Big data & Predictive Analytics Final project

Analisis Tren dan Pola Kualitas Udara

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

Anggota kelompok:
1. Ekananda Naufal Arif Wicaksana, 23.11.5387
2. Anugrah Awan Cahya Putra, 23.11.5380
3. Hafidz Ar Rofi, 23.11.5400

Program studi Informatika
Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta

2025

Daftar Isi

1. Latar belakang

Kualitas udara menjadi isu kritis global karena dampaknya terhadap kesehatan masyarakat dan lingkungan. Polusi udara menyebabkan berbagai penyakit pernapasan, kardiovaskular, dan dapat mempengaruhi produktivitas ekonomi. Oleh karena itu, diperlukan sistem monitoring dan prediksi yang dapat membantu pengambilan keputusan kebijakan lingkungan.

Ancaman Kesehatan Langsung Polusi udara merupakan salah satu penyebab utama kematian prematur di dunia. Partikel halus (PM2.5) dapat menembus paru-paru dan masuk ke aliran darah, menyebabkan penyakit jantung, stroke, kanker paru-paru, dan penyakit pernapasan kronis. Analisis diperlukan untuk mengidentifikasi area dan waktu dengan risiko kesehatan tertinggi.

Populasi Rentan Anak-anak, lansia, dan penderita penyakit kronis sangat rentan terhadap polusi udara. Analisis membantu mengidentifikasi kapan dan di mana populasi ini perlu perlindungan khusus, seperti pembatasan aktivitas outdoor atau penutupan sekolah.

Tujuan dari penelitian yang kami lakukan antara lain Mengidentifikasi Tren Temporal Kualitas Udara Menentukan apakah kualitas udara di suatu wilayah mengalami perbaikan, penurunan, atau stabil dalam periode waktu tertentu. Ini penting untuk mengevaluasi efektivitas kebijakan lingkungan yang telah diterapkan. Menganalisis Pola dan Variabilitas Mengidentifikasi pola musiman, harian, atau mingguan dalam data kualitas udara. Misalnya, apakah polusi lebih tinggi pada hari kerja dibanding weekend, atau pada musim tertentu. Membangun Model Prediktif Mengembangkan model yang dapat memprediksi kualitas udara berdasarkan variabel-variabel yang dapat diukur atau diprediksi, seperti kondisi meteorologi, aktivitas lalu lintas, atau faktor temporal.

2. Metode

2.1. Alur final project

Proyek ini dilaksanakan melalui tahapan sistematis:

- **Pengumpulan Data:** Dataset kualitas udara dari Kaggle menggunakan API.
- **Persiapan dan Pembersihan Data:** Pembersihan data, penanganan missing values dan duplikat.
- **Rekayasa Fitur (*Feature Engineering*):** Konversi tanggal ke datetime dan ekstraksi fitur temporal.
- **Analisis Data Eksplorasi (EDA):** Analisis statistik deskriptif dan visualisasi untuk memahami pola data.
- **Pembuatan Model *Machine Learning*:**
  1. **Pemilihan Fitur dan Target:** (PM2.5, PM10) dan target (AQI)
  2. **Pembagian Dataset:** (train/validation/test).
  3. **Pelatihan Model:** Regresi Linier
- **Evaluasi Model:** R^2 dan RMSE.
- **Penyimpanan Model:** Model disimpan dalam format .pkl.

2.2. Dataset

Dataset "Air Quality Dataset" dari Kaggle berformat .csv dengan kolom::

- **Date:** Tanggal pencatatan data.
- **City:** Nama kota tempat data diukur.
- **Polutan:** Serangkaian kolom yang merepresentasikan konsentrasi berbagai jenis polutan udara, yaitu:
  - CO (Karbon Monoksida)
  - NO2 (Nitrogen Dioksida)
  - SO2 (Sulfur Dioksida)
  - O3 (Ozon)
  - PM2.5 (Partikulat < 2.5 mikrometer)
  - PM10 (Partikulat < 10 mikrometer)
- **AQI (Air Quality Index):** Variabel target yang merupakan indeks komposit untuk mengukur tingkat kualitas udara secara keseluruhan.

Dataset dipilih karena lengkap, mencakup polutan utama, dan bersih dari missing values..

2.3. EDA

Tahapan analisis eksplorasi data:

1. **Inspeksi Awal dan Statistik Deskriptif:** Pemeriksaan dimensi, statistik deskriptif dengan df.describe()
2. **Pembersihan Data:** Verifikasi duplikat dan nilai kosong
3. **Rekayasa Fitur:** Konversi tanggal dan ekstraksi tahun, bulan, hari
4. **Visualisasi Data:**
    - Tren bulanan per kota (line plot)
    - Perbandingan antar kota (bar plot)
    - Matriks korelasi polutan dan AQI (heatmap)

3. Eksperimen

- **Proses Eksperimen**
    1. **Persiapan Data dan Eksplorasi EDA**

       Tahap ini adalah fondasi dari seluruh project, di mana saya menyiapkan dan memahami dataset.
        - **Memuat Data**: Dataset Air_Quality.csv ke DataFrame
        - **Inspeksi Awal**: Pemeriksaan dimensi dan statistik deskriptif
        - **Pembersihan Data**: Cek duplikat dan missing values (dataset sudah bersih)
        - **Feature Engineering**: Konversi Date ke datetime dan ekstraksi fitur temporal
        - **Visualisasi**: Agregasi data, tren bulanan, perbandingan antar kota, dan analisis korelasi
    2. **Visualisasi Data**

       Pada tahap ini, saya mengubah data mentah menjadi wawasan visual untuk memahami pola dan hubungan antar variabel.
        - **Agregasi Data**: saya mengelompokkan data berdasarkan bulan dan kota lalu menghitung rata-rata dari setiap polutan. Ini memungkinkan Anda membandingkan tren kualitas udara antar kota dari waktu ke waktu.
        - **Visualisasi Tren Bulanan**: saya membuat **grafik garis (line plot)** yang menunjukkan tren rata-rata setiap polutan (CO, NO2, SO2, O3, PM2.5 dan PM10) per bulan untuk setiap kota.
        - **Visualisasi Perbandingan Rata-rata**: saya juga membuat **diagram batang (*bar plot*)** untuk membandingkan tingkat rata-rata polutan di enam kota secara langsung.
        - **Analisis Korelasi**: saya menghitung matriks korelasi antar polutan dan AQI, kemudian memvisualisasikannya menggunakan **heatmap**. Ini membantu saya melihat seberapa kuat hubungan antar variabel, misalnya, polutan mana yang paling berpengaruh terhadap nilai AQI.
    3. **Pembuatan Model *Machine Learning***

       Di sini Anda membangun model prediktif untuk memprediksi nilai AQI berdasarkan tingkat polutan.
        - **Pemilihan Fitur**: X = PM2.5, PM10; y = AQI
        - **Pembagian Dataset**: Train/validation/test split
        - **Pelatihan**: Model Regresi Linear dengan model.fit()
    4. **Evaluasi Model**

Tahap terakhir adalah mengukur seberapa baik model saya bekerja.

- **Metrik**: MSE, RMSE, $R^2 = 0.72$ (menjelaskan 72% variasi AQI)
- **Visualisasi**: Scatter plot prediksi vs aktual
- **Penyimpanan**: Model disimpan sebagai model_aqi.pkl

- **Tools yang digunakan**

  Berikut adalah pustaka (*library*) Python dan *tools* lain yang saya manfaatkan dalam proyek ini:

  1. **Platform**: Google Colab
  2. **Data**: Kaggle API, Pandas
  3. **Visualisasi**: Matplotlib, Seaborn
  4. **Machine Learning**: Scikit-learn (LinearRegression, train_test_split, metrik evaluasi)
  5. **Lainnya**: NumPy, Joblib, Streamlit

4. Hasil dan Evaluasi

- **Hasil Eksperimen**

  Eksperimen ini berhasil mengungkap beberapa wawasan penting mengenai kualitas udara di enam kota yang dianalisis. Berikut adalah temuan utamanya:

  1. **Hubungan Kuat Antar Polutan**: Ditemukan bahwa ada **korelasi positif yang kuat** antara sebagian besar polutan dengan nilai **AQI (Indeks Kualitas Udara)**. Artinya, ketika tingkat polutan seperti **PM2.5 dan PM10** meningkat, nilai AQI juga cenderung ikut naik secara signifikan. Ini mengonfirmasi bahwa polutan-polutan tersebut adalah pendorong utama buruknya kualitas udara.
  2. **Pola Tren Bulanan**: Visualisasi data menunjukkan adanya **pola musiman atau tren bulanan** pada tingkat polusi di beberapa kota. Misalnya, ada kota yang tingkat polusinya cenderung memuncak pada pertengahan tahun, sementara kota lain mungkin lebih stabil. Ini menunjukkan bahwa faktor seperti cuaca atau aktivitas musiman kemungkinan besar mempengaruhi kualitas udara.
  3. **Perbandingan Antar Kota**: Analisis perbandingan menunjukkan adanya **perbedaan yang jelas** dalam tingkat polusi rata-rata di antara enam kota tersebut. Beberapa kota secara konsisten menunjukkan tingkat polusi yang lebih tinggi dibandingkan yang lain, menyoroti adanya perbedaan kualitas udara yang signifikan secara geografis.

  Secara singkat, eksperimen ini berhasil membuktikan bahwa dengan menganalisis data polutan, kita tidak hanya bisa memahami kondisi kualitas udara saat ini, tetapi juga memprediksi nilainya dengan cukup akurat.

- **Evaluasi Model dan Hasilnya**

  Setelah model *machine learning* (Regresi Linear) selesai dilatih, saya melakukan evaluasi untuk mengukur seberapa baik performanya. Evaluasi ini sangat penting untuk memastikan model tersebut dapat diandalkan.

  ❖ Metode Evaluasi yang Digunakan:

  ➢ **R-squared (R^2)**: Metrik ini mengukur seberapa besar persentase variasi atau perubahan nilai AQI yang bisa dijelaskan oleh model saya. Skalanya dari 0 hingga 1, di mana 1 berarti model sempurna.

- ➢ **Root Mean Squared Error (RMSE)**: Metrik ini mengukur rata-rata kesalahan prediksi yang dibuat oleh model. Semakin kecil nilainya, semakin akurat prediksinya.

❖ Hasil Evaluasi:

- ➢ **Skor R-squared (R^2) = 0.72**:
  - ■ **Artinya**: Model saya mampu **menjelaskan sekitar 72% dari variasi data AQI**. Ini adalah hasil yang **sangat baik** dan menunjukkan bahwa fitur-fitur polutan yang saya gunakan sangat relevan untuk memprediksi kualitas udara.
- ➢ **Skor Root Mean Squared Error (RMSE) = 13.34**:
  - ■ **Artinya**: Rata-rata, prediksi yang dibuat oleh model saya memiliki **kesalahan sekitar 13.34 poin dari nilai AQI sebenarnya**. Mengingat rentang nilai AQI dalam dataset Anda cukup lebar (dari 8 hingga 188), angka kesalahan ini tergolong **rendah**, yang menandakan akurasi yang baik.
- ➢ **Performa Konsisten**: Salah satu temuan terpenting adalah skor evaluasi pada **data latih dan data uji hampir identik**. Ini adalah indikator kuat bahwa model Anda **tidak *overfitting*** dan mampu melakukan generalisasi dengan baik pada data baru yang belum pernah dilihat sebelumnya.

5. Kesimpulan

5.1. Kesimpulan

Proyek ini berhasil menunjukkan bagaimana data kualitas udara dari berbagai kota global dapat dianalisis secara menyeluruh menggunakan pendekatan Big Data dan Machine Learning. Dengan memanfaatkan data historis dari enam kota besar dunia, kami menemukan bahwa:

- **Terdapat hubungan korelatif yang kuat antara polutan utama (terutama PM2.5 dan PM10) terhadap indeks kualitas udara (AQI).** Hal ini berarti semakin tinggi konsentrasi partikel polutan, semakin buruk kualitas udara suatu wilayah.
- **Tren musiman** dan pola temporal dalam kualitas udara terlihat jelas, yang menunjukkan bahwa faktor waktu, seperti bulan atau musim, sangat mempengaruhi tingkat polusi.
- Model prediktif berbasis **Regresi Linear** mampu memprediksi nilai AQI dengan baik, ditunjukkan oleh skor evaluasi yang cukup tinggi:
    - **R^2 sebesar 0.72**, menunjukkan model dapat menjelaskan 79% variasi dalam data.
    - **RMSE sebesar 13.34**, yang menandakan bahwa rata-rata kesalahan prediksi cukup rendah.
- **Visualisasi interaktif dan dashboard** yang dibangun menggunakan Streamlit memungkinkan pengguna untuk mengeksplorasi data dan melakukan prediksi AQI secara langsung berdasarkan input PM2.5 dan PM10, sehingga memperkuat aspek usability dan user-centered design dari proyek ini.

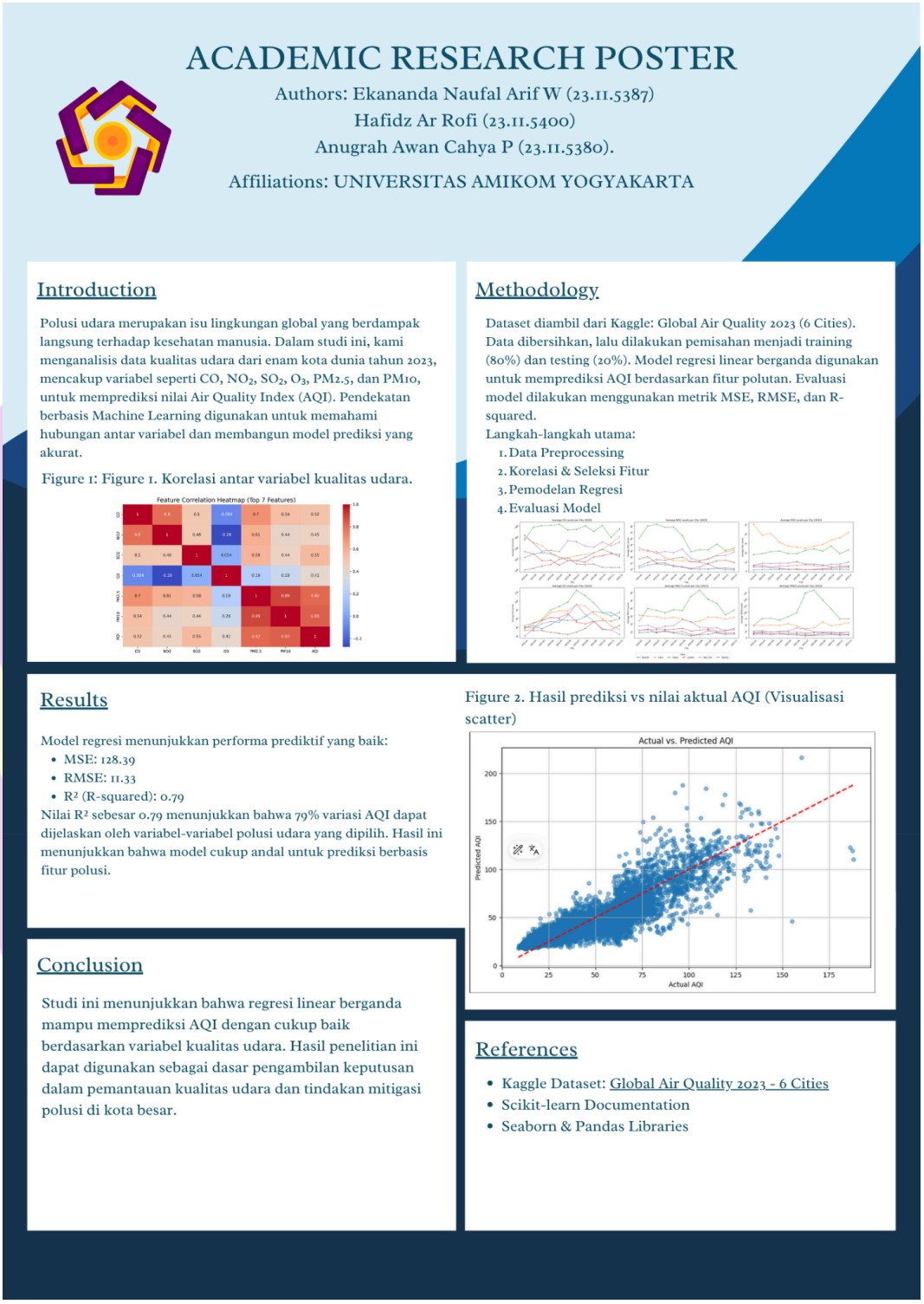
Dengan kombinasi antara analisis eksploratif, pemodelan prediktif, dan penyajian interaktif melalui dashboard, proyek ini tidak hanya menghasilkan hasil analitis yang kuat, tetapi juga solusi yang aplikatif untuk pemantauan dan prediksi kualitas udara.

5.2. Kontribusi

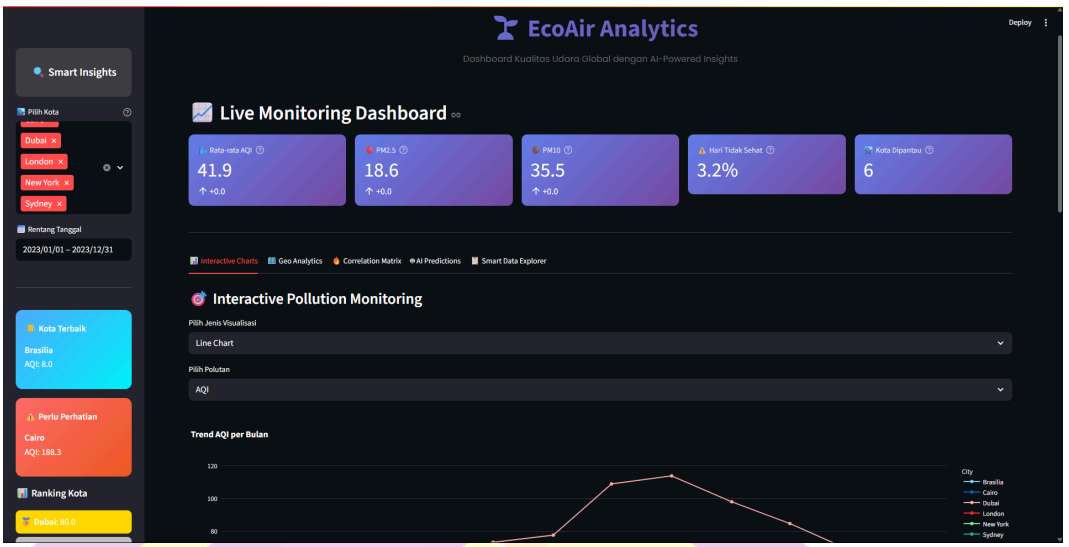
Ekananda Naufal Arif Wicaksana pada final project ini berperan membuat visualisasi, membuat model machine learning, dan membuat dashboard. Anugrah Awan Cahya Putra berperan membuat poster dan laporan. Hafidz: Ar Rofi berperan membuat poster dan laporan.

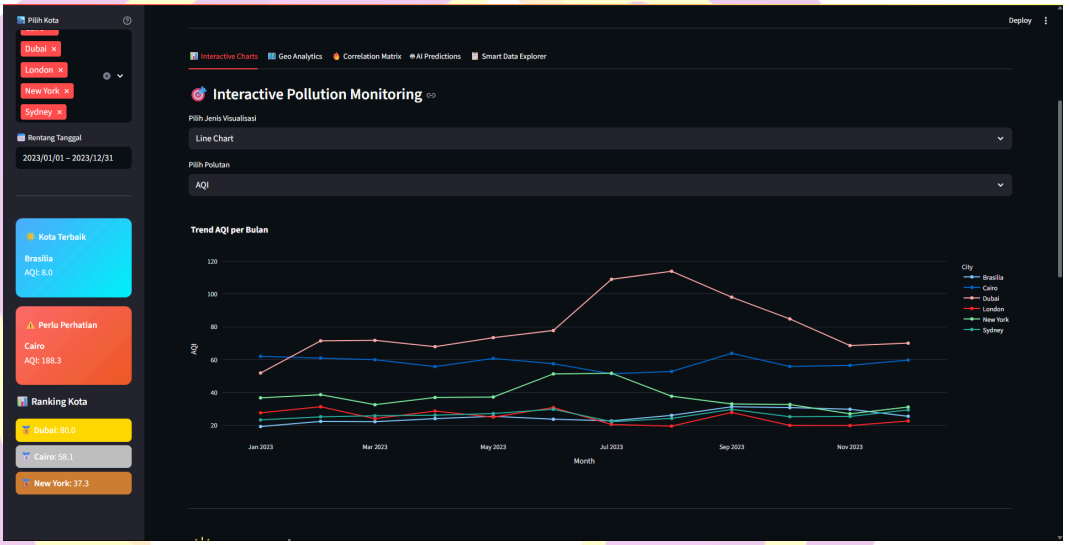
6. Lampiran

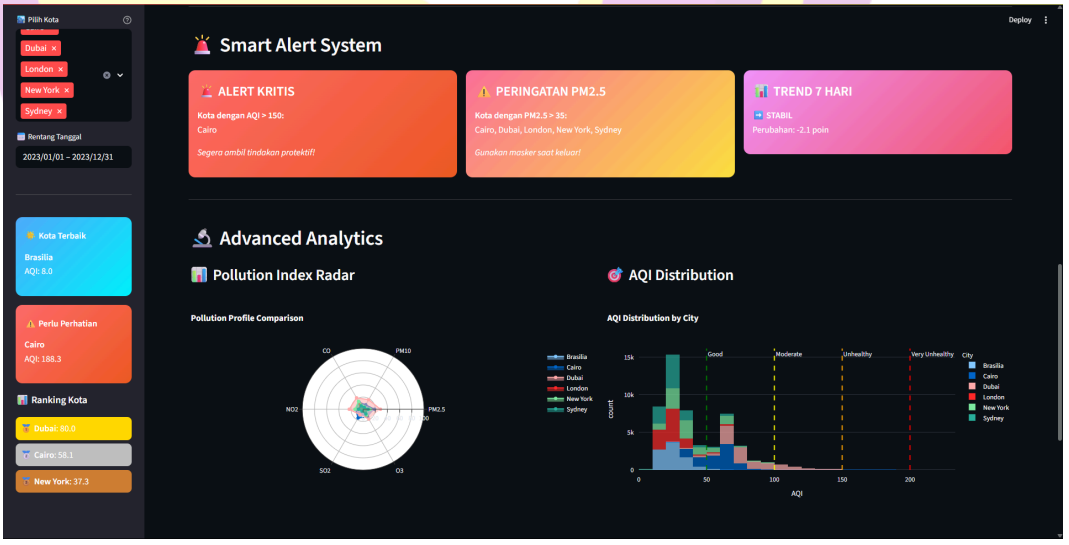
- Poster:

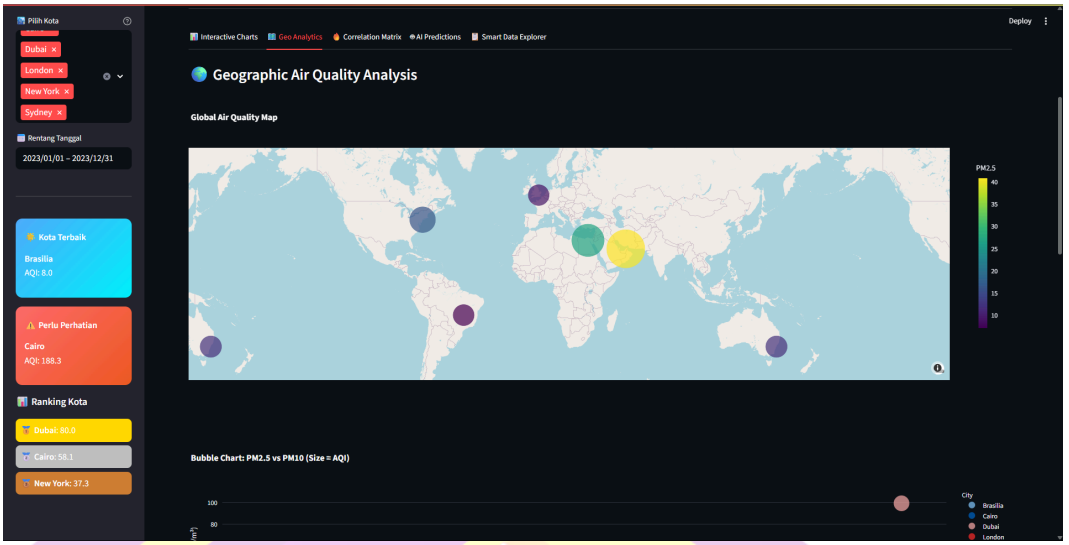


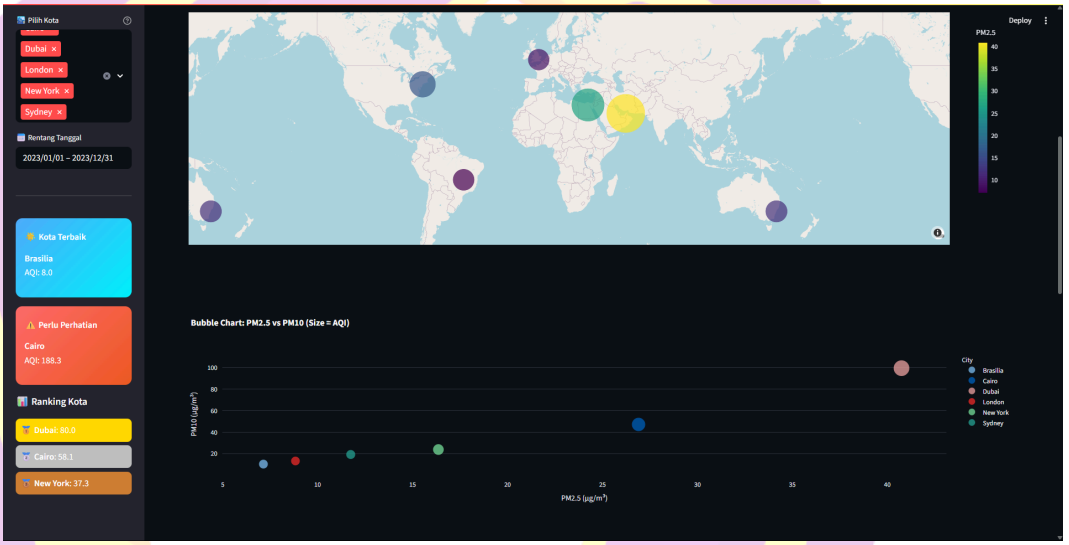
- Dashboard:

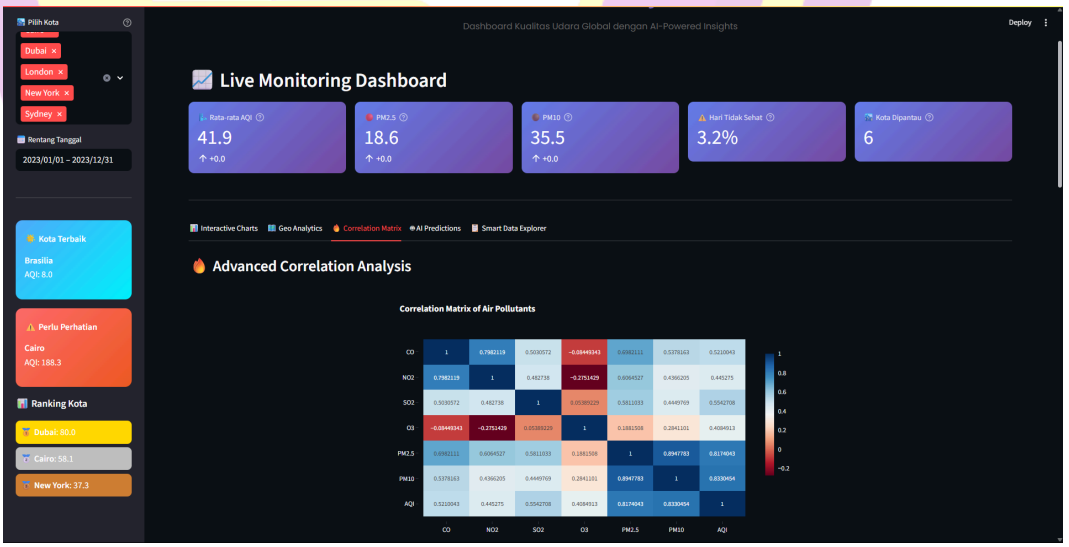

EcoAir Analytics
Dashboard Kualitas Udara Global dengan AI-Powered Insights
Deploy
Smart Insights
Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Live Monitoring Dashboard
Rata-rata AQI
41.9
+0.0
PM2.5
18.6
+0.0
PM10
35.5
+0.0
Hari Tidak Sehat
3.2%
Kota Dipantau
6
Interactive Charts
Geo Analytics
Correlation Matrix
AI Predictions
Smart Data Explorer
Interactive Pollution Monitoring
Pilih Jenis Visualisasi
Line Chart
Pilih Polutan
AQI
Trend AQI per Bulan
City
Brasilia
Cairo
Dubai
London
New York
Sydney

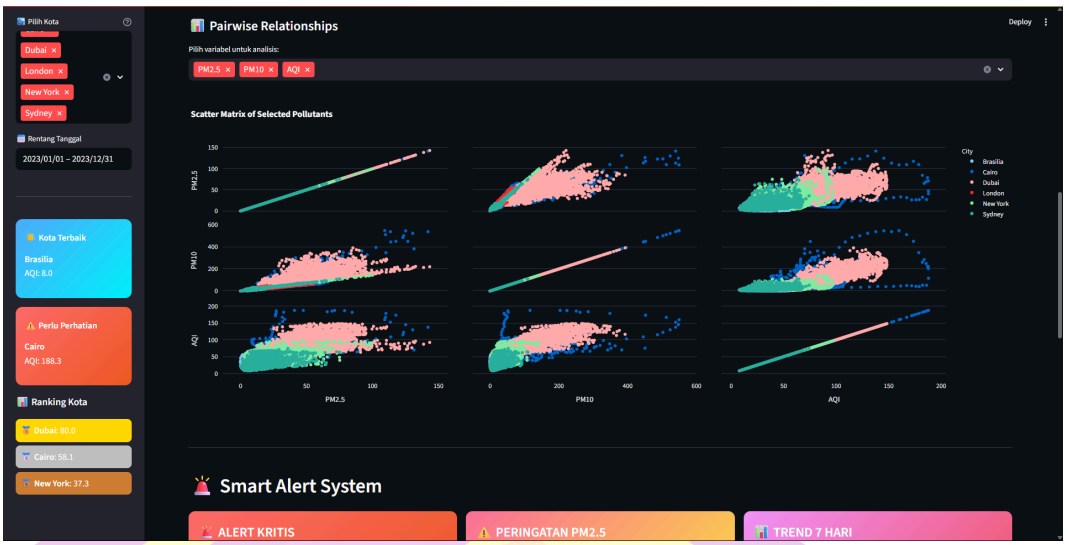

Deploy
Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Interactive Charts
Geo Analytics
Correlation Matrix
AI Predictions
Smart Data Explorer
Interactive Pollution Monitoring
Pilih Jenis Visualisasi
Line Chart
Pilih Polutan
AQI
Trend AQI per Bulan
AQI
Month
Jan 2023
Mar 2023
May 2023
Jul 2023
Sep 2023
Nov 2023
City
Brasilia
Cairo
Dubai
London
New York
Sydney

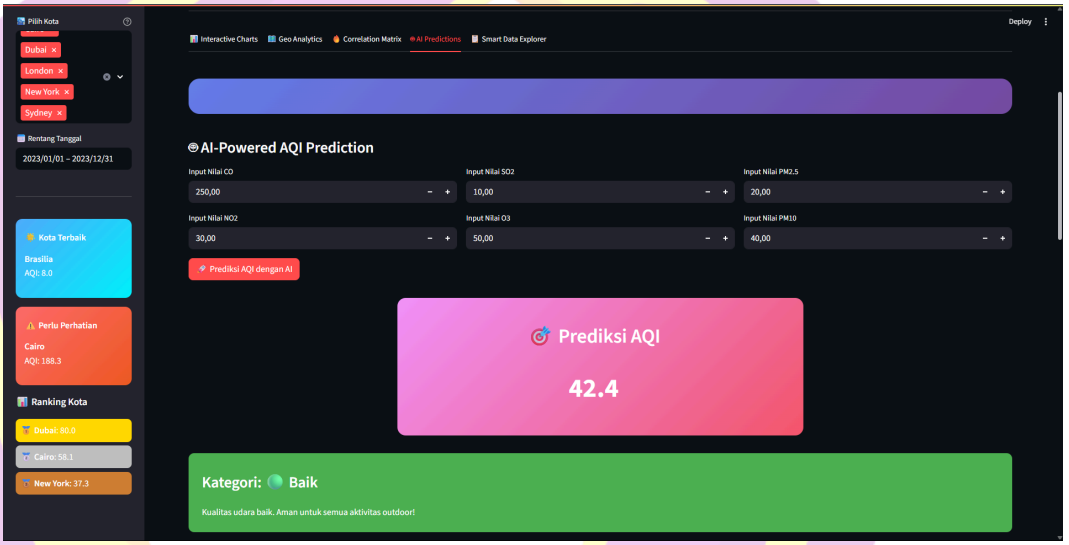

Deploy
Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Smart Alert System
ALERT KRITIS
Kota dengan AQI > 150:
Cairo
Segera ambil tindakan protektif!
PERINGATAN PM2.5
Kota dengan PM2.5 > 35:
Cairo, Dubai, London, New York, Sydney
Gunakan masker saat keluar!
TREND 7 HARI
STABIL
Perubahan: -2.1 poin
Advanced Analytics
Pollution Index Radar
Pollution Profile Comparison
CO
PM10
PM2.5
O3
SO2
NO2
AQI Distribution
AQI Distribution by City
Good
Moderate
Unhealthy
Very Unhealthy
count
AQI
City
Brasilia
Cairo
Dubai
London
New York
Sydney

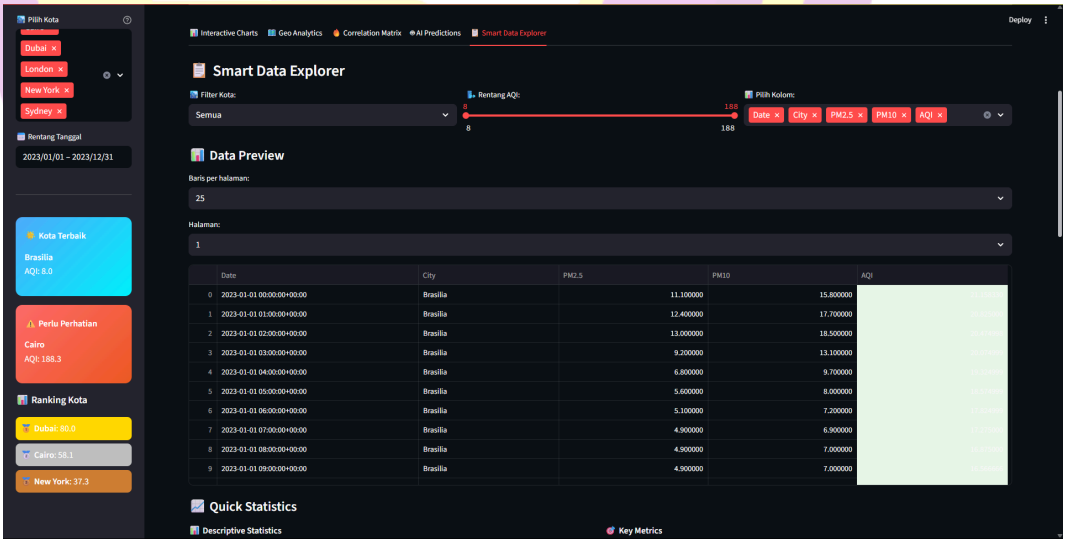

Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Deploy
Interactive Charts
Geo Analytics
Correlation Matrix
AI Predictions
Smart Data Explorer
Geographic Air Quality Analysis
Global Air Quality Map
PM2.5
Bubble Chart: PM2.5 vs PM10 (Size = AQI)
City
Brasilia
Cairo
Dubai
London

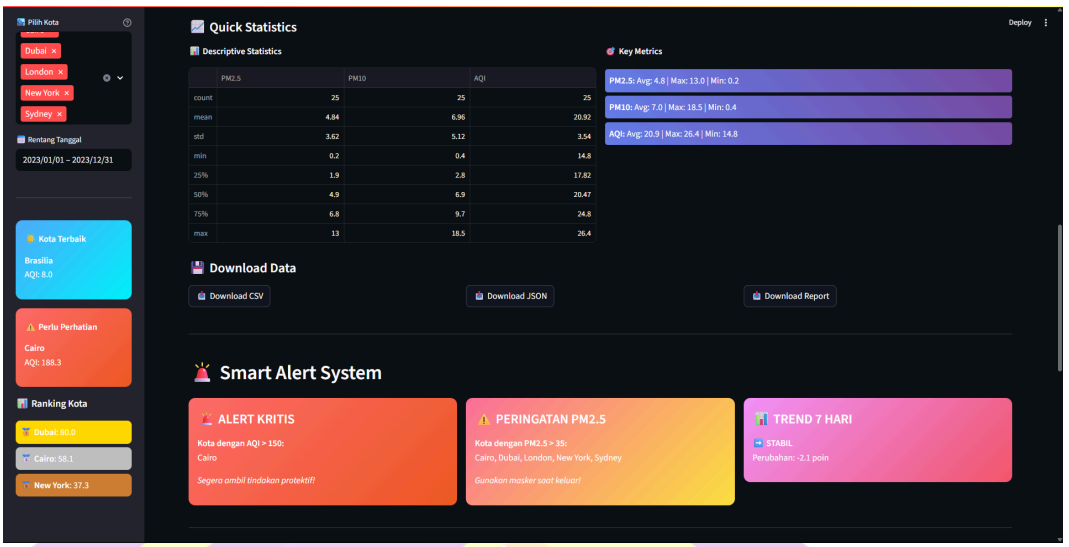

Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Deploy
PM2.5
Bubble Chart: PM2.5 vs PM10 (Size = AQI)
PM10 (μg/m³)
PM2.5 (μg/m³)
City
Brasilia
Cairo
Dubai
London
New York
Sydney


Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Dashboard Kualitas Udara Global dengan AI-Powered Insights
Deploy
Live Monitoring Dashboard
Rata-rata AQI
41.9
PM2.5
18.6
PM10
35.5
Hari Tidak Sehat
3.2%
Kota Dipantau
6
Interactive Charts
Geo Analytics
Correlation Matrix
AI Predictions
Smart Data Explorer
Advanced Correlation Analysis
Correlation Matrix of Air Pollutants
CO
NO2
SO2
O3
PM2.5
PM10
AQI


Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Pairwise Relationships
Deploy
Pilih variabel untuk analisis:
PM2.5
PM10
AQI
Scatter Matrix of Selected Pollutants
City
Brasilia
Cairo
Dubai
London
New York
Sydney
Smart Alert System
ALERT KRITIS
PERINGATAN PM2.5
TREND 7 HARI


Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Deploy
Interactive Charts
Geo Analytics
Correlation Matrix
AI Predictions
Smart Data Explorer
AI-Powered AQI Prediction
Input Nilai CO
250,00
Input Nilai SO2
10,00
Input Nilai PM2.5
20,00
Input Nilai NO2
30,00
Input Nilai O3
50,00
Input Nilai PM10
40,00
Prediksi AQI dengan AI
Prediksi AQI
42.4
Kategori: Baik
Kualitas udara baik. Aman untuk semua aktivitas outdoor!


Pilih Kota
Dubai
London
New York
Sydney
Rentang Tanggal
2023/01/01 – 2023/12/31
Kota Terbaik
Brasilia
AQI: 8.0
Perlu Perhatian
Cairo
AQI: 188.3
Ranking Kota
Dubai: 80.0
Cairo: 58.1
New York: 37.3
Deploy
Interactive Charts
Geo Analytics
Correlation Matrix
AI Predictions
Smart Data Explorer
Smart Data Explorer
Filter Kota:
Semua
Rentang AQI:
Pilih Kolom:
Date
City
PM2.5
PM10
AQI
Data Preview
Baris per halaman:
25
Halaman:
1
Date
City
PM2.5
PM10
AQI
0 2023-01-01 00:00:00+00:00 Brasilia 11.100000 15.800000
1 2023-01-01 01:00:00+00:00 Brasilia 12.400000 17.700000
2 2023-01-01 02:00:00+00:00 Brasilia 13.000000 18.500000
3 2023-01-01 03:00:00+00:00 Brasilia 9.200000 13.100000
4 2023-01-01 04:00:00+00:00 Brasilia 6.800000 9.700000
5 2023-01-01 05:00:00+00:00 Brasilia 5.600000 8.000000
6 2023-01-01 06:00:00+00:00 Brasilia 5.100000 7.200000
7 2023-01-01 07:00:00+00:00 Brasilia 4.900000 6.900000
8 2023-01-01 08:00:00+00:00 Brasilia 4.900000 7.000000
9 2023-01-01 09:00:00+00:00 Brasilia 4.900000 7.000000
Quick Statistics
Descriptive Statistics
Key Metrics



- Google colab: FP_BigData.ipynb
- Dashboard: https://airqualitypredicted.streamlit.app/
- GitHub: https://github.com/Naufallwcksn/FinalProject_BigData.git