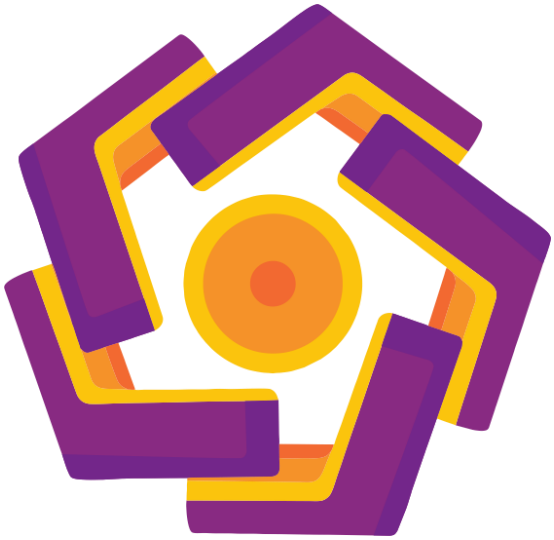
Big data & Predictive Analytics Final project

Prediksi & Analisis Kasus COVID-19 di Indonesia

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



Dibuat Oleh:

Agi Muhammad Tengku Aqamaddin 23.11.5518

Hasbie Surya Lovadeza 22.11.5042

Universitas Amikom Yogyakarta

Tahun Ajaran 2024/2025

DAFTAR ISI

1. Latar Belakang 1
2. Metode 2
2.1 Alur Final Project 2
2.2 Dataset 2
2.3 EDA (Exploratory Data Analysis) 2
3. Eksperimen 3
4. Hasil dan Evaluasi 4
5. Kesimpulan 5
5.1 Kesimpulan 5
5.2 Kontrubusi 5
6. Lampiran 6

1. Latar Belakang

COVID-19 merupakan pandemi global yang berdampak signifikan di Indonesia, baik dari segi kesehatan, ekonomi, maupun sosial. Dengan jumlah kasus yang terus berkembang setiap hari, sangat penting melakukan analisis data secara menyeluruh agar pemerintah dan masyarakat dapat memahami tren dan pola penyebaran virus. Penelitian ini bertujuan untuk memanfaatkan data COVID-19 Indonesia melalui eksplorasi data (EDA), visualisasi, serta membangun model prediksi sederhana menggunakan regresi linier. Dengan pendekatan ini, diharapkan dapat membantu menghasilkan insight yang berguna dalam pengambilan keputusan berbasis data.

2. Metode

2.1 Alur Final Project

Final project dilakukan melalui beberapa tahap:

1. Pengambilan serta pencarian data COVID-19 Indonesia dari GitHub.
2. Pembersihan dan pengecekan dataset (data cleaning).
3. Eksplorasi Data (EDA) untuk melihat distribusi, tren, dan korelasi.
4. Membangun model prediksi (Regresi Linier).
5. Evaluasi model.
6. Visualisasi hasil.
7. Penyusunan laporan dan poster.

2.2 Dataset

1. Dataset diambil dari: GitHub – indahsh/Indonesia-Covid19
2. Data berisi informasi harian COVID-19 di Indonesia, periode 1 Maret 2020 – 15 September 2022.
3. Jumlah data: ±31.822 baris dan 38 kolom (tanggal, provinsi, total kasus, kematian, recovery, dll).
4. Format: CSV, sudah cukup bersih namun tetap dicek duplikat & missing value.

2.3 EDA (Exploratory Data Analysis)

Tahap EDA dilakukan untuk memahami data secara menyeluruh:

1. Membuat grafik tren kasus baru harian.
2. Membuat histogram distribusi total kasus.
3. Membuat scatterplot hubungan antara total kasus dan total kematian.
4. Menghitung korelasi antar variabel penting dan divisualisasikan dengan heatmap.
5. Cek data outlier, missing value, dan duplikasi data.

3. Eksperimen

Eksperimen dilakukan dengan menggunakan:

1. Model: Regresi Linier sederhana (Total Cases → Total Deaths).
2. Library dan tools: Python, Pandas, Seaborn, Matplotlib, scikit-learn (LinearRegression, train_test_split, mean_absolute_error, r2_score).
3. Langkah-langkah eksperimen:
    1. Pisahkan data menjadi data latih (train) dan data uji (test) dengan rasio 80:20.
    2. Melatih model regresi menggunakan data latih.
    3. Menghitung koefisien dan intercept.
    4. Memprediksi data uji dan menghitung metrik evaluasi: MAE & R2 score.

4. Hasil dan Evaluasi

1. Koefisien model regresi linier: menunjukkan setiap kenaikan 1 kasus total berdampak pada kenaikan kematian sebesar nilai koefisien.
2. Intersep model: baseline prediksi jumlah kematian.
3. Mean Absolute Error (MAE): menggambarkan rata-rata selisih prediksi dan nilai aktual.
4. R2 score: menggambarkan seberapa baik model menjelaskan variansi data.
   Contoh hasil:
   Koefisien: [0.02737275]
   Intercept: 206.8613921375636
   MAE: 1422.5790566769163
   R2 Score: 0.939372727667665
5. Visualisasi prediksi vs data aktual menunjukkan prediksi cukup dekat dengan data sebenarnya.

5. **Kesimpulan**

5.1 Kesimpulan

Hasil analisis menunjukkan adanya korelasi positif yang kuat antara jumlah total kasus dan total kematian COVID-19 di Indonesia.Model regresi linier sederhana memiliki performa cukup baik. Visualisasi tren kasus harian, distribusi total kasus, serta scatterplot prediksi vs aktual membantu memahami dinamika pandemi di Indonesia. Prediksi model dapat digunakan sebagai referensi awal untuk memperkirakan jumlah kematian, namun tetap memerlukan data terbaru agar lebih akurat.Penelitian ini juga menunjukkan bahwa regresi linier sederhana sudah cukup efektif untuk memodelkan hubungan antara total kasus dan kematian, namun model dapat dikembangkan lebih lanjut dengan menambah variabel lain seperti jumlah testing, tingkat vaksinasi, atau faktor demografis agar hasil prediksi lebih komprehensif.

5.2 Kontribusi

Kontribusi Agi Muhammad, adalah mencari dan memilah beberapa dataset dangan jumlah minimal 1000 untuk *final project*, melakukan eksperimen prediksi dengan regresi linier sederhana dan atau regresi linier berganda dari dataset, serta membersihkan dataset csv. Kontribusi Hasbie Surya, adalah membantu dalam melakukan eksperimen prediksi, membuat laporan *final project* dan membuat poster dalam eksperimen yang dilakukan.

6. Lampiran

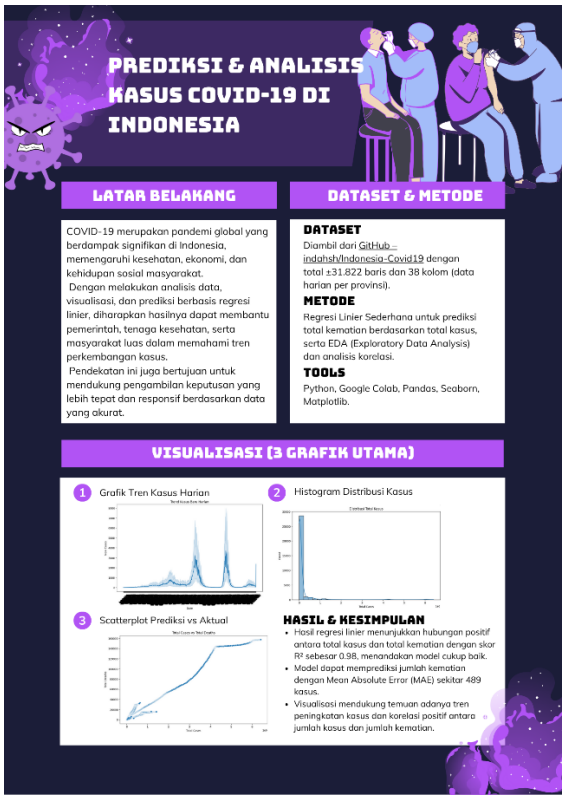
Link Google Collab : https://colab.research.google.com/drive/14TjxbGwfy8diP_vOD_Rn-8V9oODOVfCQ?usp=sharing

Link Dataset bersih : https://drive.google.com/file/d/1I4KeEMjdX8rB68DSupLbA_j-YQTnsrHV/view?usp=sharing

Link Github : https://github.com/ATengkuuu/Final_Project_Big_Data

Poster

https://drive.google.com/file/d/1CCRPjgAjfoB-ULEGvgWy8SpbJC88WcHs/view?usp=sharing



Lampiran Google Collab