Slides Pemaparan

Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

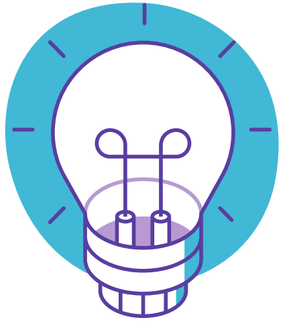
Tim Akane

Seleksi Data Science Academy COMPFEST 15 Tahun 2023

Latar Belakang

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

- Pada kurun waktu tertentu di masa pandemi, terjadi perbedaan yang cukup signifikan pada tingkat mobilitas di DKI Jakarta.
- Sejumlah negara di dunia, termasuk salah satunya Indonesia, mengalami penurunan tingkat polusi udara saat pandemi corona (Covid-19) mewabah [1].
- Kualitas udara terburuk pada tahun 2021 terjadi di Kota Jakarta [2].
- Gagasan penulis membentuk model prediksi kualitas udara menggunakan data mobilitas yang dapat digunakan pasca masa pandemi.



- Apakah pandemi COVID-19 yang telah terjadi mempengaruhi mobilitas di wilayah DKI Jakarta? Bagaimana bentuk korelasinya?

- Apakah angka mobilitas selama masa pandemi berhubungan dengan kualitas udara di wilayah DKI Jakarta? Bagaimana bentuk korelasinya?

- Dari model machine learning Linear Regression, XGBoost, dan XGBoost dengan hyperparameter tuning, manakah yang paling baik digunakan untuk memprediksi data mobilitas dan kualitas udara di wilayah DKI Jakarta?

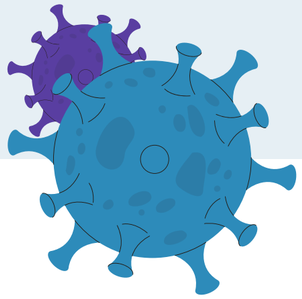
Rumusan Masalah

→ **Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta**

Hipotesis

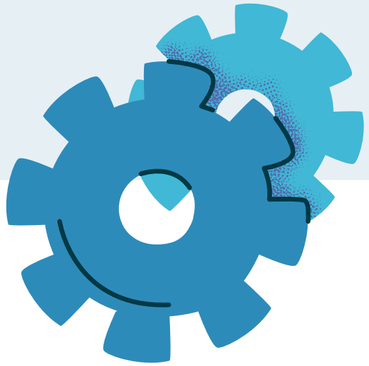
→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

- Data pandemi COVID-19 dan data mobilitas masyarakat di wilayah DKI Jakarta **saling berhubungan** dimana **data COVID-19 menyebabkan data mobilitas komunitas** dengan delay/lag tertentu.



- Tingkat mobilitas masyarakat berdasarkan data **memiliki hubungan** dengan kualitas udara di wilayah DKI Jakarta yang diwakili oleh jumlah polutan. Dimana **data mobilitas komunitas menyebabkan perubahan data kualitas udara** dengan delay/lag tertentu

- Model **XGBoost dengan hyperparameter tuning** akan menjadi model dengan **performa paling baik** yang dapat memprediksi data mobilitas masyarakat dan kualitas udara di wilayah DKI Jakarta.



Data COVID-19

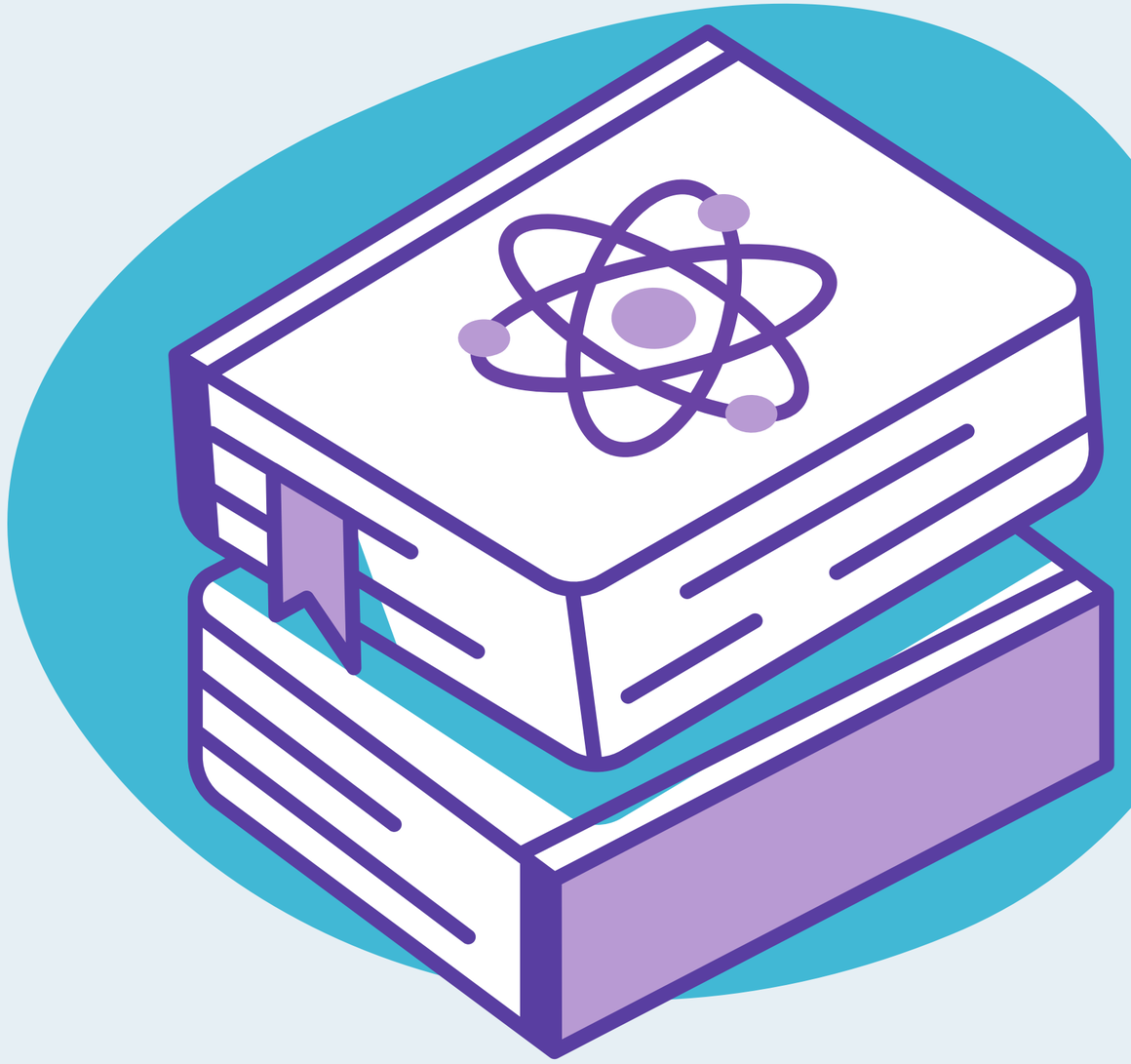
Data ini diperoleh dari platform Kaggle dengan judul **COVID-19 Indonesia Dataset** [3] yang berisi data histori COVID-19 di Indonesia secara deret waktu mulai bulan Maret 2020 - September 2022.

Data Mobilitas Komunitas

Data ini didapat dari **COVID-19 Community Mobility Reports** [4] oleh Google yang berisi data perubahan persentase mobilitas masyarakat di beberapa kategori tempat, mulai dari bulan Februari 2020 - Oktober 2022.

Data Kualitas Udara

Data ini diperoleh dari **Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta** [5] [6] melalui platform Open Data Jakarta yang berisi hasil pengukuran parameter dan kategori pencemaran udara di DKI Jakarta mulai bulan Januari 2020 - Desember 2021.

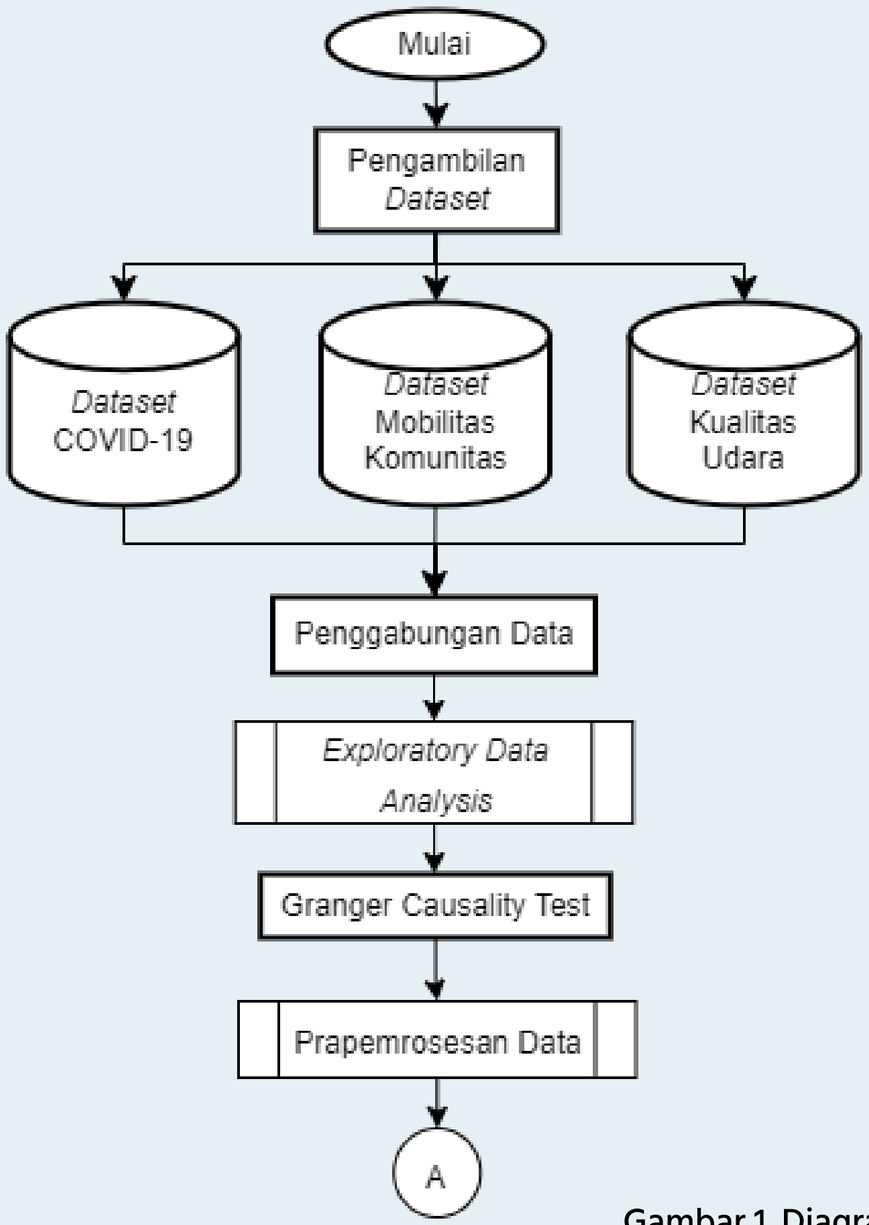


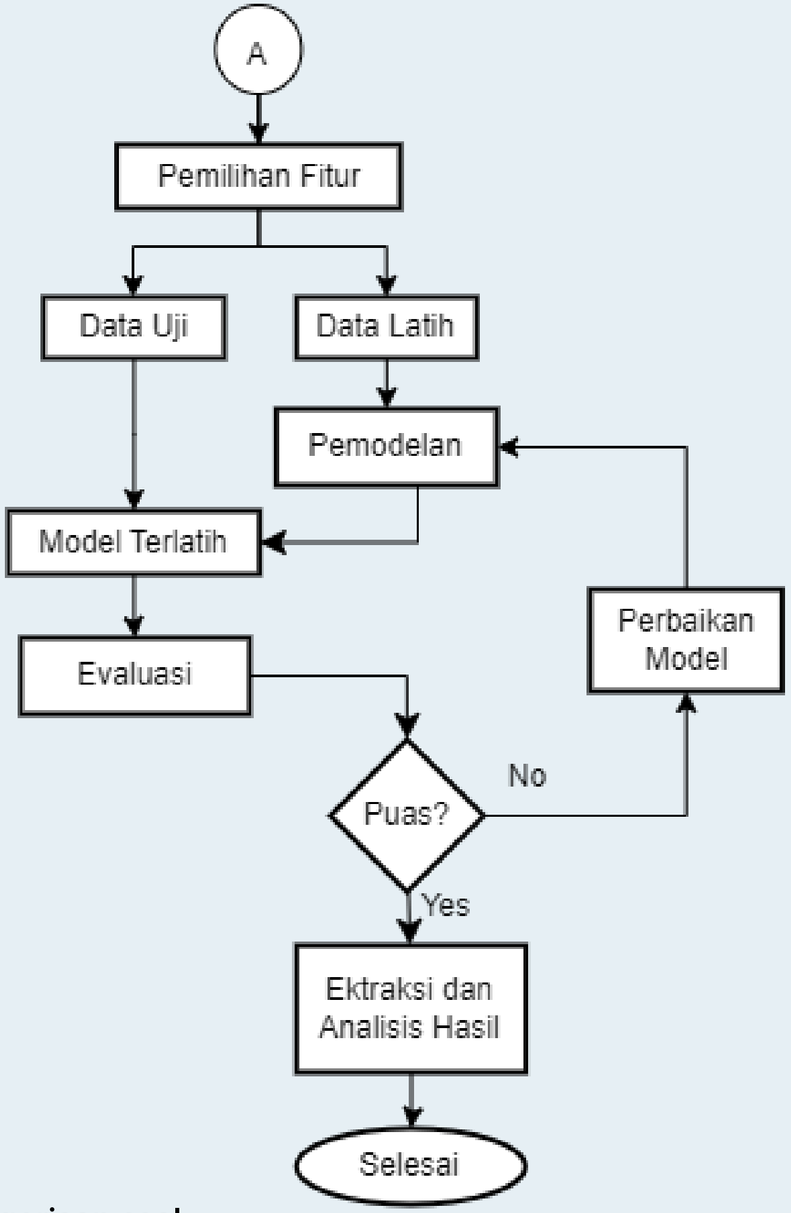
Dataset

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

Diagram Alir

Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta





Gambar 1. Diagram alir pengerjaan soal

Metodologi - Granger Causality Test

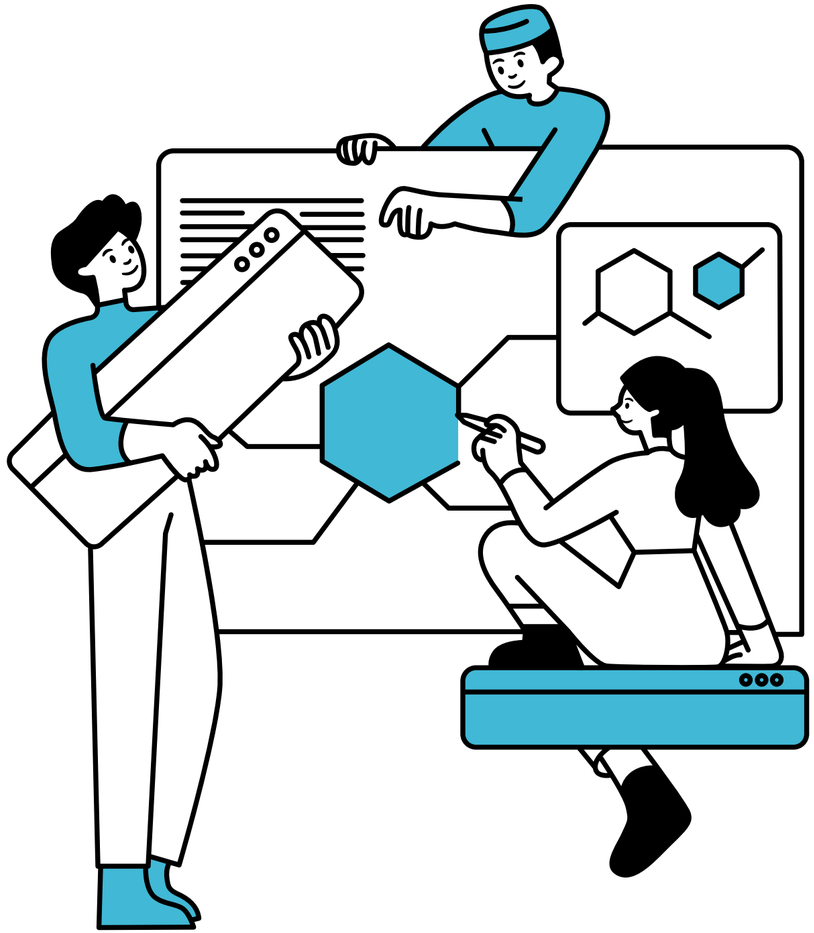
→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

Untuk membuktikan adanya **korelasi antara dua variabel**, misal X dan Y dapat digunakan **Granger causality test**. Uji t-statistik dan F-statistik digunakan dalam pembuktian bahwa variabel X memberikan informasi yang signifikan secara statistik terhadap nilai variabel Y di masa mendatang [7].

Dalam pengerjaan soal ini akan dilakukan pengecekan terhadap korelasi ketiga dataset yang digunakan dengan membuktikan **apakah variabel COVID-19 menyebabkan variabel mobilitas** dan **apakah variabel mobilitas menyebabkan variabel kualitas udara** menggunakan Granger causality test. Hasil tes ini akan dibandingkan dengan pengamatan terhadap tren berdasarkan plot analisis multivariabel pada dataset.

Metodologi - Model Machine Learning

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

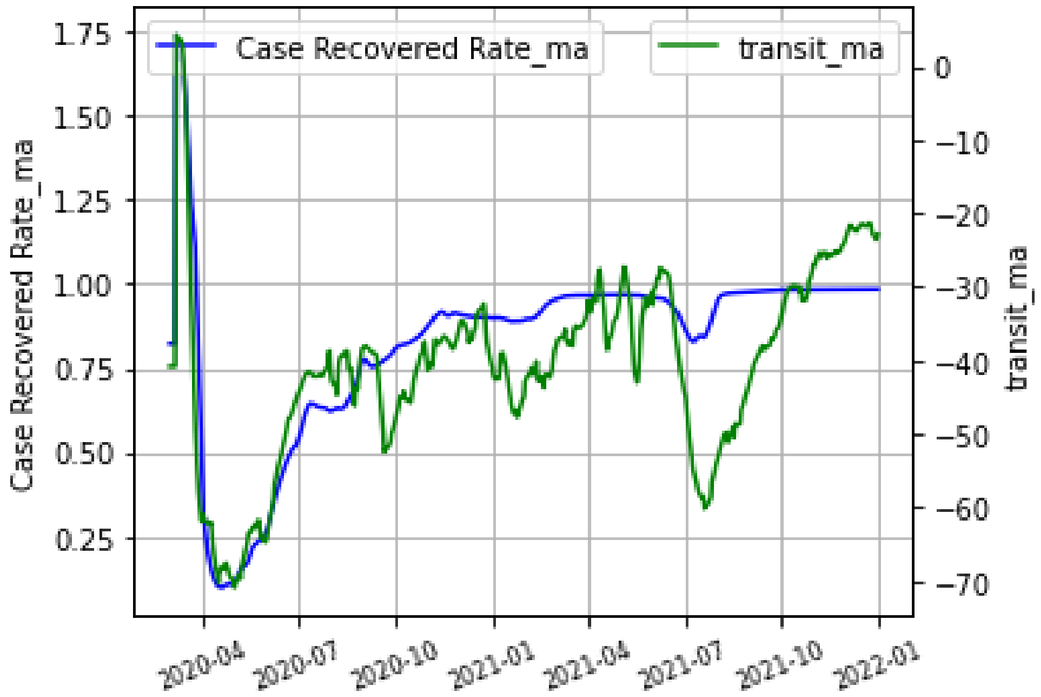


Model Regresi Linear

Regresi linear adalah teknik pemodelan dimana variabel dependen diprediksi berdasarkan satu atau lebih variabel independen. Analisis regresi linier adalah yang paling banyak digunakan dari semua teknik statistik [8].

Model Extreme Gradient Boosting (XGBoost)

XGBoost adalah salah satu implementasi algoritma Gradient Boosted Trees yang paling populer dan efisien, metode supervised learning yang didasarkan pada perkiraan fungsi dengan mengoptimalkan fungsi loss tertentu serta menerapkan beberapa teknik regularisasi [9]. Menurut [10], XGBoost adalah pendekatan yang ampuh untuk membangun model regresi supervised.



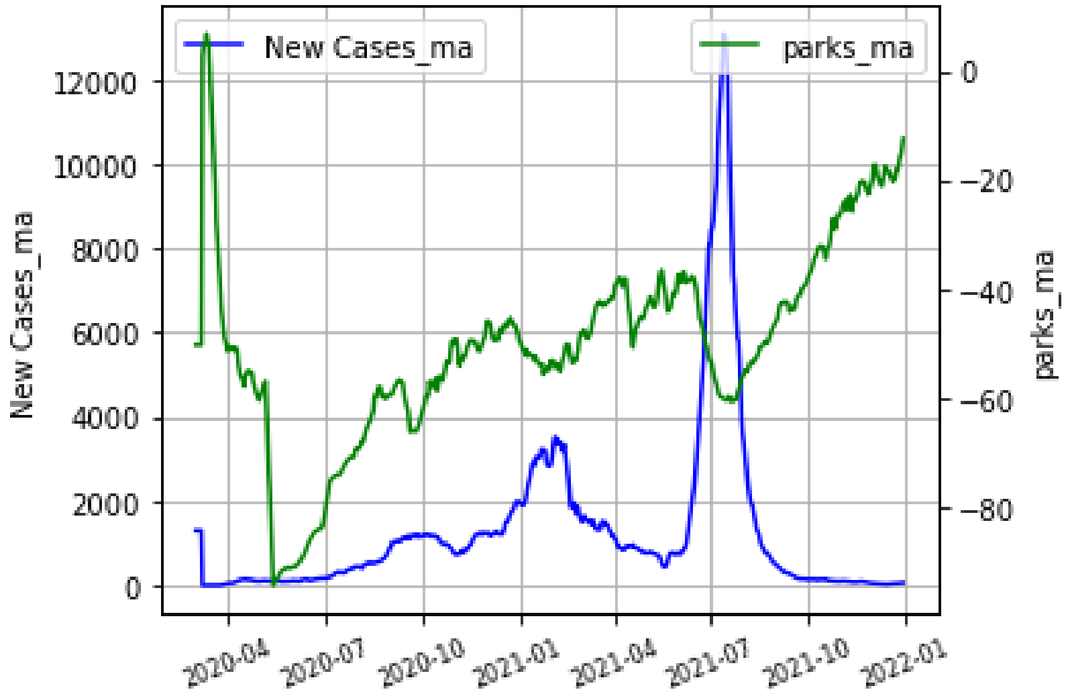
Gambar 2. Plot deret waktu multivariabel antara data Case Recovered Rate dan transit yang keduanya telah diberi perlakuan moving average.

Gambar 2 menunjukkan plot deret waktu multivariabel antara data COVID-19 di kategori **Case Recovered Rate** dan data mobilitas di kategori **transit** yang telah diberi perlakuan **moving average** agar tren dapat terlihat dengan lebih jelas. Kedua variabel dipilih sebagai **contoh** dari data COVID-19 dan mobilitas untuk dianalisis trennya.

Berdasarkan gambar tersebut, sejak awal tren kolom transit terlihat berjalan **cukup selaras** dengan tren kolom Case Recovered Rate. Meskipun pada pertengahan tampak beberapa ketidakstabilan kolom transit yang berbeda dengan kolom Case Recovered Rate, tetapi **perbedaan** tersebut **tidak terlalu signifikan** dan masih sejalur dengan tren penurunan dan kenaikan kolom Case Recovered Rate. Secara umum dapat diindikasikan bahwa kedua kolom **saling berhubungan** dengan jenis **korelasi positif**.

Analisis Awal - Pengenalan Tren

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta



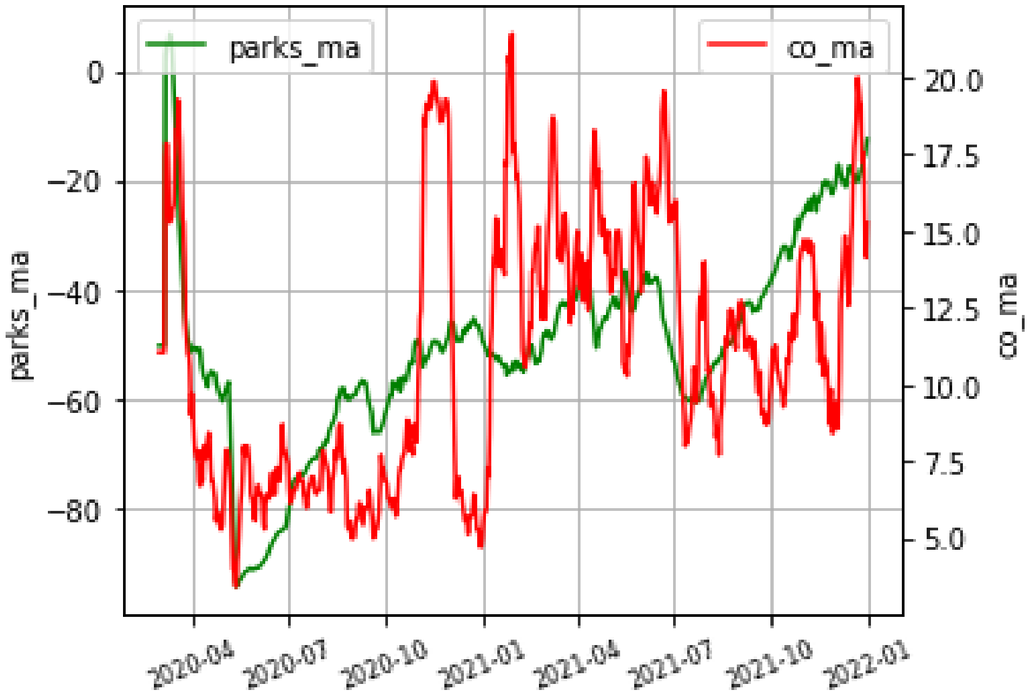
Gambar 3. Plot deret waktu multivariabel antara data New Cases dan parks yang keduanya telah diberi perlakuan moving average.

Gambar 3 menunjukkan plot deret waktu multivariabel antara data COVID-19 di kategori **New Cases** dan data mobilitas di kategori **parks** yang telah diberi perlakuan **moving average** agar tren dapat terlihat dengan lebih jelas. Kedua variabel dipilih sebagai **contoh** dari data COVID-19 dan mobilitas untuk dianalisis trennya.

Di awal-pertengahan tahun 2020 ketika **mulai ada kasus baru** COVID-19 di DKI Jakarta, **mobilitas menurun** drastis bahkan hampir menyentuh angka -100%. Setelah masa itu, angka mobilitas mulai naik kembali. Namun, ketika ada **kenaikan angka New Cases**, plot **mobilitas kembali turun** dan membentuk lembah seperti pada 2020-10, 2021-02, dan 2021-08. Hal ini **mengindikasikan adanya hubungan** antara kedua kolom dengan jenis **korelasi negatif**.

Analisis Awal - Pengenalan Tren

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta



Gambar 4. Plot deret waktu multivariabel antara data parks dan co yang keduanya telah diberi perlakuan moving average.

Gambar 4 menunjukkan plot deret waktu multivariabel antara data mobilitas di kategori **parks** dan data kualitas udara di kategori **co** yang telah diberi perlakuan **moving average** agar tren dapat terlihat dengan lebih jelas. Kedua variabel dipilih sebagai **contoh** dari data mobilitas dan kualitas udara untuk dianalisis trennya.

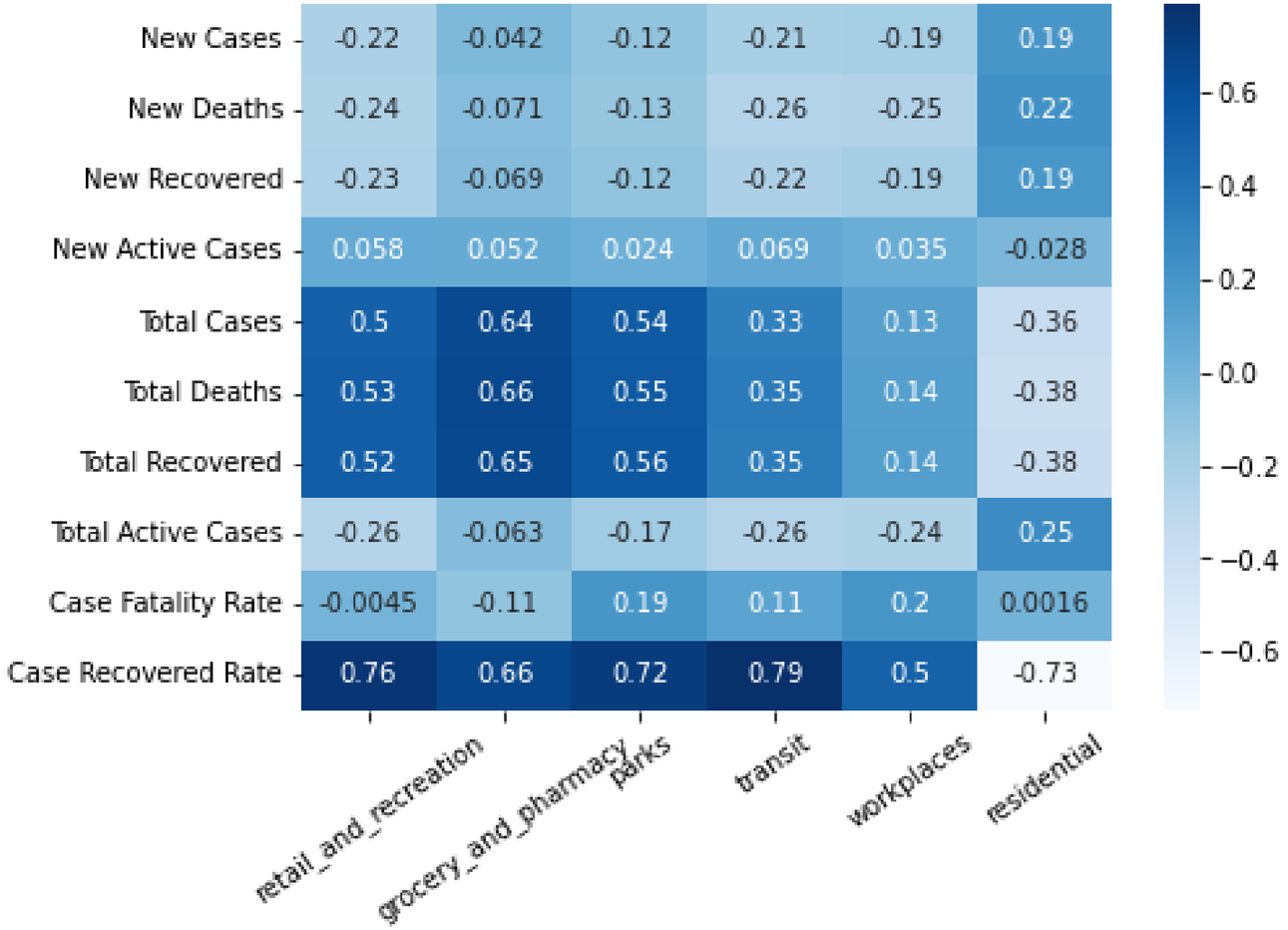
Di awal-pertengahan tahun 2020 ketika **mobilitas menurun drastis**, **angka co juga menurun drastis**. Setelah itu, angka **mobilitas mulai naik kembali** diikuti juga dengan **naiknya angka co**. Pada 2021-07, ketika **penurunan di mobilitas** terjadi, angka **co juga mengalami penurunan**. Hal ini **mengindikasikan** adanya hubungan antara kedua data dengan jenis **korelasi positif**.

Analisis Awal - Pengenalan Tren

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

Analisis - Matriks Korelasi

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta



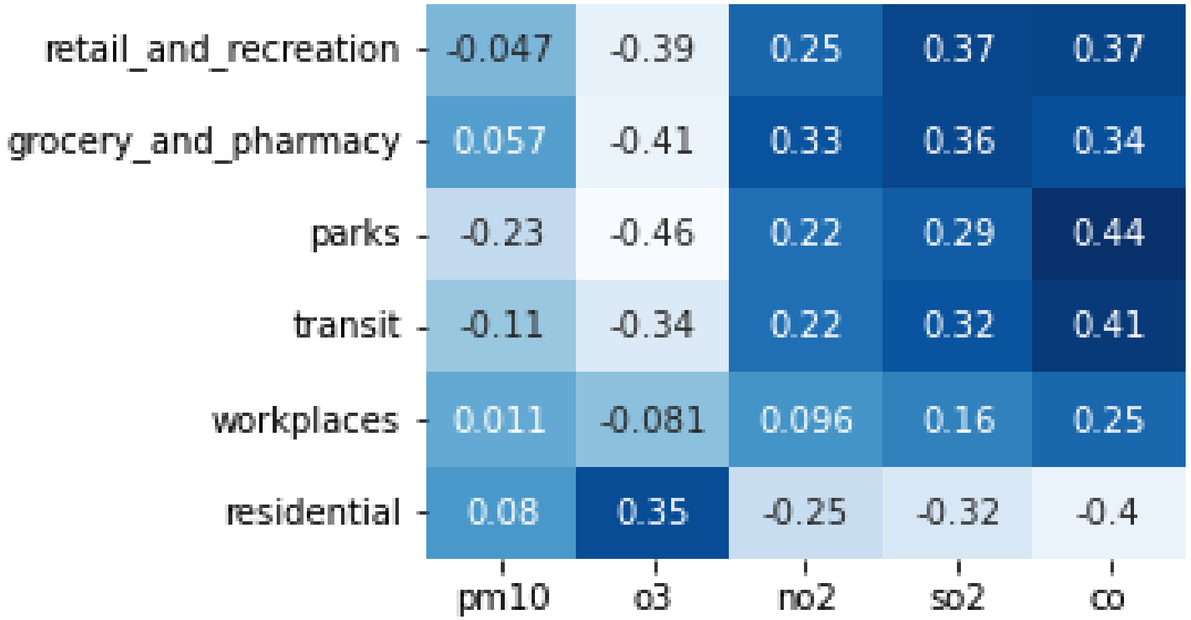
Gambar 5. Matriks korelasi antara data COVID-19 dan data mobilitas.

Gambar 5 menunjukkan matriks korelasi antara data **COVID-19** dan data **mobilitas**. Berdasarkan gambar tersebut, nilai absolut korelasi **tertinggi** pada setiap kolom mobilitas terdapat pada **Case Recovered Rate** dengan jenis **korelasi positif**. Kolom-kolom lainnya juga cenderung memiliki **korelasi positif**, kecuali pada kolom **New Cases, New Death, New Recovered, dan Total Active Cases** yang memiliki **korelasi negatif** dengan data COVID-19.

Selain itu, terdapat **perbedaan** pada kategori **residential** yang memiliki nilai absolut korelasi cenderung sama, tetapi **jenis korelasinya berbeda** dengan kolom mobilitas lainnya. Hal ini mungkin terjadi karena ketika mobilitas di tempat umum/**kategori lain menurun** berarti orang-orang lebih banyak berada di kawasan residential/pemukiman yang **menyebabkan mobilitas di residential meningkat**.

Analisis - Matriks Korelasi

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta



Gambar 6. Matriks korelasi antara data mobilitas dan data parameter kualitas udara.

Gambar 6 menunjukkan matriks korelasi antara data **mobilitas** dan data **kualitas udara**. Berdasarkan gambar tersebut, kategori **parks** memiliki peran penting karena memiliki nilai absolut korelasi **tertinggi** pada tiga parameter kualitas udara yaitu **pm10**, **o3**, dan **co**.

Parameter **pm10** memiliki nilai absolut **korelasi yang cenderung kecil** dan tidak terlalu berpengaruh dibandingkan parameter-parameter lainnya. **Sebagian besar** hubungan di matriks berjenis **korelasi positif**, **kecuali** pada kolom **o3** yang memiliki jenis **korelasi negatif**. Hal ini dimungkinkan karena dengan **banyaknya orang** yang hanya beraktivitas di sekitar **kawasan rumah/residential** maka jumlah **polusi udara** akan semakin **menurun**.

	retail_and_recreation	grocery_and_pharmacy	parks	transit	workplaces	residential
New Cases	4		4	2	2	
New Deaths	6		5		1	
New Recovered					1	
New Active Cases	3		4	7	5	
Total Cases	1	1	1	1	2	
Total Deaths	1	1	1	1	1	
Total Recovered	1	1	1	1	1	1
Total Active Cases	3		5	7	1	1
Case Fatality Rate	1	2	3	2	7	5
Case Recovered Rate	1	1	1	1	1	1

Tabel 1. Hasil Granger causality test antara data COVID-19 dan data mobilitas.

Tabel 1 menunjukkan hasil Granger causality test antara data **COVID-19** dan data **mobilitas.** Berdasarkan tabel tersebut, dapat diambil informasi bahwa **sebagian besar** variabel COVID-19 **menyebabkan (cause)** perubahan pada variabel mobilitas. Meskipun pada kolom **grocery_and_pharmacy**, **residential**, dan **New Recovered** terdapat **cukup banyak** sel berwarna **merah** yang menandakan bahwa variabel COVID-19 **tidak menyebabkan (don't cause)** variabel mobilitas.

Sel Hijau: Variabel COVID-19 **menyebabkan** mobilitas dengan nilai sebagai lag /delay (hari)

Sel Merah: Variabel COVID-19 **tidak menyebabkan** mobilitas

Analisis Lanjut - Granger causality

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

	pm10	o3	no2	so2	co
retail_and_recreation	4	1		1	1
grocery_and_pharmacy	2	1	3	1	1
parks	1	1			1
transit	2	1	3		1
workplaces	2		7		1
residential	2	1	2		1

Tabel 2. Hasil Granger causality test antara data mobilitas dan data kualitas udara.

Tabel 2 menunjukkan hasil Granger causality test antara data **mobilitas** dan data **kualitas udara.** Berdasarkan tabel tersebut, dapat diambil informasi bahwa **sebagian besar** variabel mobilitas **menyebabkan (cause)** perubahan pada variabel kualitas udara. Meskipun pada **kebanyakan** kolom **so2** dan **beberapa** kolom **no2** dan **o3** memiliki sel berwarna **merah** yang menandakan bahwa variabel COVID-19 **tidak menyebabkan (don't cause)** variabel mobilitas.

Sel Hijau: Variabel mobilitas **menyebabkan** kualitas udara dengan nilai sebagai lag /delay (hari)

Sel Merah: Variabel mobilitas **tidak menyebabkan** kualitas udara

Analisis Lanjut - Granger causality

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

Analisis Lanjut - Model Prediksi

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

	linear	xgb	xgb_tuning
retail_and_recreation	56.896685	27.828166	16.506339
grocery_and_pharmacy	62.190291	41.862614	33.226408
parks	145.614830	39.292409	31.366192
transit	59.617229	23.763470	20.378411
workplaces	129.549104	143.786291	112.362952
residential	13.352074	13.144491	10.777151

Tabel 3. Performa MSE model prediksi data mobilitas.

	linear	xgb	xgb_tuning
pm10	144.223763	152.697195	144.186031
o3	241.690309	264.664884	199.772687
no2	109.849908	87.676533	73.291859
so2	144.709981	107.974694	88.890529
co	20.121542	23.042128	19.752860

Tabel 4. Performa MSE model prediksi data kualitas udara.

Tabel 3 dan tabel 4 berturut-turut menunjukkan performa tiap model pada prediksi kolom mobilitas dan polusi udara. Terlihat jelas **penurunan nilai MSE** dari **model linear ke model XGBoost**, serta **XGBoost dengan hyperparameter tuning**. Contohnya pada prediksi kolom "retail_and_recreation" yang mendapati penurunan **MSE** hingga 70%. Penurunan ini dimungkinkan karena metode prediksi dengan model linear tidak cocok diaplikasikan ke beberapa kolom, sedangkan model **XGBoost lebih tidak rentan** dengan perbedaan sifat data.

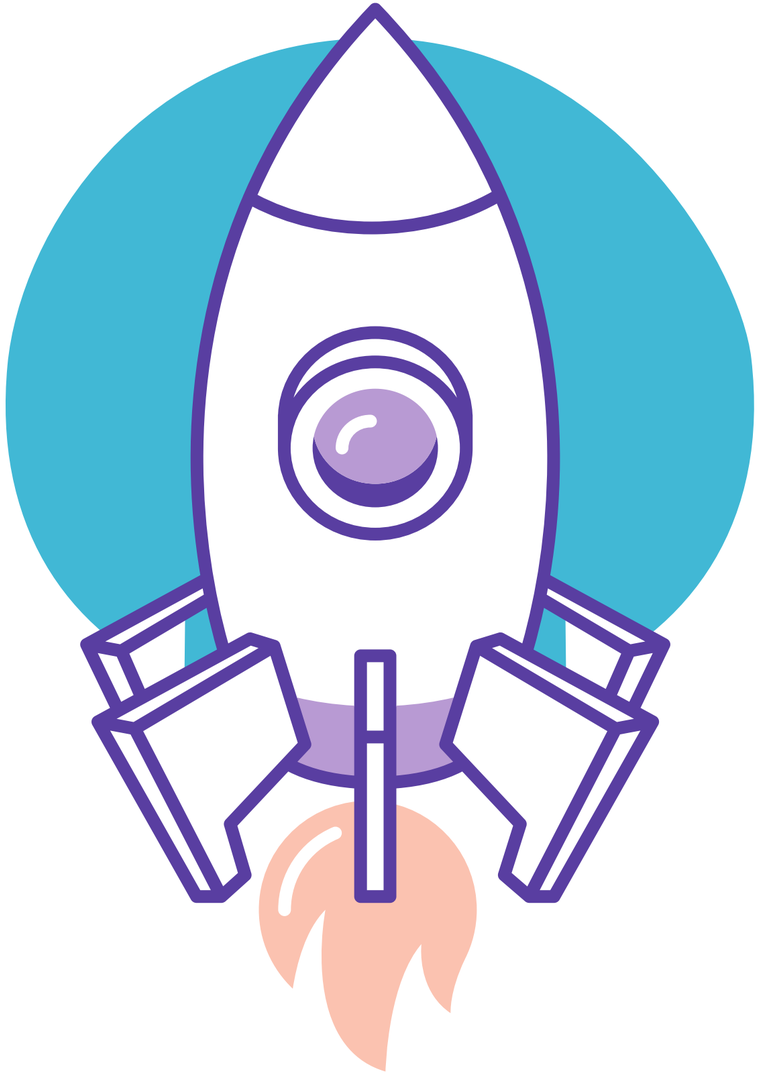
Simpulan

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

- Pandemi **COVID-19 mempengaruhi sebagian besar mobilitas** di wilayah DKI Jakarta, dengan delay atau lag yang berbeda-beda. **Semakin buruk kondisi pandemi**, maka akan **semakin kecil angka mobilitas** di DKI Jakarta.

- **Mobilitas** selama masa pandemi juga **berhubungan** dengan **sebagian besar parameter kualitas udara** di wilayah DKI Jakarta. **Semakin banyak mobilitas** yang terjadi, **semakin meningkat pula angka polutan parameter kualitas udara.**

- Dibandingkan dengan model Linear Regression dan XGBoost dengan parameter default, **XGBoost dengan hyperparameter tuning** menghasilkan **performa yang paling baik** dalam prediksi data mobilitas komunitas dan data kualitas udara di wilayah DKI Jakarta, ditunjukkan dengan hasil



Kurangnya data Indeks Standar Pencemaran Udara (ISPU) DKI Jakarta setelah tahun 2021 menjadi hambatan dalam menganalisis relasi antar dataset pada masa COVID-19 dengan masa post-COVID-19. Oleh karena itu, **penambahan jangka waktu dataset** menjadi salah satu poin penting untuk analisis selanjutnya.

Selain itu, dalam pengerjaan soal ini **penambahan langkah statistik lain sebelum Granger causality test** seperti Augmented Dickey-Fuller (ADF) test, Differencing, maupun Smoothing juga dapat dilakukan [11].

Rekomendasi

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

Referensi

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

[1] D. R. Utomo, “Foto : Kualitas Udara Sebelum dan Sesudah Pandemi Mewabah Dunia,” merdeka.com, Apr. 21, 2020. https://www.merdeka.com/foto/dunia/1169376/20200421202544-kualitas-udara-sebelum-dan-sesudah-pandemi-mewabah-dunia-002-.html (accessed Jul. 14, 2023).

[2] H. D. Atmanti and R. Y. Prakoso, “Polusi Udara Meningkat Pasca Pandemi COVID-19 (Studi Kasus Kota Semarang),” 2021.

[3] Hendratno, “COVID-19 Indonesia Dataset,” www.kaggle.com, 2022. https://www.kaggle.com/datasets/hendratno/covid19-indonesia (accessed Jul. 04, 2023).

[4] Google, “COVID-19 Community Mobility Report,” COVID-19 Community Mobility Report, 2020. https://www.google.com/covid19/mobility/ (accessed Jul. 04, 2023).

[5] Dinas Lingkungan Hidup, “Indeks Standar Pencemaran Udara (ISPU) Tahun 2020 - Open Data Jakarta,” data.jakarta.go.id, 2021. https://data.jakarta.go.id/dataset/indeks-standar-pencemaran-udara-ispu-tahun-2020 (accessed Jul. 04, 2023).

Referensi

→ Model Prediksi dan Analisis Korelasi antara COVID-19, Mobilitas, serta Kualitas Udara di DKI Jakarta

[6] Dinas Lingkungan Hidup, “Indeks Standar Pencemaran Udara (ISPU) Tahun 2021 - Open Data Jakarta,” data.jakarta.go.id, 2021. https://data.jakarta.go.id/dataset/indeks-standar-pencemaran-udara-ispu-tahun-2021 (accessed Jul. 04, 2023).

[7] V. A. Profillidis and G. N. Botzoris, “Chapter 7 - Econometric, Gravity, and the 4-Step Methods,” ScienceDirect, Jan. 01, 2019. https://www.sciencedirect.com/science/article/abs/pii/B9780128115138000078 (accessed Jul. 13, 2023).

[8] K. Kumari and S. Yadav, “(PDF) Linear regression analysis study,” ResearchGate, Jan. 2018. https://www.researchgate.net/publication/324944461_Linear_regression_analysis_study

[9] D. Leventis, “XGBoost Mathematics Explained,” Medium, Jan. 02, 2022. https://dimleve.medium.com/xgboost-mathematics-explained-58262530904a

[10] “XGBoost for Regression,” GeeksforGeeks, Aug. 29, 2020. https://www.geeksforgeeks.org/xgboost-for-regression/

[11] S. Li, “A Quick Introduction On Granger Causality Testing For Time Series Analysis,” Medium, Dec. 23, 2020. https://towardsdatascience.com/a-quick-introduction-on-granger-causality-testing-for-time-series-analysis-7113dc9420d2 (accessed Jul. 14, 2023).

Terima Kasih

Tim Akane

- → Muhammad **Dafa** Wisnu Galih
- → **Nail**faaz
- → Ahmad Wildan Jauharul **Fuad**