data ini memiliki ekor panjang di sisi kanan.

Dari grafik ini dapat disimpulkan bahwa distribusi hasil panen cenderung condong ke kanan (*right-skewed*) karena:

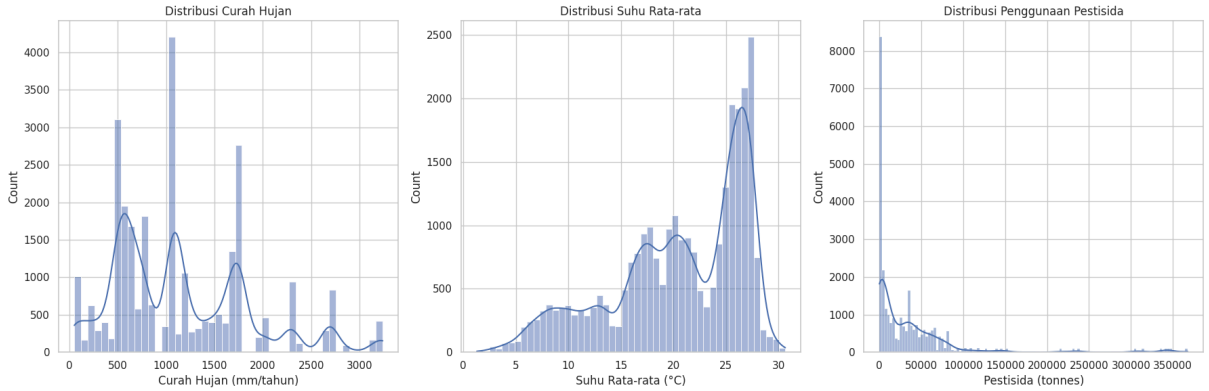
- Mayoritas data terkumpul di rentang nilai yang lebih rendah (**0-100.000 hg/ha**).
- Nilai **mean (77.053,33 hg/ha) lebih besar dari median (38.295,00 hg/ha)**, menunjukkan adanya data ekstrem yang mempengaruhi rata-rata.
- Terdapat beberapa **nilai ekstrem di sebelah kanan**, yang menarik kurva distribusi ke arah kanan.
- **Frekuensi tertinggi berada di rentang sekitar 20.000-40.000 hg/ha**, menunjukkan bahwa sebagian besar pengamatan memiliki hasil panen dalam rentang tersebut.

Dari analisis ini, dapat disimpulkan bahwa data hasil panen memiliki distribusi yang tidak merata dan terdapat kemungkinan perlunya transformasi data sebelum diterapkan ke dalam model regresi.

2) Eksplorasi Data dan Pra-pemrosesan

Pada tahap ini, dilakukan analisis lebih lanjut terhadap dataset untuk memahami struktur data, mengidentifikasi pola, serta menangani potensi permasalahan seperti data yang hilang, duplikasi, outlier, dan skala variabel. Langkah-langkah eksplorasi meliputi pemeriksaan nilai yang hilang dan duplikasi, analisis korelasi antar variabel, deteksi serta penanganan outlier, serta normalisasi atau standarisasi data jika diperlukan. Proses ini bertujuan untuk memastikan bahwa data dalam kondisi optimal sebelum digunakan dalam pemodelan regresi, sehingga dapat meningkatkan akurasi dan interpretasi hasil analisis.

- Hipotesis pendekatan segmentasi dalam pemodelan hasil panen



Gambar 2.1

Dalam analisis awal terhadap dataset hasil panen, kami menemukan tantangan signifikan berupa **korelasi yang rendah** antara variabel target (hasil panen) dan prediktor lingkungan, yaitu:

- Korelasi **curah hujan** dengan hasil panen: ≈ **0.00**
- Korelasi **suhu rata-rata** dengan hasil panen: ≈ **-0.11**
- Korelasi **penggunaan pestisida** dengan hasil panen: ≈ **0.06**

Rendahnya korelasi ini menunjukkan bahwa model prediktif seperti **regresi linear** dan **regresi polinomial**, yang sangat bergantung pada kekuatan hubungan antar variabel, akan

mengalami kesulitan dalam menghasilkan prediksi yang akurat jika diterapkan langsung pada dataset global tanpa penyaringan.

Hipotesis Segmentasi

Untuk mengatasi hal ini, kami mengajukan **hipotesis segmentasi**, yakni bahwa rendahnya korelasi ini disebabkan oleh **tingginya heterogenitas** dalam dataset secara global. Oleh karena itu, kami mengeksplorasi dua pendekatan segmentasi untuk meningkatkan kualitas pemodelan:

1. Segmentasi berdasarkan Negara
   Setiap negara memiliki karakteristik unik seperti iklim, praktik pertanian, dan geografi. Dengan menganalisis data dari satu negara secara terpisah (misalnya Albania), hubungan antara variabel lingkungan dan hasil panen diharapkan menjadi lebih jelas dan kuat.

2. Segmentasi berdasarkan Jenis Tanaman
   Tiap tanaman memiliki kebutuhan lingkungan dan toleransi yang berbeda. Misalnya, tanaman kentang mungkin lebih sensitif terhadap suhu dibandingkan tanaman sereal. Oleh karena itu, dengan memfokuskan analisis pada satu jenis tanaman (misalnya Kentang), pola-pola spesifik yang sebelumnya tersamarkan bisa diidentifikasi dengan lebih akurat.

Justifikasi Pendekatan Segmentasi

Pendekatan ini didukung oleh beberapa alasan logis:

- **Homogenitas Ekologis:**
  Dalam satu negara, kondisi tanah dan iklim relatif seragam sehingga mengurangi variasi acak yang tidak relevan dengan hasil panen.

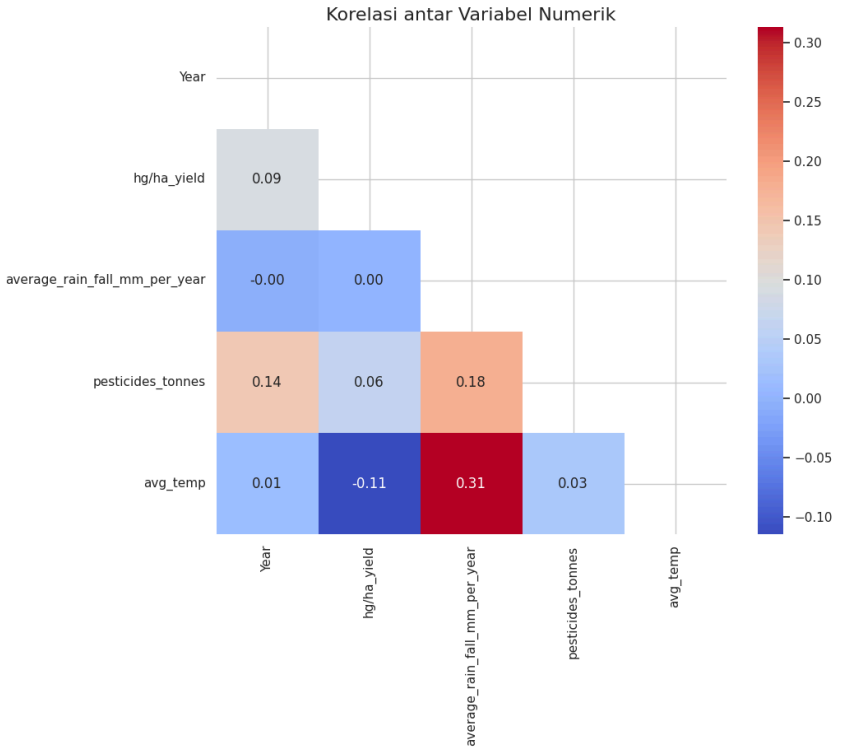
- **Keseragaman Praktik Pertanian:**
  Petani di wilayah geografis yang sama cenderung menerapkan teknik bercocok tanam yang serupa, mengurangi perbedaan dalam faktor manusia.

- **Spesifikasi Biologis Tanaman:**
  Setiap tanaman memiliki rentang optimal untuk suhu, curah hujan, dan nutrisi. Segmentasi per tanaman memungkinkan model fokus pada parameter yang relevan.

- **Kesesuaian dengan Model Regresi:**
  Regresi linear dan polinomial memerlukan adanya hubungan yang cukup kuat antar variabel. Segmentasi membantu memperkuat hubungan tersebut dalam subset data yang lebih homogen dan memungkinkan pengenalan hubungan non-linear yang lebih nyata.

- Dataset Penuh

s



Gambar 2.6 Korelasi antar variabel numerik

Secara umum, hubungan antara variabel target (`hg/ha_yield` atau hasil panen) dan variabel prediktor seperti **curah hujan**, **suhu rata-rata**, serta **penggunaan pestisida** tergolong **lemah**, yang mengindikasikan bahwa hubungan antar variabel ini kemungkinan **tidak bersifat linear sepenuhnya**.

Beberapa poin penting dari hasil analisis korelasi:

- **Korelasi Terkuat: Suhu dan Penggunaan Pestisida (0.31)**

  Ini menunjukkan bahwa daerah dengan suhu rata-rata yang lebih tinggi cenderung menggunakan lebih banyak pestisida, yang mungkin disebabkan oleh peningkatan hama/penyakit pada kondisi yang lebih hangat.

- **Suhu dan Hasil Panen (-0.11)**

  Terdapat korelasi negatif lemah antara suhu dan hasil panen. Artinya, peningkatan suhu sedikit banyak mungkin berkontribusi pada penurunan produktivitas tanaman secara umum.

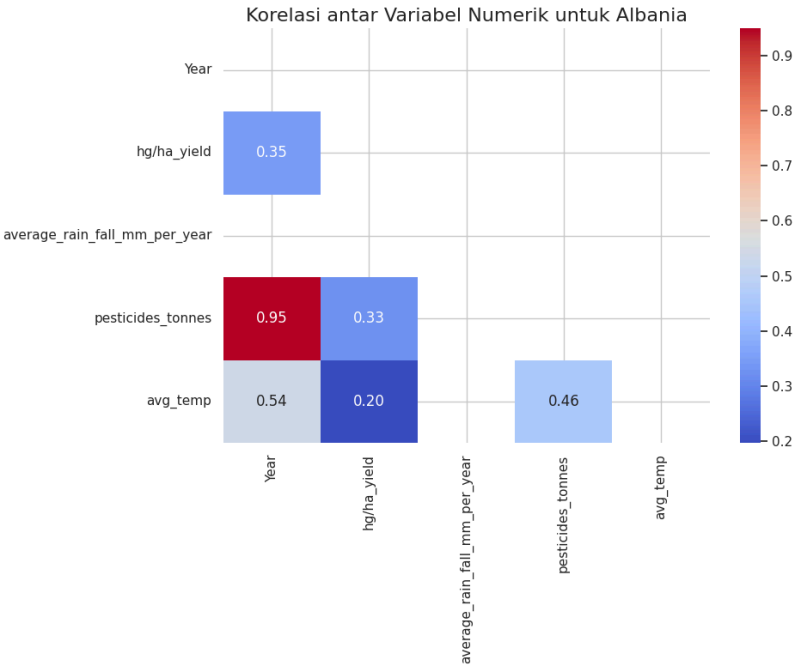
- **Curah Hujan dan Hasil Panen (0.00)**

  Tidak ditemukan hubungan yang berarti antara curah hujan dan hasil panen. Korelasi ini hampir nol, menandakan bahwa variabel ini mungkin memiliki pengaruh tidak langsung atau interaksi kompleks dengan variabel lain.

- **Pestisida dan Hasil Panen (0.06)**

  Korelasi yang sangat lemah, menandakan bahwa peningkatan penggunaan pestisida tidak serta-merta berhubungan dengan hasil panen yang lebih tinggi.

- EDA Segmented (by Country)



Gambar 2.2

Berdasarkan analisis korelasi antar variabel numerik di Albania, ditemukan beberapa pola menarik:

- **Korelasi Sangat Kuat antara Tahun dan Penggunaan Pestisida (0.95)**
  Terdapat hubungan positif yang sangat kuat antara tahun dan jumlah penggunaan pestisida. Hal ini menunjukkan bahwa penggunaan pestisida di Albania meningkat secara signifikan dari waktu ke waktu.

- **Korelasi Moderat antara Tahun dan Suhu Rata-Rata (0.54)**
  Terdapat tren peningkatan suhu rata-rata tahunan di Albania, yang tercermin dari korelasi moderat antara tahun dan suhu.

- **Korelasi Lemah hingga Moderat antara Hasil Panen (Yield) dan Faktor-Faktor Lingkungan**

    - Hasil panen dengan tahun: korelasi positif lemah (0.35)

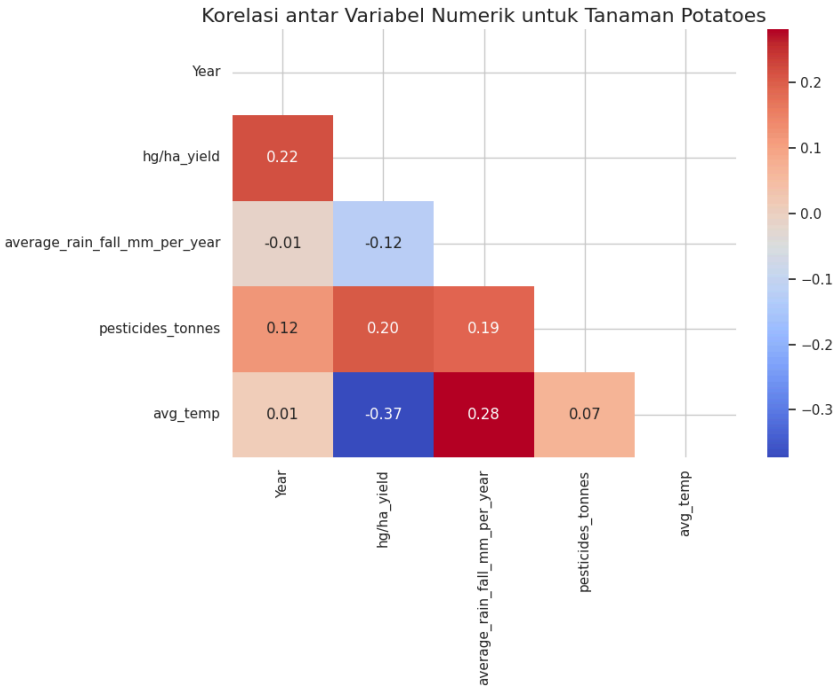
    - Hasil panen dengan pestisida: korelasi positif lemah (0.33)

    - Hasil panen dengan suhu rata-rata: korelasi sangat lemah (0.20)

- Korelasi-korelasi ini menunjukkan adanya potensi hubungan, namun belum cukup kuat untuk langsung diterapkan dalam model tanpa pertimbangan tambahan.

- **Data Curah Hujan Tidak Bervariasi**
  Variabel curah hujan (`average_rain_fall_mm_per_year`) memiliki nilai yang konstan (standar deviasi = 0), sehingga tidak bisa dihitung korelasinya. Hal ini menunjukkan tidak adanya variasi data curah hujan untuk Albania dalam dataset yang digunakan.

- EDA Segmented (by Item)



Gambar 2.3

Berdasarkan analisis korelasi untuk tanaman kentang, ditemukan beberapa pola hubungan antar variabel numerik:

- **Korelasi Negatif antara Suhu dan Hasil Panen (-0.37)**

  Terdapat hubungan negatif yang cukup signifikan antara suhu rata-rata dan hasil panen kentang. Ini mengindikasikan bahwa suhu yang lebih tinggi cenderung berdampak buruk terhadap produktivitas kentang, yang kemungkinan besar lebih cocok tumbuh di iklim yang lebih sejuk.

- **Korelasi Positif antara Suhu dan Curah Hujan (0.28)**

  Daerah atau waktu dengan suhu yang lebih tinggi juga cenderung mengalami curah hujan yang lebih tinggi, berdasarkan data dalam dataset ini.

- **Korelasi Positif antara Pestisida dan Hasil Panen (0.20)**

  Ditemukan korelasi positif lemah antara jumlah penggunaan pestisida dan hasil panen kentang, yang mungkin menunjukkan bahwa penggunaan pestisida membantu meningkatkan hasil panen, meskipun tidak terlalu dominan.

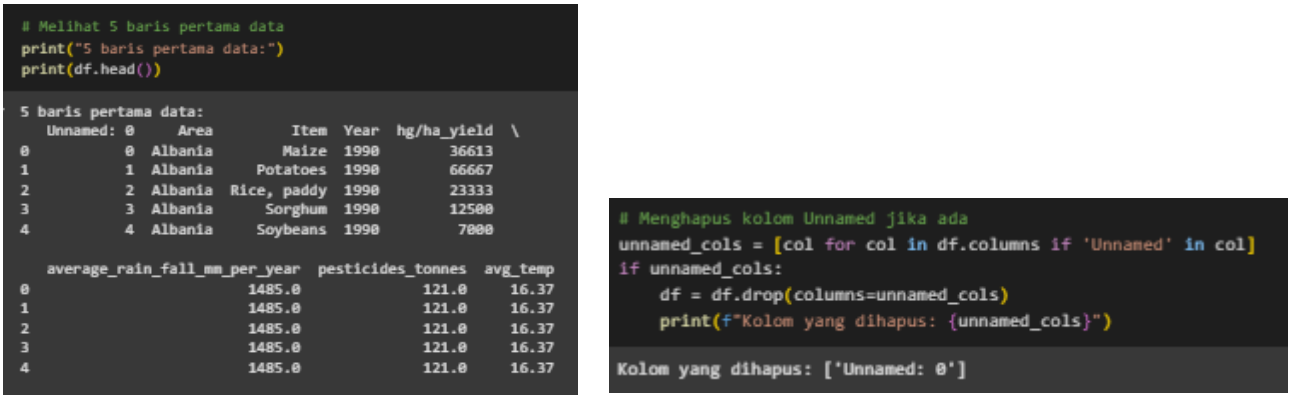
- **Korelasi Negatif antara Curah Hujan dan Hasil Panen (-0.12)**

  Terdapat korelasi negatif lemah antara curah hujan dan hasil panen kentang. Ini bisa berarti bahwa curah hujan berlebih mungkin kurang menguntungkan bagi tanaman kentang.

- **Pengaruh Tahun terhadap Hasil Panen (0.22)**

  Korelasi positif lemah antara tahun dan hasil panen menunjukkan adanya peningkatan produktivitas dari waktu ke waktu, kemungkinan dipengaruhi oleh perkembangan teknologi pertanian atau praktik budidaya yang semakin baik.

- Menghapus kolom Unnamed



```
# Melihat 5 baris pertama data
print("5 baris pertama data:")
print(df.head())
```

```
5 baris pertama data:
   Unnamed: 0     Area         Item  Year  hg/ha_yield  \
0           0  Albania        Maize  1990        36613
1           1  Albania     Potatoes  1990        66667
2           2  Albania  Rice, paddy  1990        23333
3           3  Albania      Sorghum  1990        12500
4           4  Albania     Soybeans  1990         7000

   average_rain_fall_mm_per_year  pesticides_tonnes  avg_temp
0                         1485.0              121.0     16.37
1                         1485.0              121.0     16.37
2                         1485.0              121.0     16.37
3                         1485.0              121.0     16.37
4                         1485.0              121.0     16.37
```

```
# Menghapus kolom Unnamed jika ada
unnamed_cols = [col for col in df.columns if 'Unnamed' in col]
if unnamed_cols:
    df = df.drop(columns=unnamed_cols)
    print(f"Kolom yang dihapus: {unnamed_cols}")
```

```
Kolom yang dihapus: ['Unnamed: 0']
```

Gambar 2.1

Kode yang ditampilkan digunakan untuk eksplorasi awal dataset. Pertama, perintah df.head() menampilkan lima baris pertama dari dataset, yang berisi informasi seperti wilayah (Area), jenis tanaman (Item), tahun pengamatan (Year), hasil panen (hg/ha_yield), curah hujan rata-rata (average_rain_fall_mm_per_year), jumlah penggunaan pestisida (pesticides_tonnes), dan suhu rata-rata (avg_temp). Selain itu, terlihat adanya kolom **"Unnamed: 0"**, yang kemungkinan berasal dari indeks tambahan saat membaca data dari file.

Untuk membersihkan dataset, kode berikutnya mencari kolom yang mengandung kata "Unnamed" dan menghapusnya jika ditemukan. Dalam hal ini, kolom **"Unnamed: 0"** dihapus karena tidak memiliki nilai yang relevan untuk analisis. Langkah ini membantu memastikan bahwa dataset lebih bersih dan hanya terdiri dari informasi yang benar-benar dibutuhkan untuk eksplorasi lebih lanjut serta analisis regresi yang akan dilakukan.

- Cek Missing Values

```
# Cek missing values
print("\nJumlah missing values per kolom:")
print(df.isnull().sum())
```

```
Jumlah missing values per kolom:
Area                              0
Item                              0
Year                              0
hg/ha_yield                       0
average_rain_fall_mm_per_year     0
pesticides_tonnes                 0
avg_temp                          0
dtype: int64
```

Gambar 2.2 Cek Missing Values

Kode ini digunakan untuk memeriksa apakah terdapat nilai yang hilang (missing values) dalam dataset. Perintah df.isnull().sum() menghitung jumlah nilai yang hilang di setiap kolom. Hasil yang ditampilkan menunjukkan bahwa tidak ada nilai yang hilang dalam dataset, karena semua kolom memiliki nilai nol dalam jumlah missing values.

Hal ini berarti dataset dalam kondisi yang bersih, sehingga tidak perlu dilakukan imputasi atau penghapusan data akibat nilai yang hilang. Langkah ini penting dalam eksplorasi data karena keberadaan missing values dapat mempengaruhi analisis dan hasil model prediksi nantinya.

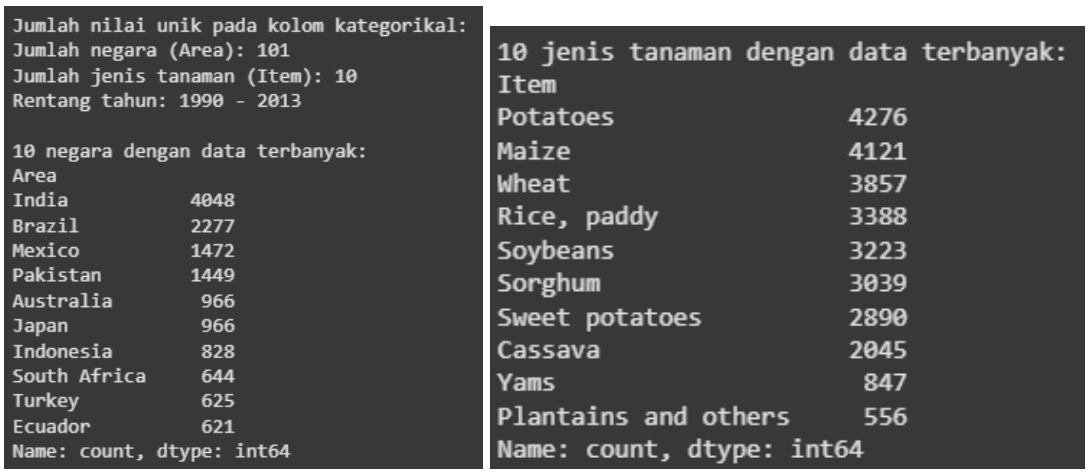
- Cek jumlah nilai unik pada kolom kategorikal

```
# Cek jumlah nilai unik pada kolom kategorikal
print("Jumlah nilai unik pada kolom kategorikal:")
print(f"Jumlah negara (Area): {df['Area'].nunique()}")S
print(f"Jumlah jenis tanaman (Item): {df['Item'].nunique()}")
print(f"Rentang tahun: {df['Year'].min()} - {df['Year'].max()}")

# Menampilkan 10 negara dengan data terbanyak
print("\n10 negara dengan data terbanyak:")
print(df['Area'].value_counts().head(10))

# Menampilkan 10 jenis tanaman dengan data terbanyak
print("\n10 jenis tanaman dengan data terbanyak:")
print(df['Item'].value_counts().head(10))
```

Gambar 2.4



Gambar 2.5

Berdasarkan eksplorasi data ini, terdapat beberapa temuan penting yang dapat diinterpretasikan lebih lanjut.

Pertama, jumlah negara yang tercatat dalam dataset adalah 101, yang menunjukkan bahwa data ini cukup luas dan mencakup berbagai wilayah di dunia. Dengan rentang tahun dari 1990 hingga 2013, dataset ini mencerminkan tren hasil panen selama lebih dari dua dekade, sehingga dapat digunakan untuk menganalisis perubahan dalam produktivitas pertanian seiring waktu.

Kedua, India memiliki jumlah entri data terbanyak, diikuti oleh Brasil dan Meksiko. Ini mengindikasikan bahwa negara-negara ini memiliki data pertanian yang lebih terdokumentasi, kemungkinan karena skala produksi pertanian yang besar atau pencatatan yang lebih aktif. Sebaliknya, negara-negara dengan jumlah entri lebih sedikit mungkin memiliki keterbatasan dalam dokumentasi atau kontribusi yang lebih kecil dalam sektor pertanian global.

Ketiga, dari segi tanaman, kentang menjadi jenis tanaman dengan data terbanyak, disusul oleh jagung dan gandum. Ini menunjukkan bahwa tanaman-tanaman ini memiliki signifikansi tinggi dalam pertanian global, baik dari segi produksi maupun konsumsi. Kehadiran tanaman seperti padi dan kedelai juga mencerminkan pentingnya komoditas pangan utama di berbagai wilayah.

Secara keseluruhan, eksplorasi ini memberikan gambaran awal mengenai distribusi data dan tren dalam sektor pertanian. Informasi ini dapat menjadi dasar untuk analisis lebih lanjut, misalnya melihat pola produksi dari waktu ke waktu, faktor-faktor yang memengaruhi hasil panen, atau tren pertanian di berbagai negara.

- Feature Selection dan Normalisasi

```python
from sklearn.preprocessing import StandardScaler

# Pilih fitur yang digunakan berdasarkan EDA
features = ['average_rain_fall_mm_per_year', 'pesticides_tonnes', 'avg_temp']
categorical_features = ['Area', 'Item']
target = 'log_yield'  # gunakan target yang sudah ditransformasi

# Standarisasi fitur numerik
scaler = StandardScaler()
df_scaled = df.copy()
df_scaled[features] = scaler.fit_transform(df[features])

# Tampilkan perbandingan sebelum dan sesudah normalisasi
print("Sebelum normalisasi:")
print(df[features].describe())

print("\nSetelah normalisasi:")
print(df_scaled[features].describe())
```

Gambar 2.6 Normalisasi

```
Sebelum normalisasi:
       average_rain_fall_mm_per_year  pesticides_tonnes      avg_temp
count                   28242.00000       28242.000000  28242.000000
mean                     1149.05598       37076.909344     20.542627
std                       709.81215       59958.784665      6.312051
min                        51.00000           0.040000      1.300000
25%                       593.00000        1702.000000     16.702500
50%                      1083.00000       17529.440000     21.510000
75%                      1668.00000       48687.880000     26.000000
max                      3240.00000      367778.000000     30.650000

Setelah normalisasi:
       average_rain_fall_mm_per_year  pesticides_tonnes      avg_temp
count                   2.824200e+04       2.824200e+04  2.824200e+04
mean                    1.449163e-16       1.408908e-16  4.750034e-16
std                     1.000018e+00       1.000018e+00  1.000018e+00
min                    -1.546994e+00      -6.183835e-01 -3.048608e+00
25%                    -7.833986e-01      -5.899975e-01 -6.083909e-01
50%                    -9.306286e-02      -3.260209e-01  1.532609e-01
75%                     7.311134e-01       1.936526e-01  8.646112e-01
max                     2.945823e+00       5.515571e+00  1.601310e+00
```

Gambar 2.7 Output setelah normalisasi

Normalisasi yang dilakukan menggunakan **StandardScaler** bertujuan untuk menyelaraskan skala fitur numerik dengan mean 0 dan standar deviasi 1.

Dari tabel statistik:

- **Sebelum normalisasi**, nilai rata-rata (*mean*) dari setiap fitur memiliki skala yang berbeda, misalnya *average_rain_fall_mm_per_year* berkisar antara 51 hingga 3240 mm, sedangkan *pesticides_tonnes* memiliki nilai jauh lebih besar.

- **Setelah normalisasi**, rata-rata setiap fitur mendekati 0, dengan standar deviasi mendekati 1. Nilai minimum dan maksimum juga telah disesuaikan ke dalam skala yang lebih seragam.

3) Implementasi Model

Dalam analisis dataset Dampak Faktor Lingkungan terhadap Hasil Panen di Berbagai Negara, kami telah mengimplementasikan dua jenis model untuk memprediksi hasil pertanian yang dilakukan dengan tiga strategi yaitu:

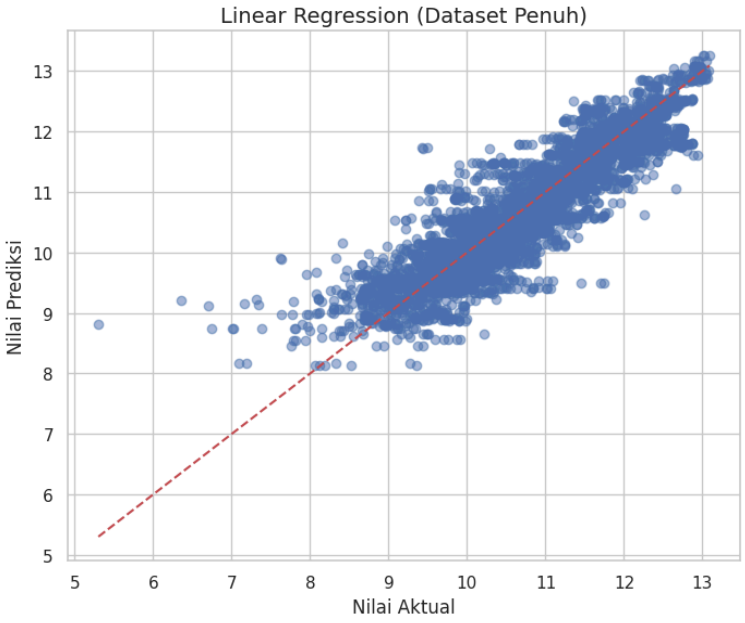
- Model dengan Dataset Penuh - Model dilatih dengan seluruh dataset tanpa ada filter area atau jenis tanaman
- Model Per Negara (Albania) - Model hanya dilatih dengan data dari negara Albania.
- Model Per Tanaman (Kentang) - Model dilatih menggunakan data dari satu jenis tanaman saja, yaitu kentang.

a) Model Linear Regression

Model linear regression digunakan sebagai baseline dengan memetakan hubungan linear antara faktor lingkungan dan hasil panen.

Implementasi pada 3 strategi:

➢ **Strategi 1: Model dengan Dataset Penuh**



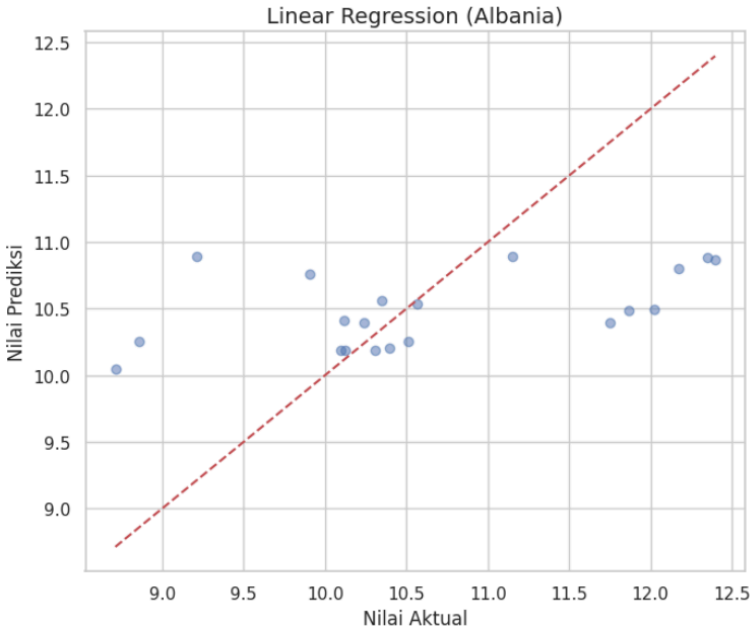
Gambar 3.1 *Linear Regression (Dataset Penuh)*

Evaluasi Model Linear Regression:

- Root Mean Squared Error (RMSE): 0.4578
- Mean Absolute Error (MAE): 0.3477
- Koefisien Determinasi (R^2): 0.8290

Model regresi linear menunjukkan performa yang cukup baik dengan nilai R^2 sebesar 0.8290, yang berarti model mampu menjelaskan sekitar 82.9% variabilitas dalam data. Meskipun masih ada error, model ini dapat menjadi baseline yang cukup baik sebelum mencoba model yang lebih kompleks.

➢ **Strategi 2: Model Per Negara (Contoh: Albania)**



Gambar 3.2 *Linear Regression (Albania)*

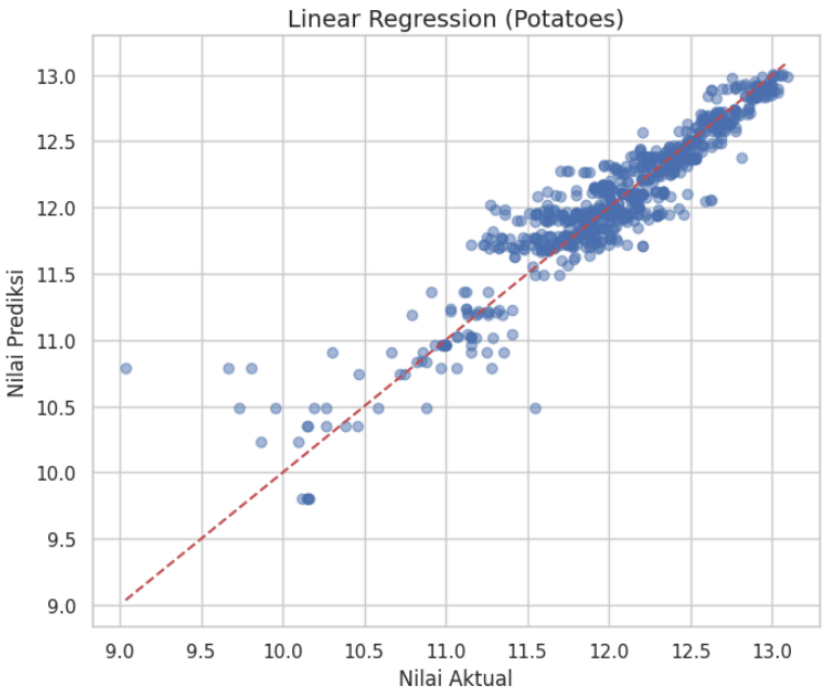
Evaluasi Model Linear Regression:

- Root Mean Squared Error (RMSE): 1.0039
- Mean Absolute Error (MAE): 0.7798
- Koefisien Determinasi (R^2): 0.1624

Hasil evaluasi negara Albania menunjukkan bahwa model linear kurang mampu menjelaskan hubungan antara variabel dalam dataset ini (R^2 hanya 0.1624). Hal ini mengindikasikan bahwa:

1. Data untuk Albania mungkin memiliki pola yang berbeda dari dataset global.
2. Ukuran sampel mungkin terlalu kecil sehingga model tidak bisa belajar dengan baik.
3. Faktor lingkungan di Albania mungkin tidak memiliki hubungan linier dengan hasil panen, sehingga model linear kurang efektif.

➢ **Strategi 3: Model Per Jenis Tanaman (Contoh: Kentang)**



Gambar 3.3 *Linear Regression (Kentang)*

Evaluasi Model Linear Regression:

- Root Mean Squared Error (RMSE): 0.2021
- Mean Absolute Error (MAE): 0.1380
- Koefisien Determinasi (R^2): 0.8761

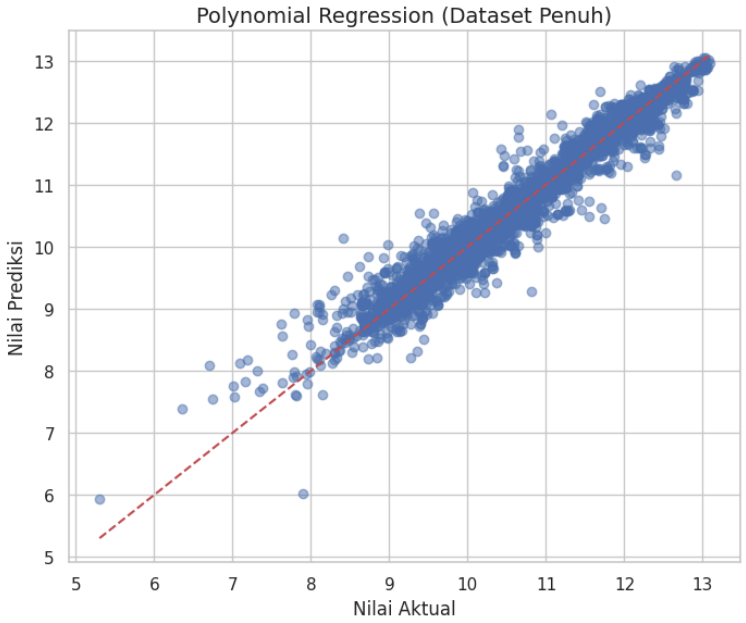
Untuk jenis tanaman kentang, model linear bekerja dengan baik, dengan R^2 sebesar 0.8761, yang berarti sekitar 87.6% variabilitas data dapat dijelaskan oleh model. Ini menunjukkan bahwa hasil panen kentang memiliki hubungan yang relatif linier dengan faktor lingkungan dibandingkan dengan dataset Albania.

b) Model Polynomial Regression

Model polynomial regression dibangun untuk menangkap hubungan non-linear antara fitur dan target. Model ini dibungkus dalam pipeline bersama preprocessing agar proses pelatihan tetap konsisten.

Implementasi pada 3 strategi:

➢ **Strategi 1: Model dengan Dataset Penuh**



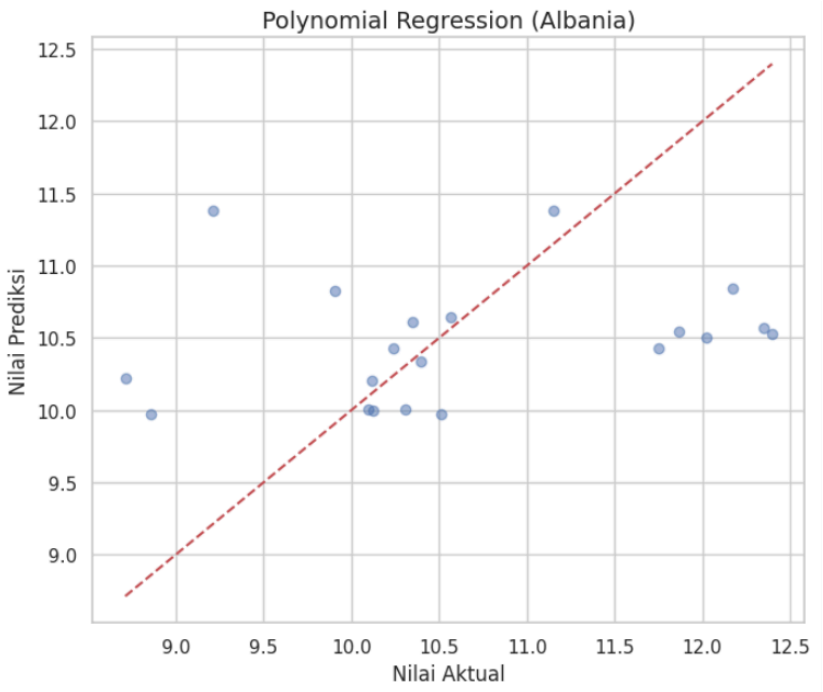
Gambar 3.4 *Polynomial Regression (Dataset Penuh)*

Evaluasi Model Linear Regression:

- Root Mean Squared Error (RMSE): 0.2148
- Mean Absolute Error (MAE): 0.1400
- Koefisien Determinasi (R^2): 0.9623

Regresi polinomial memberikan peningkatan performa yang signifikan dibandingkan regresi linear (R^2 naik dari 0.8290 menjadi 0.9623), yang menunjukkan bahwa hubungan antara faktor lingkungan dan hasil panen memang lebih kompleks daripada hubungan linier.

➢ Strategi 2: Model Per Negara (Contoh: Albania)



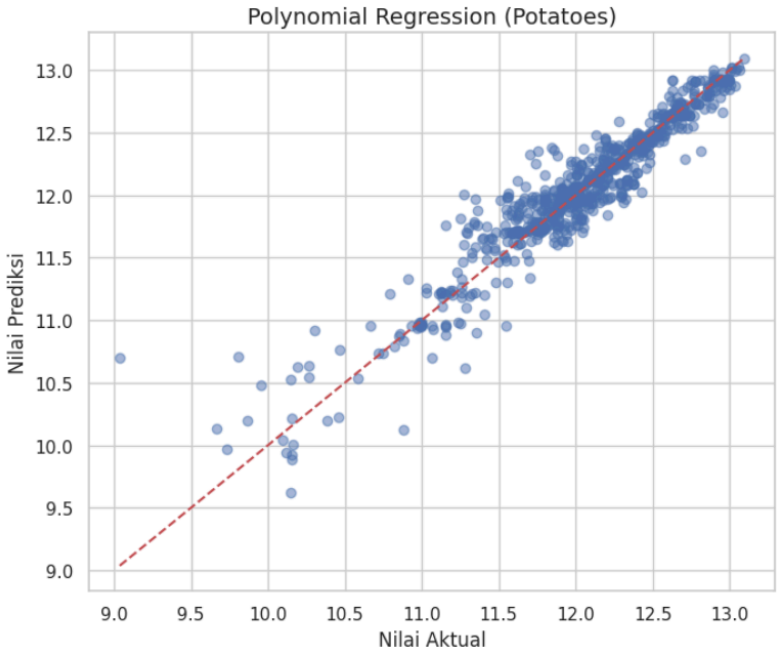
Gambar 3.5 *Polynomial Regression (Albania)*

Evaluasi Model Linear Regression:

- Root Mean Squared Error (RMSE): 1.0943
- Mean Absolute Error (MAE): 0.8412
- Koefisien Determinasi (R^2): 0.0049

Pada negara Albania, polynomial regression memiliki R^2 yang sangat rendah, yaitu 0.0049, yang mana hasilnya lebih buruk dari linear regression. Hal ini mengindikasikan bahwa model polinomial overfitting pada data latih atau bahwa pola dalam data negara Albania sulit dipelajari dengan regresi polinomial. Hal ini disebabkan oleh jumlah data yang terlalu sedikit, sehingga model tidak cukup belajar.

➢ **Strategi 3: Model Per Jenis Tanaman (Contoh: Kentang)**



Gambar 3.6 *Polynomial Regression (Kentang)*

Evaluasi Model Linear Regression:

- Root Mean Squared Error (RMSE): 0.1774
- Mean Absolute Error (MAE): 0.1156
- Koefisien Determinasi (R^2): 0.9046

Regresi polinomial juga memberikan peningkatan performa pada dataset tanaman kentang (R^2 naik dari 0.8761 menjadi 0.9046). Ini menunjukkan bahwa untuk dataset ini, hubungan non-linear cukup kuat sehingga polinomial lebih mampu menangkap pola dibandingkan regresi linear.

c) Pemilihan Derajat Polinomial

Model Polynomial Regression dibangun dengan menggunakan derajat 2 untuk semua strategi. Pemilihan derajat 2 didasarkan pada pertimbangan berikut:

- Derajat yang terlalu tinggi dapat menyebabkan overfitting pada data latih.
- Derajat 2 memungkinkan model untuk menangkap interaksi non-linier dasar antar fitur, tanpa membuat kompleksitas model menjadi terlalu tinggi.
- Hasil evaluasi menunjukkan bahwa derajat ini cukup untuk memberikan peningkatan akurasi dibandingkan model linier murni, terutama dalam beberapa strategi.

Pemilihan derajat ini dapat dievaluasi ulang dengan pendekatan sistematis seperti pencarian grid (GridSearchCV) jika diperlukan untuk optimasi lebih lanjut.

d) Teknik Validasi

Untuk mengukur performa model dan mencegah overfitting, digunakan teknik validasi sebagai berikut:

- Train-Test Split

  Dataset dibagi menjadi dua bagian:

  - Data latih (80%) digunakan untuk melatih model.
  - Data uji (20%) digunakan untuk mengevaluasi kinerja model.

  Train-test split digunakan secara konsisten dalam semua strategi, baik untuk dataset penuh, subset negara, maupun subset tanaman.

- Standarisasi Fitur

  Untuk model yang hanya menggunakan data numerik (seperti di strategi Albania), dilakukan standarisasi menggunakan 'StandardScaler' untuk memastikan semua fitur berada pada skala yang sama.

- Pipeline dan ColumnTransformer

  Untuk strategi dengan fitur campuran (numerik dan kategorikal), seperti pada model tanaman kentang, digunakan 'Pipeline' dan 'ColumnTransformer' untuk memastikan proses preprocessing dilakukan secara konsisten selama pelatihan dan prediksi.

4) Evaluasi Model

a) **Metrik Evaluas**i

Dalam evaluasi model-model regresi untuk dataset hasil panen, digunakan tiga metrik utama:

i. RMSE (Root Mean Squared Error)

RMSE adalah akar kuadrat dari rata-rata error kuadrat antara nilai prediksi model dan nilai aktual. Semakin kecil nilai RMSE, semakin baik model RMSE memberikan bobot lebih besar pada error yang besar (karena dikuadratkan). RMSE memiliki satuan yang sama dengan variabel target. RMSE erguna untuk mendeteksi error besar yang tidak diinginkan.

ii. MAE (Mean Absolute Error)

MAE adalah rata-rata dari nilai absolut error antara nilai prediksi model dan nilai aktual. Semakin kecil nilai MAE, semakin baik model. MAE memperlakukan semua error dengan bobot yang sama. MAE juga memiliki satuan yang sama dengan variabel target. Lebih mudah diinterpretasikan karena secara langsung menunjukkan rata-rata besarnya deviasi.

iii. R^2 (Coefficient of Determination)

R^2 mengukur proporsi variasi dalam variabel dependen yang dapat dijelaskan oleh variabel independen dalam model. Nilai berkisar antara 0 hingga 1 (dalam beberapa kasus bisa negatif) Semakin mendekati 1, semakin baik model.

- $R^2 = 0$ berarti model tidak menjelaskan variasi data sama sekali
- $R^2 = 1$ berarti model menjelaskan semua variasi data (fit sempurna)
- $R^2 = 0.7$ berarti model menjelaskan 70% variasi dalam data

b) Perbandingan Kinerja Regresi linear dan polinomial

Tabel berikut menyajikan perbandingan kinerja antara model Linear Regression dan Polynomial Regression berdasarkan dataset penuh, dataset per negara (contoh: Albania), dan dataset per jenis tanaman (contoh: Kentang).

❖ **Dataset Penuh**

Strategi	Model	RMSE	MAE	R^2
Dataset Penuh	Linear	0.4578	0.3477	0.8290
	Polynomial	0.2148	0.1400	0.9623

Pada dataset penuh, Regresi Polinomial menunjukkan performa yang jauh lebih baik dibandingkan Regresi Linear:

- RMSE: Regresi Polinomial (0,2148) menunjukkan error 53% lebih rendah dibandingkan Regresi Linear (0,4578)
- MAE: Regresi Polinomial (0,1400) menunjukkan error 60% lebih rendah dibandingkan Regresi Linear (0,3477)

- R^2: Regresi Polinomial mencapai nilai 0,9623 (sangat baik) dibandingkan Regresi Linear dengan 0,8290

Hubungan antar variabel pada dataset penuh bersifat non-linear. Model polinomial mampu menangkap kerumitan data secara lebih efektif dibandingkan model linear sederhana.

❖ **Per Negara (Albania)**

Strategi	Model	RMSE	MAE	R^2
Per Negara (Albania)	Linear	1.0039	0.7798	0.1624
	Polynomial	1.0943	0.8412	0.0049

Pada dataset spesifik Albania, kedua model menunjukkan performa yang kurang memuaskan:

- RMSE: Regresi Linear (1,0039) justru sedikit lebih baik dibandingkan Regresi Polinomial (1,0943)
- MAE: Regresi Linear (0,7798) juga lebih rendah dibandingkan Regresi Polinomial (0,8412)
- R^2: Kedua model menunjukkan nilai R^2 yang sangat rendah, dengan Regresi Linear (0,1624) masih lebih baik dibandingkan Regresi Polinomial yang hampir nol (0,0049)

Nilai R^2 yang sangat rendah menunjukkan bahwa model gagal menjelaskan variabilitas data. Hal ini disebabkan oleh jumlah data yang terlalu kecil pada dataset per negara (Albania) dan overfitting pada model polinomial karena jumlah data sedikit.

❖ **Per Tanaman (Kentang)**

Strategi	Model	RMSE	MAE	R^2
Per Jenis Tanaman (Kentang)	Linear	0.2021	0.1380	0.8761
	Polynomial	0.1774	0.1156	0.9046

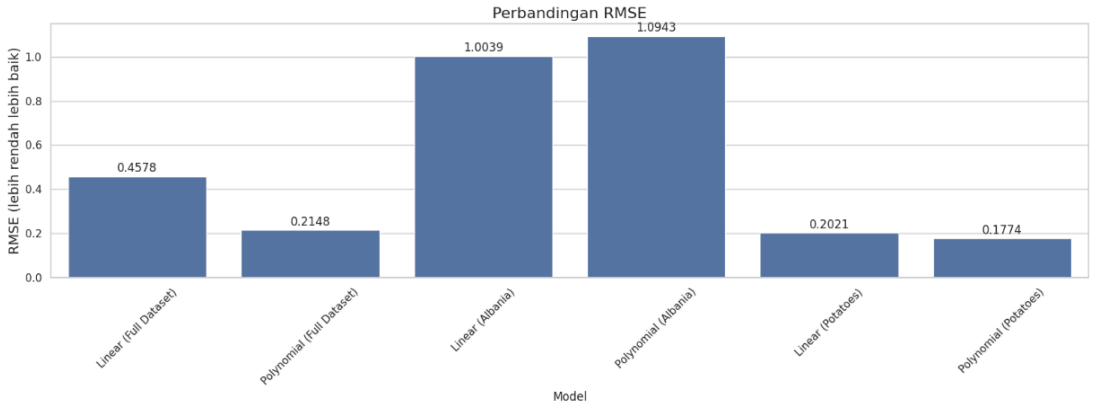
Untuk dataset per jenis tanaman, kedua model menunjukkan performa yang cukup baik, dengan Regresi Polinomial sedikit unggul:

- RMSE: Regresi Polinomial (0,1774) lebih rendah dibandingkan Regresi Linear (0,2021)
- MAE: Regresi Polinomial (0,1156) lebih rendah dibandingkan Regresi Linear (0,1380)
- R^2: Kedua model menunjukkan nilai R^2 yang baik, dengan Regresi Polinomial (0,9046) sedikit lebih tinggi dibandingkan Regresi Linear (0,8761)

Produksi kentang dipengaruhi oleh faktor-faktor non-linear, seperti interaksi antara suhu, curah hujan, dan teknik pertanian. Model polinomial lebih cocok menangkap pola, meskipun Regresi Linear juga sudah memberikan performa yang baik.

c) Evaluasi Performa Model pada Berbagai Strategi

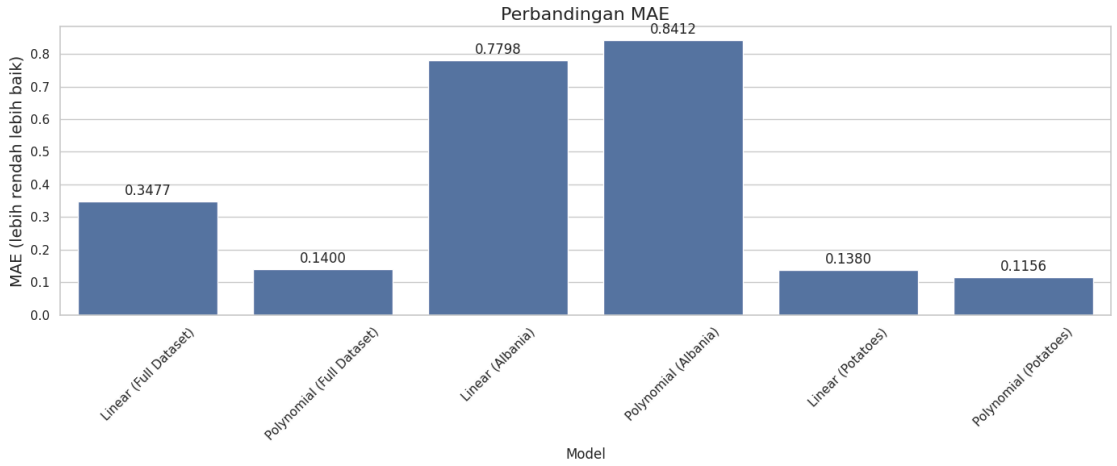
i. RMSE (Root Mean Square Error)



Gambar 4.1 *RMSE*

Pada metrik RMSE, semakin kecil nilainya maka semakin baik performa model dalam melakukan prediksi. Strategi Per Tanaman memiliki nilai RMSE terkecil (0.1774), yang menunjukkan bahwa model paling akurat saat difokuskan pada masing-masing jenis tanaman. Selanjutnya, Dataset Penuh memiliki nilai RMSE sebesar 0.2148, sementara strategi Per Negara mencatatkan performa terburuk dengan RMSE sebesar 1.0039.

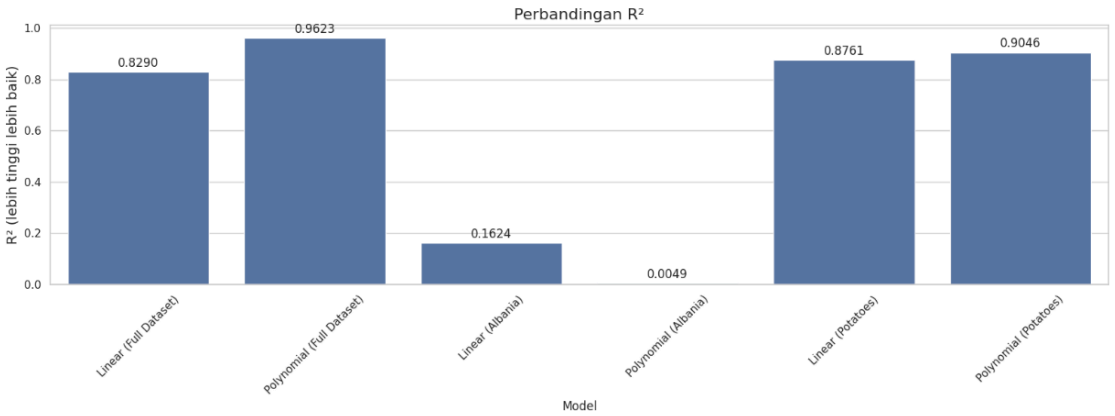
ii. MAE (Mean Absolute Error)



Gambar 4.2 *MAE*

Sama seperti RMSE, pada metrik MAE semakin kecil nilainya maka semakin baik. Strategi Per Tanaman kembali unggul dengan nilai MAE terkecil (0.1156), menandakan bahwa rata-rata selisih antara prediksi dan nilai aktual sangat kecil. Dataset Penuh berada di posisi kedua dengan MAE sebesar 0.1400, sedangkan strategi Per Negara menunjukkan hasil terburuk dengan MAE sebesar 0.7798.

iii. R^2 (Koefisien Determinasi)



Gambar 4.3 *R^2*

Pada metrik R^2, semakin besar nilainya maka semakin baik karena menunjukkan seberapa besar variasi data yang bisa dijelaskan oleh model. Strategi Dataset Penuh memiliki nilai R^2 tertinggi (0.9623), yang berarti model mampu menjelaskan sebagian besar variasi dalam data. Strategi Per Tanaman juga menunjukkan performa sangat baik dengan nilai R^2 sebesar 0.9046. Sebaliknya, strategi Per Negara hanya menghasilkan R^2 sebesar 0.1624, yang berarti model

tidak efektif dalam menjelaskan variasi data jika dibatasi hanya pada masing-masing negara.

5) Analisis Hasil

a) Interpretasi Koefisien Regresi

- Regresi Linear, koefisien regresi merepresentasikan seberapa besar pengaruh suatu fitur terhadap hasil panen. Koefisien positif menunjukkan bahwa kenaikan pada fitur tersebut meningkatkan hasil panen, sedangkan koefisien negatif menunjukkan penurunan.
- Regresi Polinomial, interpretasi menjadi lebih kompleks karena terdapat fitur-fitur hasil transformasi seperti kuadrat, kubik, dan interaksi antar variabel. Meskipun sulit untuk diinterpretasikan secara langsung, polinomial tetap memberikan keunggulan dalam akurasi karena mampu menangkap hubungan non-linear antar variabel.

b) Kesimpulan

Berdasarkan hasil analisis, model Polynomial Regression menunjukkan performa prediktif yang lebih unggul dibandingkan Linear Regression, terutama pada dataset penuh dan dataset per jenis tanaman, mengindikasikan bahwa hubungan antara faktor lingkungan dan hasil panen bersifat non-linear. Model per jenis tanaman (contohnya kentang) memberikan hasil terbaik, menunjukkan bahwa pemodelan spesifik per tanaman mampu menangkap pola dengan lebih akurat. Sebaliknya, model per negara seperti Albania justru menunjukkan performa buruk, disebabkan oleh jumlah data yang terbatas dan potensi overfitting, sehingga tidak cukup mewakili variasi pola. Analisis korelasi juga mengindikasikan bahwa faktor lingkungan tertentu seperti curah hujan tidak memiliki hubungan linear yang kuat terhadap hasil panen, sementara suhu cenderung memiliki pengaruh negatif. Tren historis menunjukkan peningkatan hasil panen dari waktu ke waktu, yang kemungkinan dipengaruhi oleh peningkatan teknologi pertanian. Secara keseluruhan, pendekatan pemodelan yang mempertimbangkan kompleksitas hubungan antar variabel dan segmentasi data yang tepat (misalnya per tanaman) mampu meningkatkan akurasi prediksi hasil panen secara signifikan.