Big data & Predictive Analytics Final project

Pemodelan Prediktif Hasil Panen Padi Menggunakan Data Lingkungan dan Agrikultur

Dosen pengampu:
Mulia Sulistiyono, S.Kom., M.Kom.

Anggota kelompok:
1. Egidius Dicky Narendra Ba'as, 23.11.5490
2. Farhan Ardiansyah, 23.11.5464
3. Garda Fitrananda, 23.11.5440
4. Rayan, 23.11.5486

Program studi Informatika
Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta
2025

Daftar Isi

1. Latar belakang

Padi adalah salah satu komoditas pangan strategis yang menjadi sumber makanan pokok bagi sebagian besar penduduk dunia, terutama di kawasan Asia seperti Indonesia. Sebagai negara agraris, keberhasilan sektor pertanian, khususnya dalam menjaga stabilitas produksi padi, sangat menentukan tingkat ketahanan pangan nasional. Namun, dalam beberapa dekade terakhir, dunia pertanian menghadapi tantangan besar. Perubahan iklim, pertumbuhan jumlah penduduk, serta tuntutan efisiensi dalam penggunaan lahan dan sumber daya menjadi faktor yang sangat mempengaruhi produktivitas pertanian.

Fluktuasi cuaca, meningkatnya suhu global, dan intensitas musim hujan yang tak menentu menjadikan proses produksi pangan semakin sulit diprediksi. Di sisi lain, praktik pertanian modern seperti penggunaan bahan kimia atau pestisida harus dijalankan secara hati-hati agar tidak menimbulkan dampak negatif jangka panjang terhadap lingkungan. Dalam konteks ini, memahami hubungan antara kondisi lingkungan dan hasil panen menjadi semakin penting sebagai dasar pengambilan keputusan yang lebih cerdas dan berkelanjutan.

- **Alasan Analisis pada bidang ini**

  Kemajuan teknologi informasi telah membuka peluang baru dalam dunia pertanian, terutama melalui penerapan analisis data. Dengan memanfaatkan data historis dan pendekatan prediktif, kita dapat melihat pola-pola penting yang sebelumnya sulit teridentifikasi. Analisis ini sangat berguna untuk membantu petani merencanakan musim tanam secara lebih presisi, memberi masukan kepada pembuat kebijakan, serta memberikan panduan bagi peneliti dalam mengkaji faktor-faktor yang memengaruhi hasil pertanian.

  Lebih jauh, pendekatan ini berpotensi besar dalam mengantisipasi dampak perubahan iklim, mendorong efisiensi penggunaan sumber daya, serta meningkatkan ketahanan pangan jangka panjang. Oleh karena itu, penelitian di bidang ini menjadi semakin relevan dan bernilai strategis.

- **Tujuan Penelitian**

  Penelitian ini bertujuan untuk membangun model prediktif yang dapat memperkirakan hasil panen padi dengan memanfaatkan informasi dari kondisi lingkungan dan praktik agrikultur. Dengan model ini, diharapkan dapat dihasilkan wawasan yang berguna untuk mendukung pengambilan keputusan di sektor pertanian, serta membantu mengembangkan strategi pertanian yang lebih adaptif dan berkelanjutan.

2. Metode

2.1. Alur Final Project

Alur kerja dari final project ini adalah sebagai berikut:

1. **Data Collection**
   Mendownload dataset agrikultur dan lingkungan terkait hasil panen dari sumber kaggle.
2. **Data Cleaning & Preprocessing**
   - Menggabungkan data dari berbagai sheet.
   - Menghilangkan missing value dan duplikasi.
   - Normalisasi nama kolom dan tipe data.
   - Penyatuan format agar kompatibel dengan tahap eksplorasi dan modeling.
3. **Exploratory Data Analysis (EDA)**
   - Visualisasi distribusi hasil panen.
   - Analisis hubungan antara suhu, curah hujan, dan penggunaan pestisida terhadap hasil panen.
   - Deteksi outlier dan korelasi awal antar variabel.
4. **Eksperimen Model**
   - Korelasi antar fitur numerik.
   - Regresi linier sederhana dan multivariat.
5. **Evaluasi Model**
   - Pengukuran akurasi model dengan MAE, RMSE, dan R^2.
   - Analisis kesalahan prediksi.
6. **Deploy dan Visualisasi**
   - Visualisasi hasil prediksi dalam dashboard menggunakan Streamlit.

2.1. Dataset

Dataset yang digunakan berasal dari Kaggle yang dimuat oleh Rishi Patel: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/data

Dataset ini berisi data hasil panen padi dan faktor lingkungan dari berbagai negara sejak tahun 1961.

Kolom-kolom utama:

- Area Code: Kode wilayah negara.
- Country: Nama negara.
- Crop_Item: Jenis tanaman (fokus: padi).
- Year: Tahun data dicatat.
- Yield_Value (ton/hektar): Hasil panen dalam ton per hektar.
- Pesticide_Value (ton): Jumlah penggunaan pestisida per tahun.
- Avg_Rainfall_mm (mm/year): Rata-rata curah hujan tahunan.
- Avg_Temperature_celsius (°C): Rata-rata suhu tahunan.

Area Code	Country	Crop_Item	Year	Yield_Value (ton/hektar)	Pesticide_Value (ton)	Avg_Rainfall_mm (mm/year)	Avg_Temperature_celsius (°C)
2	Afghanistan	Rice, paddy	1961	15190	36861.81242	1149.958504	14.23
2	Afghanistan	Rice, paddy	1962	15190	36861.81242	1149.958504	14.1
2	Afghanistan	Rice, paddy	1963	15190	36861.81242	1149.958504	15.01
2	Afghanistan	Rice, paddy	1964	17273	36861.81242	1149.958504	13.73

2.2. EDA

Pada tahap eksplorasi data dilakukan langkah-langkah berikut:

- **Descriptive Statistics**
  Menghitung nilai rata-rata, median, standar deviasi dari tiap fitur numerik.
- **Distribusi dan Korelasi**
  - Visualisasi histogram untuk Yield_Value.
  - Heatmap korelasi antara Yield, Temperature, Rainfall, dan Pesticide_Use.
  - Scatter plot antara variabel X dan Y.
- **Outlier Detection**
  Menggunakan boxplot untuk mendeteksi nilai ekstrem terutama di suhu dan pestisida.
- **Insight Awal**
  - Korelasi positif antara curah hujan dan hasil panen.
  - Korelasi negatif kecil antara suhu tinggi dan hasil panen.

3. Eksperimen

Eksperimen dilakukan untuk membangun model prediktif menggunakan pendekatan regresi linier.

Library yang digunakan:

- pandas, numpy: manipulasi data
- matplotlib, seaborn: visualisasi
- sklearn: model regresi dan evaluasi

Tahapan:

1. **Feature Selection**
   - Fitur input: Avg_Temperature, Avg_Rainfall, Pesticide_Value
   - Target: Yield_Value
2. **Modeling**
   - Regresi Linier Sederhana: satu variabel prediktor.
   - Regresi Linier Berganda: tiga variabel input.
3. **Training & Testing**
   - Membagi data menjadi 80% training dan 20% testing.

4. Hasil dan Evaluasi

Evaluasi dilakukan pada Model Utama yang menggunakan semua fitur. Data dibagi menjadi 80% data latih dan 20% data uji untuk mengukur performa model pada data yang belum pernah dilihat sebelumnya.

Hasil Evaluasi:

- R-squared (R^2): 0.2239
- Mean Absolute Error (MAE): 13228.09 ton/ha
- Root Mean Squared Error (RMSE): 16545.45 ton/ha

Interpretasi:

- Nilai R-squared sebesar 0.2239 menunjukkan bahwa model Regresi Linier hanya mampu menjelaskan sekitar 22.4% dari variasi data hasil panen. Ini mengindikasikan bahwa hubungan antara fitur dan target kemungkinan besar tidak sepenuhnya linier.

5. Kesimpulan

5.1. Kesimpulan

- Model prediktif berbasis regresi linier dapat digunakan untuk memperkirakan hasil panen padi dengan tingkat akurasi yang cukup baik.
- Faktor lingkungan seperti curah hujan, suhu, dan penggunaan pestisida berkontribusi signifikan terhadap hasil panen.
- Pendekatan ini bisa membantu petani dan pengambil kebijakan untuk lebih siap menghadapi fluktuasi hasil panen.

5.2. Kontribusi

Nama	NIM	Kontribusi
Egidius Dicky Narendra Ba'as	23.11.5490	● Koordinator ● Data Collection ● Data Cleaning
Farhan Ardiansyah	23.11.5464	● EDA ● Visualisasi
Rayan	23.11.5486	● Eksperimen Korelasi ● Model Regresi
Garda Fitrananda	23.11.5440	● Evaluasi Model ● Dashboard ● Poster (bersama)

6. Lampiran

- **Dataset Bersih:**
  rice_paddy_crop_yield.csv - Google Drive
  (https://drive.google.com/file/d/1wmfKmXTWh0S6Z_Rk3gRx5XVJO12RJmxE/view)
- **Data Collection dan Cleaning:**
  Data_Cleaning.ipynb - Colab
  (https://drive.google.com/file/d/1SoBf-yTqd5M7na-34EOfTp9tBAEe722a/view?usp=drive_link)
- **EDA dan Visualisasi:**
  EDA.ipynb - Colab
  (https://colab.research.google.com/drive/12eh0iOoBehomlgYG8R2dR-NSbJQqg7uG?usp=sharing)
- **Eksperimen Korelasi dan Model Regresi Linier:**
  Model Regresi.ipynb - Colab
  (https://colab.research.google.com/drive/1u0K9lLOkk3C0DRo3yPD2kR6WII2hoLL9?usp=sharing)

- **Evaluasi Model :**
  Evaluasi_Model.ipynb - Colab
  (https://colab.research.google.com/drive/16kxwl7jXJbFek-g7Iz2FI6IkheS8Q2Vu?usp=sharing)
- **Dashboard :**
  Dashboard Prediksi Panen Padi · Streamlit
  (https://dashboard-prediksi-panen.streamlit.app/)
- **Poster:**
  Poster Prediksi Panen Hasil Padi - Canva Poster (A3 Portrait)
  (https://www.canva.com/design/DAGtfDDZN6s/RBPBGjkEMd2GwriinTd3iQ/edit?utm_content=DAGtfDDZN6s&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)