GreatEdu
Home for All Education Needs

MSIB
magang dan studi independen bersertifikat
Kampus Merdeka
INDONESIA JAYA

DATA SCIENCE FINAL PROJECT

Analisis dan Prediksi Kualitas Udara di Seoul

YTTA - Mentoring 3

#Greenceleration #MSIB6 #2024

ANGGOTA TIM

Sitanggang Immanuel
Project Manager

Filbert Leonardo
Analyst

Deri Rizky Nugraha
Modeller

Anas Putra Agazy
Visualizer

DAFTAR ISI

Business Understanding

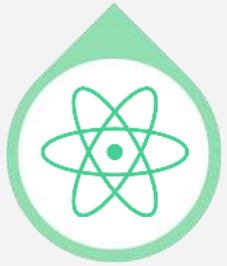
Pembahasan seputar latar belakang masalah dan tujuan yang ingin dicapai.

Data Understanding

Pembahasan seputar deskripsi dataset.

Data Preparation

Pembahasan seputar preprocessing dataset.



Modelling

Pembahasan seputar pembuatan model prediksi.

Evaluation

Pembahasan seputar nilai evaluasi kinerja model.

Deployment

Pembahasan seputar hasil analisis dan aplikasi Streamlit.

Conclusion and Recommendation

Pembahasan seputar kesimpulan dan rekomendasi.

BUSINESS UNDERSTANDING

Business Understanding

Latar Belakang

Kabut asap tebal menyelimuti Seoul, Korea Selatan, menjadi ancaman bagi kesehatan masyarakat. Terpapar polusi udara dalam jangka panjang dapat memicu berbagai penyakit pernapasan, penyakit jantung, dan kanker. Hal ini sangat berbahaya bagi kelompok rentan seperti anak-anak, lansia, dan mereka dengan kondisi kesehatan tertentu.

Permasalahan

Kurangnya data dan informasi yang akurat tentang kualitas udara di tingkat lokal menjadi tantangan utama dalam memerangi polusi udara. Hal ini membuat sulit untuk melacak sumber polusi dan mengembangkan solusi yang efektif.

Tujuan

Tujuan utama pengembangan model Machine Learning ini adalah untuk memprediksi tingkat kualitas udara di Korea Selatan menggunakan data partikel udara (polutan).

Manfaat

- Memberitahu masyarakat kapan mereka mungkin terpapar polusi berbahaya dan kapan mereka perlu melindungi diri.
- Membantu pemerintah dan lembaga lingkungan untuk bertindak cepat dan efektif dalam mengendalikan polusi.
- Membantu mengurangi dampak kesehatan dan ekonomi dari polusi udara.

DATA UNDERSTANDING

Dataset Polusi Udara

FITUR	DESKRIPSI
Measurement date	Tanggal dan waktu saat pengamatan dilakukan
Station code	Kode stasiun tempat pengamatan dilakukan
Address	Alamat tempat pengamatan dilakukan
Latitude	Lintang geografis tempat pengukuran dilakukan
Longitude	Bujur geografis tempat pengukuran dilakukan
SO2	Gas yang terbentuk dari pembakaran bahan bakar fosil yang mengandung sulfur (ppm)
NO2	Gas berwarna coklat kemerahan yang terbentuk dari oksidasi nitrogen oksida (ppm)
O3	Keberadaan ozon di udara (ppm)
CO	Gas tidak berwarna dan tidak berbau hasil pembakaran karbon yang tidak sempurna (ppm)
PM10	Partikel halus dengan diameter ≤ 10 mikrometer di udara (ug/m3)
PM2.5	Partikel halus dengan diameter ≤ 2.5 mikrometer di udara (ug/m3)

Sumber : https://www.kaggle.com/datasets/bappekim/air-pollution-in-seoul

- Pengembangan model prediksi kualitas udara di Seoul diperkuat dengan penggunaan dataset polusi udara yang komprehensif dan terpercaya.
- Dataset ini terdiri dari 647.511 baris data, dengan 11 kolom yang berisi informasi penting tentang konsentrasi polutan udara di 25 stasiun pemantauan di Seoul sejak 2017 - 2019.
- Data tersebut dikumpulkan setiap jam, memberikan gambaran detail tentang fluktuasi kualitas udara di berbagai wilayah dan waktu sepanjang hari.

Definisi CAI (Comprehensive Air-quality Index)

CAI yang juga dikenal sebagai Air Quality Index (AQI), adalah alat yang digunakan untuk mengukur dan melaporkan tingkat keparahan polusi udara. AQI menyajikan informasi kualitas udara dalam angka dan warna yang mudah dipahami, membantu masyarakat untuk:

- Memahami tingkat risiko kesehatan akibat paparan polusi udara.
- Membuat keputusan yang lebih baik terkait aktivitas luar ruangan, seperti berolahraga atau bermain di taman.
- Melindungi diri dan orang-orang terkasih dari efek berbahaya polusi udara.

Description		Good		Moderate		Unhealthy		Very unhealthy	
Values	I_{LO}	0		51		101		251	
	I_{HI}	50		100		250		500	
Concentration		BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}
SO_2(ppm)	1hr	0	0.02	0.021	0.05	0.051	0.15	0.151	1
CO(ppm)	1hr	0	2	2.1	9	9.1	15	15.1	50
O_3(ppm)	1hr	0	0.03	0.031	0.09	0.091	0.15	0.151	0.6
NO_2(ppm)	1hr	0	0.03	0.031	0.06	0.061	0.2	0.201	2
PM_{10}($\mu g/m^3$)	24hr	0	30	31	80	81	150	151	600
$PM_{2.5}$($\mu g/m^3$)	24hr	0	15	16	35	36	75	76	500

Sumber : https://airkorea.or.kr/eng/khaiInfo?PMENU_NO=166

Description	Good	Moderate	Unhealthy	Very unhealthy	
values	0~50	51~100	101~250	251~350	351~500
Health Effects	A level that will not impact patients suffering from diseases related to air pollution	A level which may have a meager impact on patients in case of chronic exposure	A level that may have harmful impacts on patients and members of sensitive groups (children, aged or weak people), and also cause the general public unpleasant feelings	A level which may have a serious impact on patients and members of sensitive groups in case of acute exposure, and that even the general public can be weakly affected	A level which may need to take emergency measures for patients and members of sensitive groups and have harmful impacts on the general public

Kategori Indeks Kualitas Udara (AQI)

DATA PREPARATION

Data Cleansing

	Polutan	Jumlah Data Anomali	Persentase (%)
0	SO2	3976	0.614044
1	NO2	3834	0.592113
2	O3	4059	0.626862
3	CO	4036	0.623310
4	PM10	3962	0.611881
5	PM2.5	3973	0.613580

Data Anomali

```
The number of duplicated value in data:
 0
```

Data Duplikat

	Missing Values	% of Missing Values
Date	0	0.00
Station code	0	0.00
District	0	0.00
Latitude	0	0.00
Longitude	0	0.00
SO2	0	0.00
NO2	0	0.00
O3	0	0.00
CO	0	0.00
PM10	0	0.00
PM2.5	0	0.00

Data Hilang

- Terdapat nilai anomali berupa nilai negatif (>0) pada data, yang kemudian diubah menjadi NAN.
- Untuk menangani data hilang tersebut, kami menggunakan metode imputasi dengan nilai median dari masing-masing polutan udara.
- Setelah itu dilakukan pengecekan data duplikat dan tidak ditemukan adanya data duplikat.

Data Transformation (1/2)

	Date	Station code	District	Latitude	Longitude	SO2	NO2	O3	CO	PM10	PM2.5
0	2017-01-01 00:00:00	101	Jongno-gu	37.57	127.01	0.00	0.06	0.00	1.20	73.00	57.00
1	2017-01-01 01:00:00	101	Jongno-gu	37.57	127.01	0.00	0.06	0.00	1.20	71.00	59.00
2	2017-01-01 02:00:00	101	Jongno-gu	37.57	127.01	0.00	0.06	0.00	1.20	70.00	59.00
3	2017-01-01 03:00:00	101	Jongno-gu	37.57	127.01	0.00	0.06	0.00	1.20	70.00	58.00
4	2017-01-01 04:00:00	101	Jongno-gu	37.57	127.01	0.00	0.05	0.00	1.20	69.00	61.00

Konversi Kolom Address ke District

Konversi ini dilakukan untuk memudahkan analisa geografis untuk masing-masing daerah.

	Date	SO2	NO2	O3	CO	PM10	PM2.5	District
0	2017-01-01	0.00	0.05	0.00	1.03	81.67	67.29	Jongno-gu
1	2017-01-02	0.01	0.05	0.02	1.02	112.17	87.17	Jongno-gu
2	2017-01-03	0.01	0.05	0.01	0.83	72.92	51.83	Jongno-gu
3	2017-01-04	0.01	0.06	0.01	0.99	51.12	34.92	Jongno-gu
4	2017-01-05	0.00	0.04	0.01	0.62	34.88	21.96	Jongno-gu

Konversi Data Per Jam ke Harian

Konversi ini dilakukan untuk mempersingkat waktu yang dibutuhkan dalam melatih model.

	Date	SO2	NO2	O3	CO	PM10	PM2.5	AQI	AQI Category	Station code	District	Latitude	Longitude
0	2017-01-01	0.00	0.05	0.00	1.03	81.67	67.29	220.55	Unhealthy	101	Jongno-gu	37.57	127.01
1	2017-01-02	0.01	0.05	0.02	1.02	112.17	87.17	251.03	Very unhealthy	101	Jongno-gu	37.57	127.01
2	2017-01-03	0.01	0.05	0.01	0.83	72.92	51.83	161.49	Unhealthy	101	Jongno-gu	37.57	127.01
3	2017-01-04	0.01	0.06	0.01	0.99	51.12	34.92	99.79	Moderate	101	Jongno-gu	37.57	127.01
4	2017-01-05	0.00	0.04	0.01	0.62	34.88	21.96	67.90	Moderate	101	Jongno-gu	37.57	127.01

Perhitungan dan Kategorisasi AQI

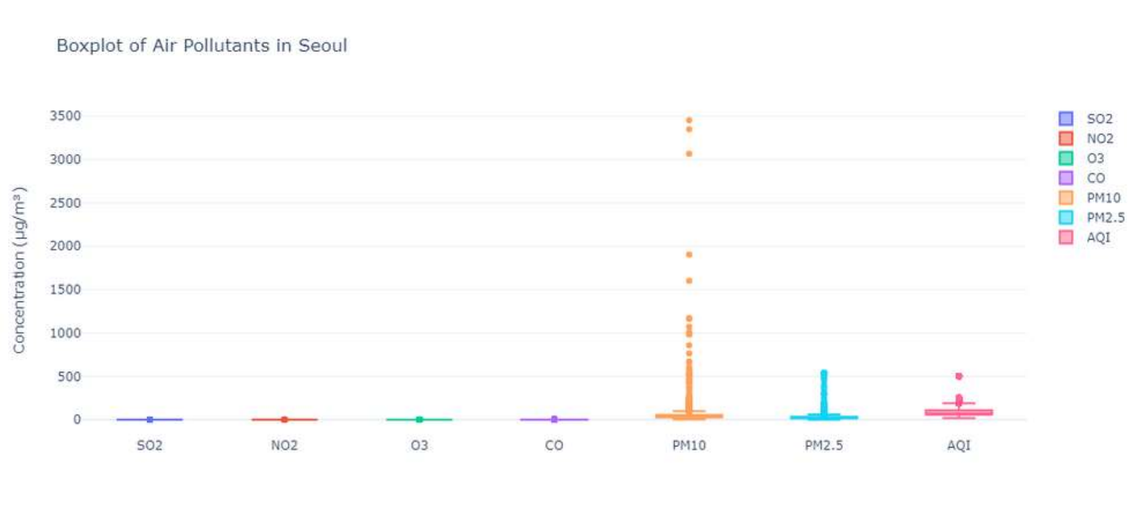
Telah dilakukan penambahan data berupa data AQI dan Kategori AQI yang dimana Kategori AQI ini akan digunakan sebagai target dalam model machine learning.

	mean	std	min	25%	50%	75%	max
SO2	0.004000	0.003000	0.001000	0.003000	0.004000	0.005000	0.230000
NO2	0.029000	0.029000	0.003000	0.019000	0.026000	0.037000	3.698000
O3	0.024000	0.020000	0.001000	0.014000	0.023000	0.033000	2.395000
CO	0.523000	0.235000	0.100000	0.375000	0.479000	0.622000	15.117000
PM10	44.386000	50.101000	1.625000	25.583000	37.750000	54.625000	3454.583000
PM2.5	25.810000	21.210000	1.042000	13.833000	21.208000	31.875000	542.375000
AQI	113.480000	113.123000	16.670000	56.070000	73.350000	108.885000	500.000000

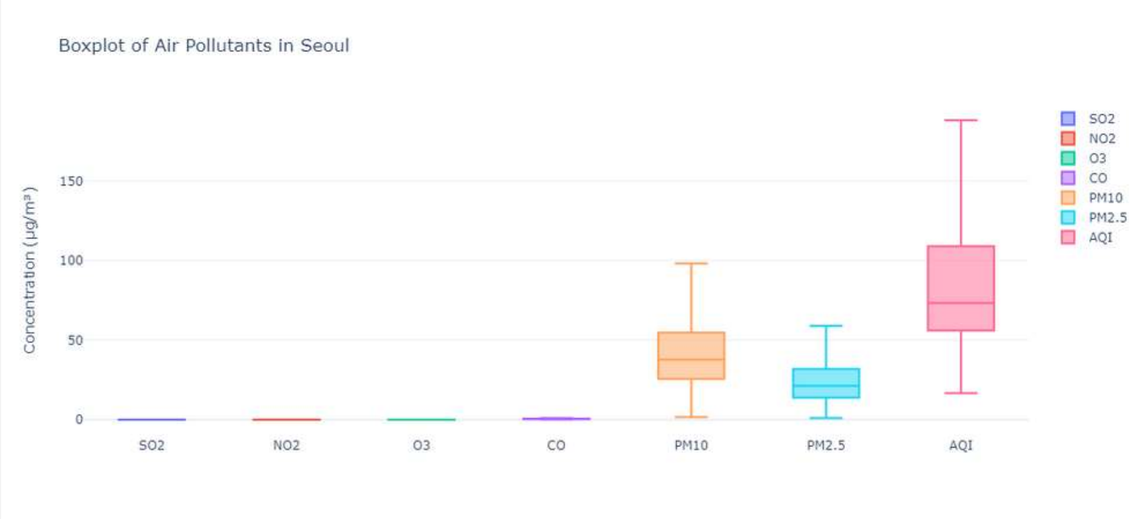
Statistik Data Numerikal

Analisis ini menemukan beberapa nilai maksimum yang dianggap sebagai outlier dan perlu ditangani untuk mencegahnya memengaruhi hasil model secara signifikan.

Data Transformation (2/2)



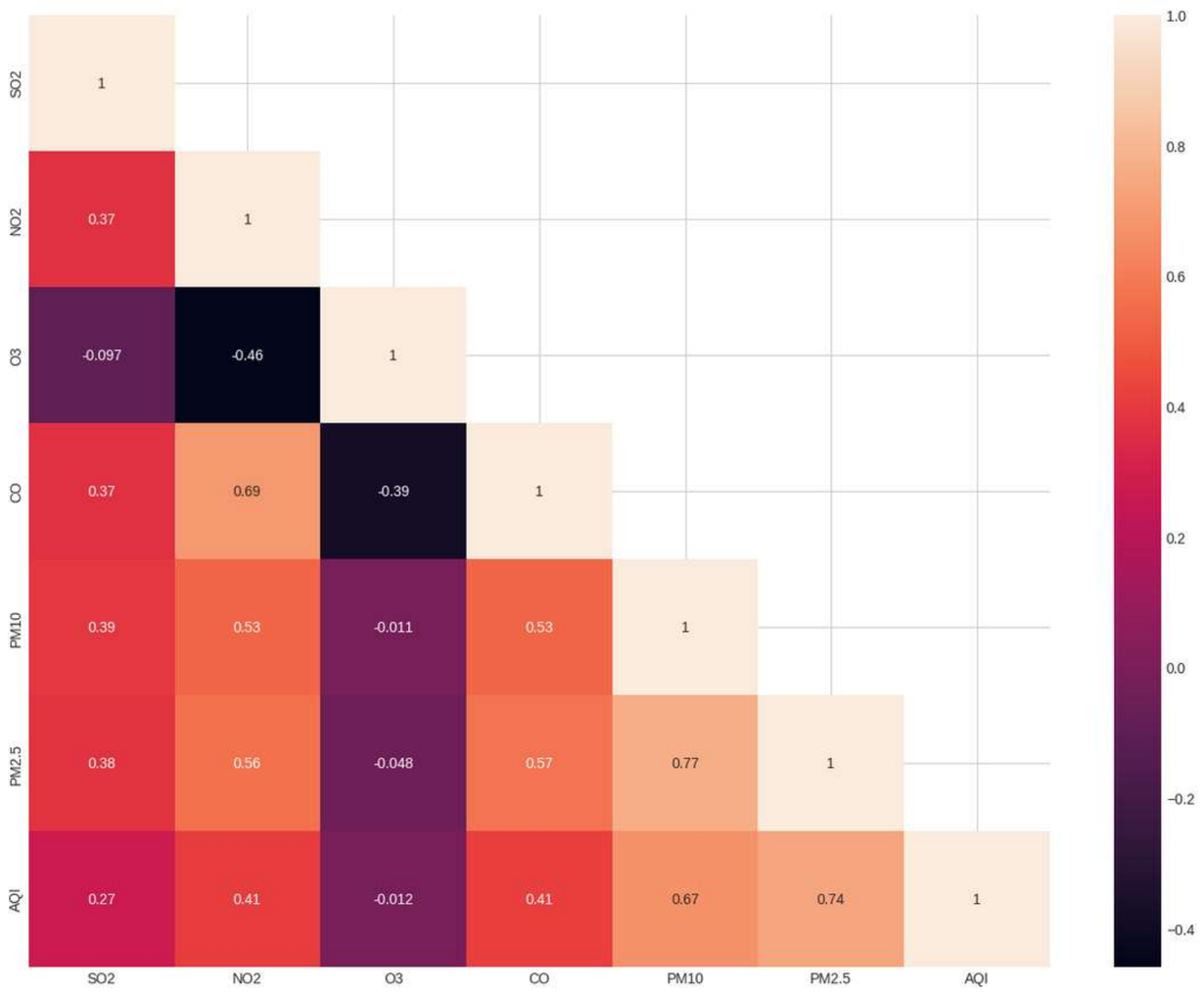
Data Outlier (Sebelum)



Data Outlier (Setelah)

Analisis data polutan udara menunjukkan adanya nilai-nilai outlier yang signifikan, yang dapat mengganggu interpretasi dan akurasi model. Untuk menangani ini, kami menggunakan metode Interquartile Range (IQR) untuk menyesuaikan data outlier. Dengan pendekatan ini, nilai-nilai yang berada di luar batas atas dan batas bawah IQR disesuaikan ke batas tersebut, sehingga menghasilkan distribusi data yang lebih representatif dan meningkatkan keandalan serta akurasi hasil pemodelan.

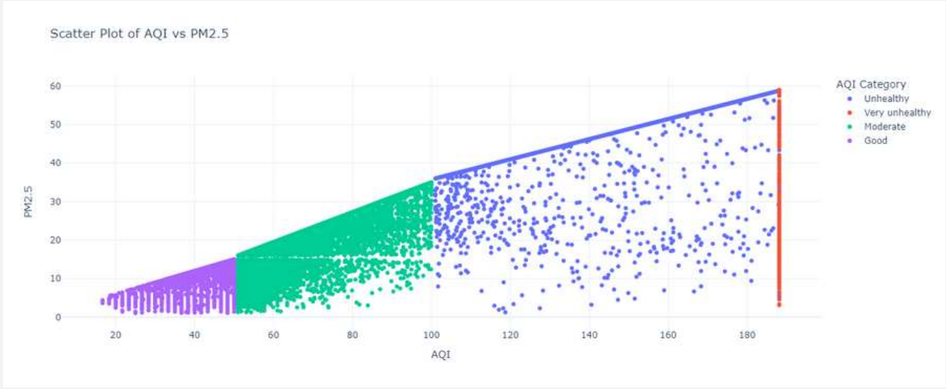
Data Visualization (1/2)



Korelasi Matriks antar Variabel

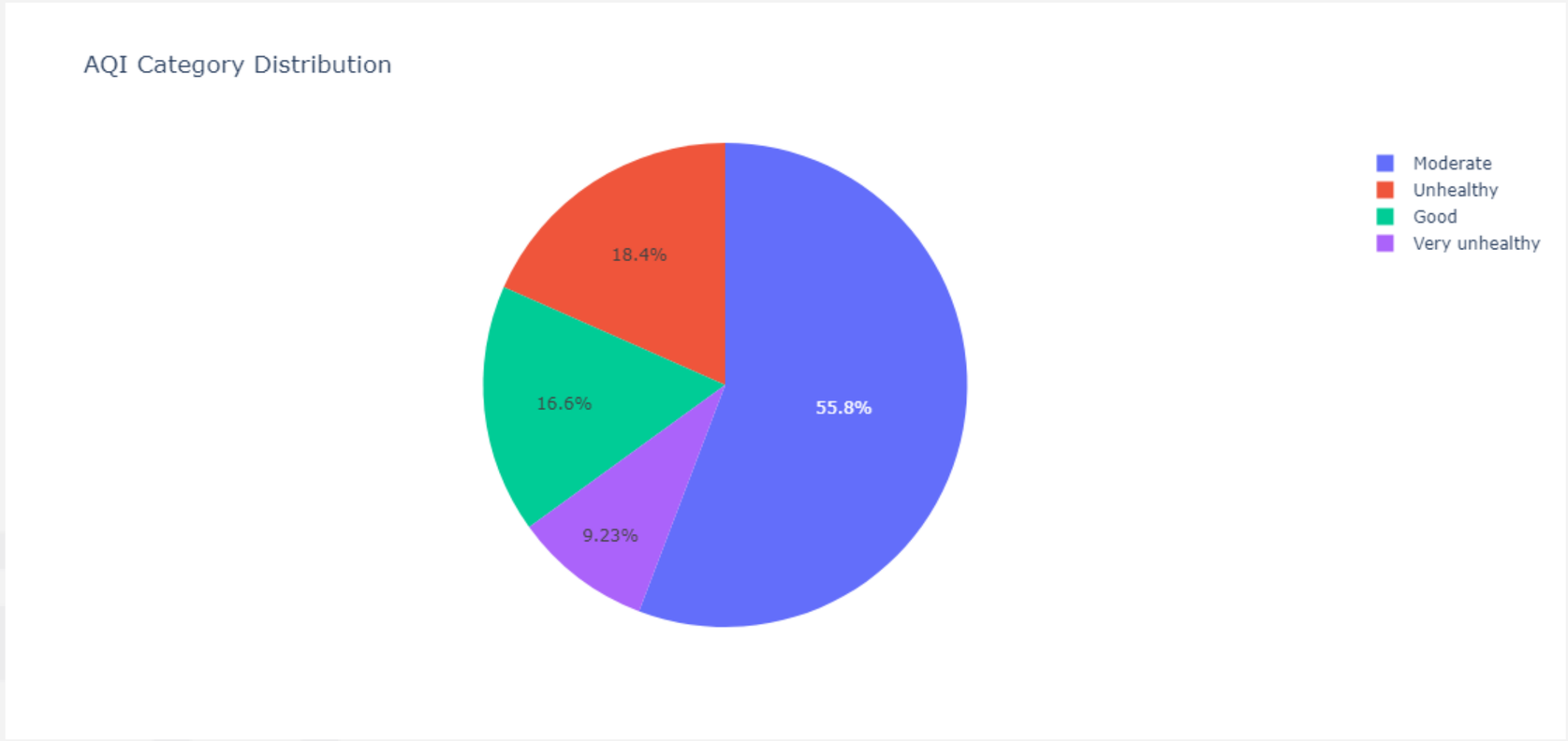
Hubungan Polusi Udara dan AQI

1. **PM2.5 dan PM10:** Memiliki korelasi kuat (0,74 dan 0,67) dengan AQI yang merupakan kontributor utama buruknya kualitas udara. Konsentrasi yang lebih tinggi meningkatkan nilai AQI.
2. **O3:** Berbeda dengan parameter lain, memiliki korelasi negatif (-0,012) dengan AQI dan semua parameter polusi udara.



Korelasi PM2.5 dan AQI

Data Visualization (2/2)



Distribusi Kategori AQI

Visualisasi tersebut menunjukkan :

- Kategori "Moderate" mendominasi dengan persentase mencapai 55,8%, jauh melebihi persentase kategori lainnya.
- Ketidakseimbangan data yang signifikan ini dapat menimbulkan potensi bias dalam pemodelan Machine Learning ketika kategori AQI digunakan sebagai variabel respon/dependen.

Hal ini dapat menghasilkan model yang kurang akurat dalam memprediksi kualitas udara untuk kategori AQI yang kurang terwakili.

MODELLING

AQI CATEGORY

Skala yang digunakan untuk melaporkan kualitas udara harian, dengan kategori yang menunjukkan tingkat bahaya terhadap kesehatan manusia.

Fitur & Target

SO2
Gas beracun yang dihasilkan dari pembakaran bahan bakar fosil yang mengandung sulfur.

NO2
Gas polutan yang dihasilkan dari pembakaran kendaraan bermotor dan industri.

CO
Gas tidak berwarna dan tidak berbau yang dihasilkan dari pembakaran tidak sempurna bahan bakar fosil.

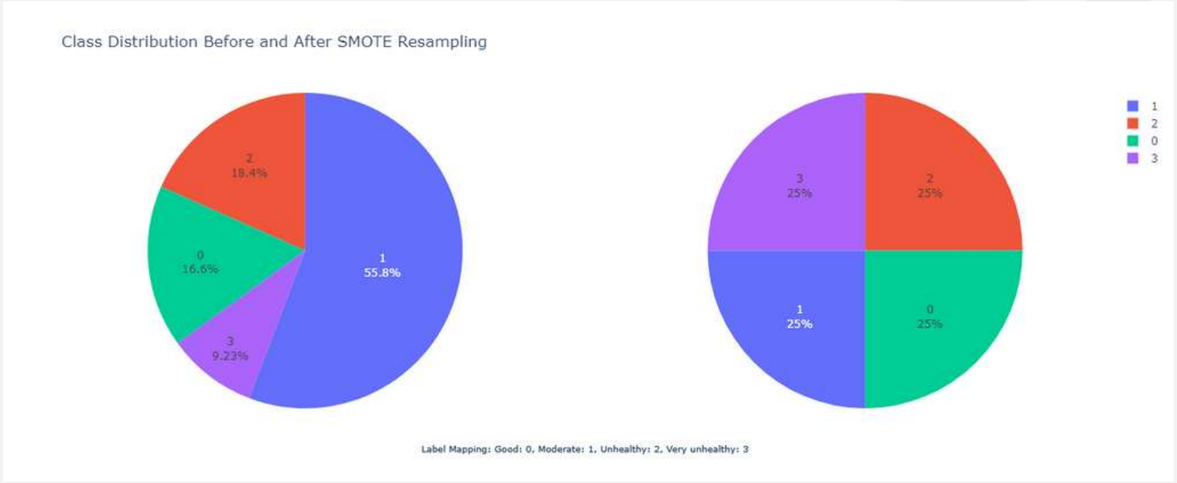
O3
Gas yang terbentuk dari reaksi kimia antara polutan di atmosfer.

PM10
Partikel udara berukuran kurang dari 10 mikrometer yang dapat masuk ke saluran pernapasan atas.

PM2.5
Partikel udara berukuran kurang dari 2.5 mikrometer yang dapat masuk ke paru-paru dan aliran darah.

Feature Engineering

	SO2	NO2	O3	CO	PM10	PM2.5	AQI_Category_Label
0	0.004	0.054	0.003	1.029	81.667	67.292	2
1	0.005	0.046	0.016	1.021	112.167	87.167	3
2	0.005	0.047	0.014	0.833	72.917	51.833	2
3	0.005	0.058	0.009	0.992	51.125	34.917	1
4	0.004	0.041	0.011	0.621	34.875	21.958	1



Label Encoding

Teknik Label Encoding diterapkan untuk mengonversi data kategorik menjadi data numerik karena model machine learning hanya dapat dilatih dengan data numerik.

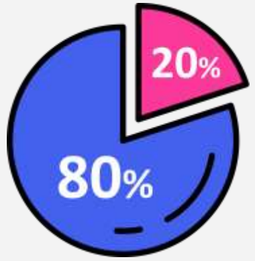
Penyeimbangan Kelas (SMOTE)

Teknik SMOTE diterapkan untuk meningkatkan jumlah data pada kategori minoritas, sehingga distribusi data menjadi seimbang dan tidak bias dalam model.

Modelling

DATA TRAIN

Jumlah data train sebanyak 46.460 data



DATA TEST

Jumlah data test sebanyak 11.616 data

LightGBM

Akurasi mencapai
97,4%

CatBoost

Akurasi mencapai
97,5%

XGBoost

Akurasi mencapai
98,1%

Random Forest Classifier

Akurasi mencapai
98,4%

5

EVALUATION

Model Evaluation

Analysis Model Random Forest Classifier

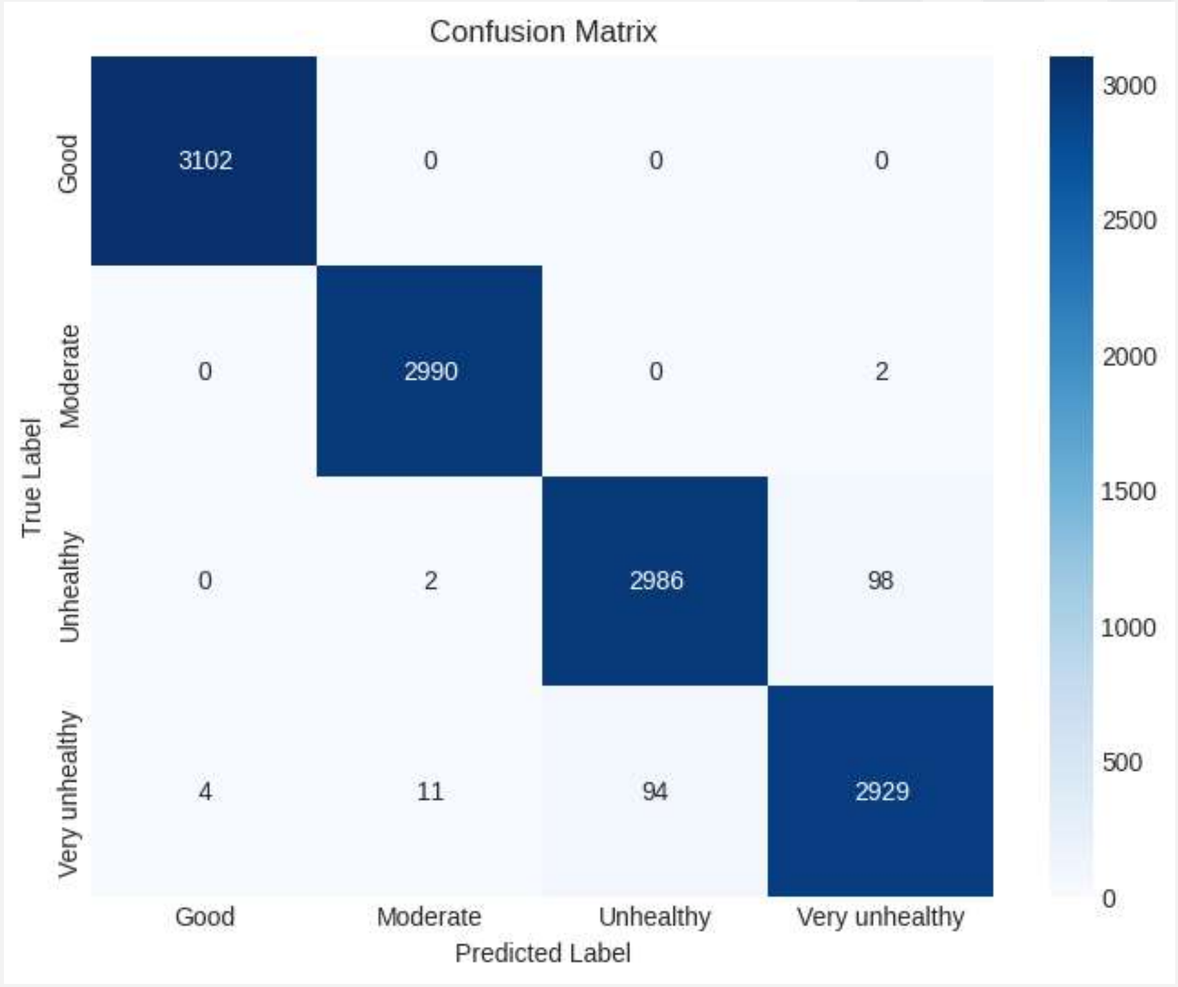
Analysis Metric Error Model Testing

	precision	recall	f1-score	support
Good	1.00	1.00	1.00	3102
Moderate	1.00	1.00	1.00	2992
Unhealthy	0.97	0.97	0.97	3086
Very unhealthy	0.97	0.97	0.97	3038
accuracy			0.98	12218
macro avg	0.98	0.98	0.98	12218
weighted avg	0.98	0.98	0.98	12218

Nilai Evaluasi Metriks

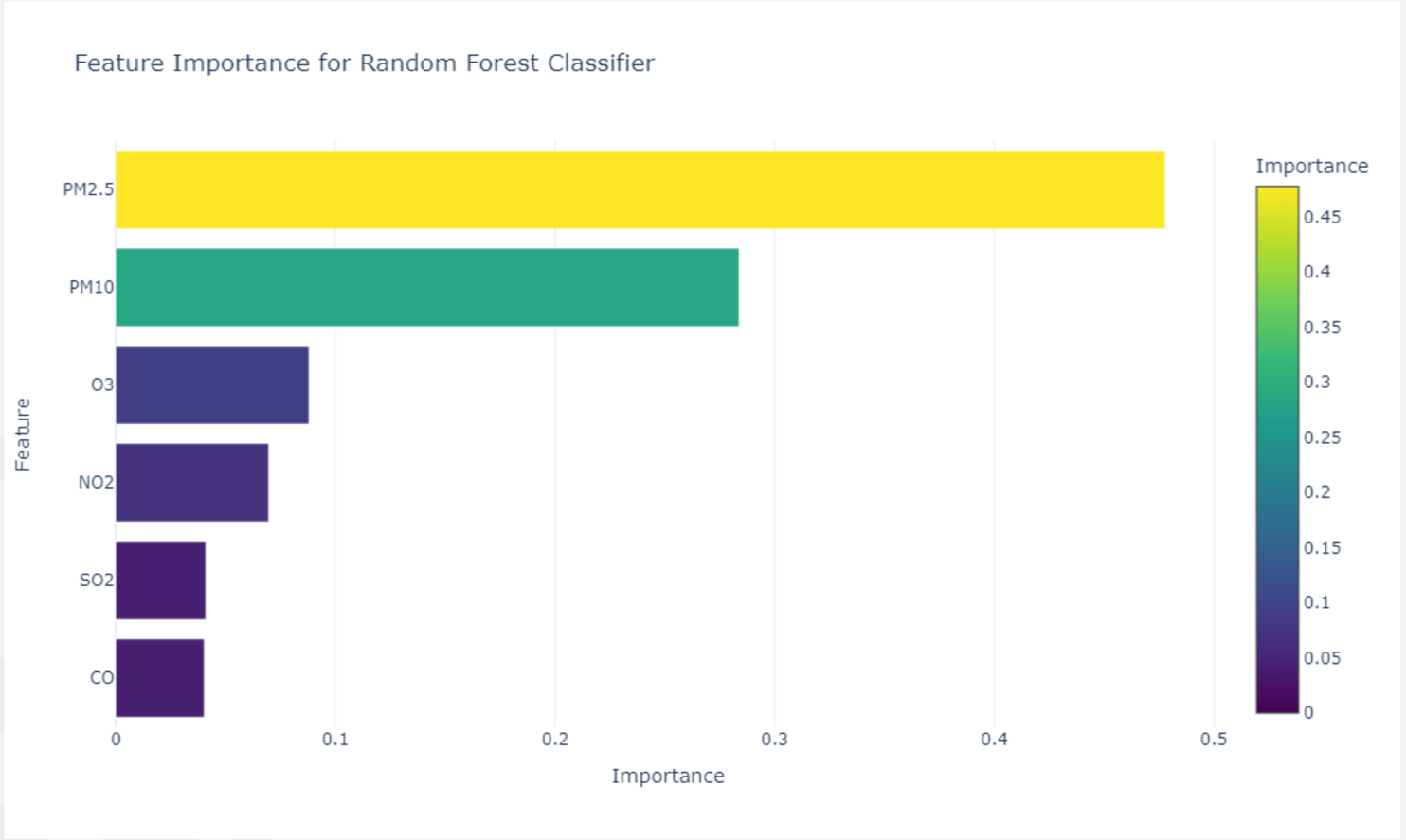
Hasil Pemodelan Kualitas Udara Seoul

- **Model Terbaik:** Random Forest Classifier
- **Performa:**
  - **Akurasi:** 98%
  - **Recall:** 98%
  - **Precision:** 98%
  - **F1 Score:** 98%



Confusion Matrix

Feature Importance

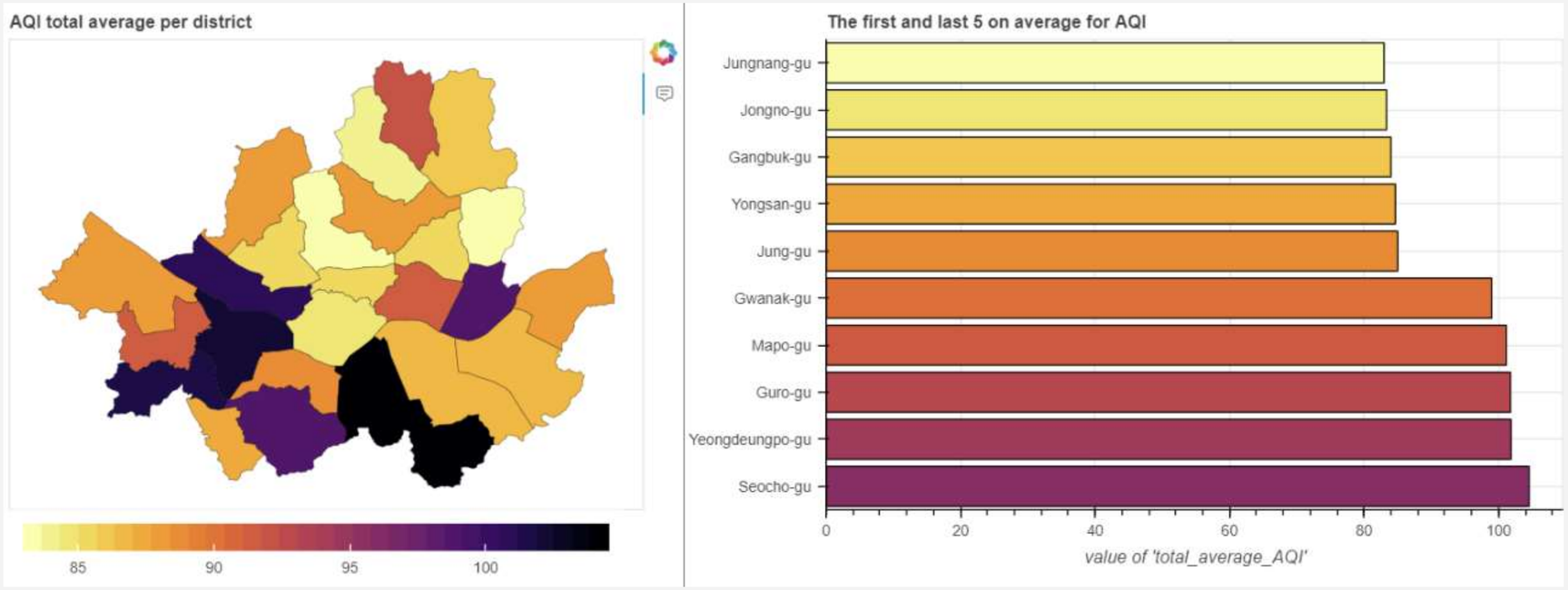


Analisis Feature Importance:

- **PM2.5:** Memiliki pengaruh paling signifikan (>0,4), menunjukkan bahwa polutan ini merupakan faktor utama dalam menentukan kategori AQI.
- **CO dan SO2:** Memiliki pengaruh paling rendah (<0,1), menunjukkan bahwa polutan ini memiliki pengaruh yang lebih kecil.

DEPLOYMENT

Hasil Analisa Kualitas AQI (1/2)



Peta Daerah dengan Tingkat Kualitas Udara Terbaik dan Terburuk

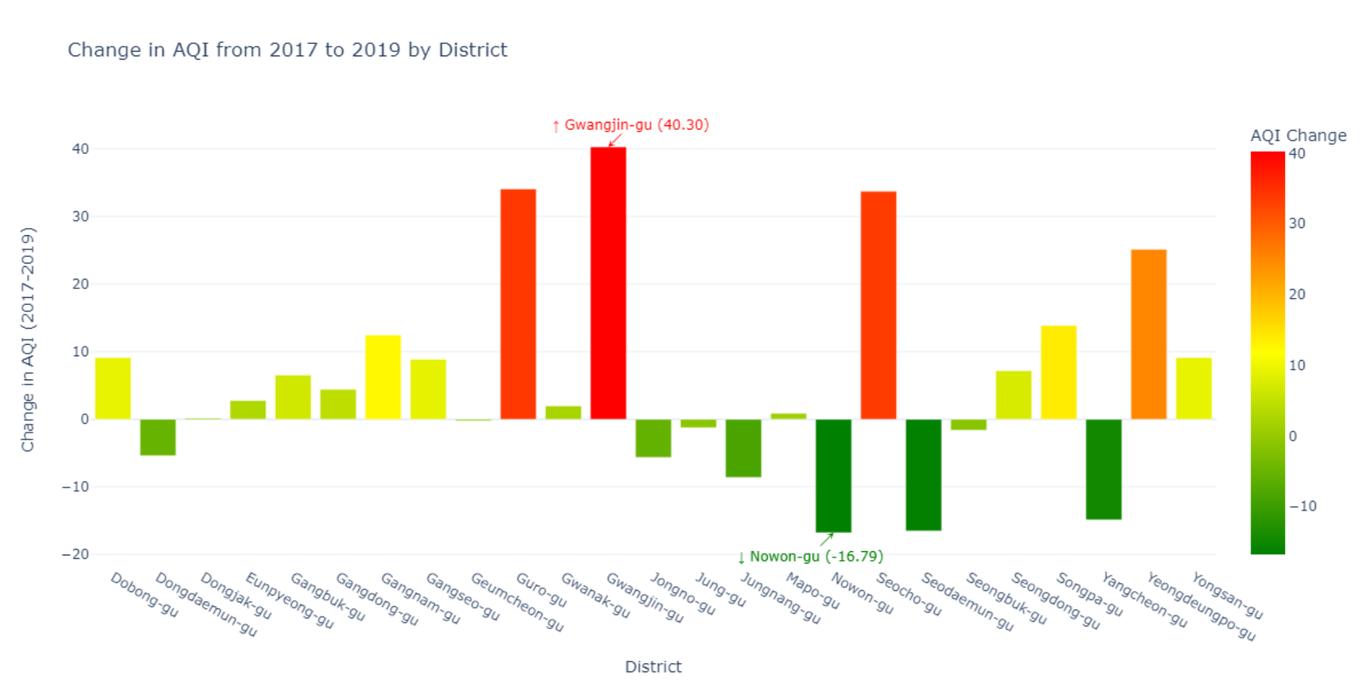
Kualitas Udara di Lima Daerah Memprihatinkan

- Seocho-gu, Yeongdeungpo-gu, Guro-gu, Mapo-gu, dan Gwanak-gu memiliki tingkat polusi parah dengan konsentrasi AQI tinggi (daerah berwarna gelap pada peta).
- Kualitas udara di daerah-daerah ini membahayakan kesehatan masyarakat.

Lima Daerah dengan Kualitas Udara Baik

- Jungnang-gu, Jongno-gu, Gangbuk-gu, Yongsan-gu, dan Jung-gu memiliki AQI terendah (daerah berwarna terang pada peta).
- Kualitas udara di daerah-daerah ini lebih sehat dan aman untuk beraktivitas.

Hasil Analisa Kualitas AQI (2/2)



Tren Perubahan AQI

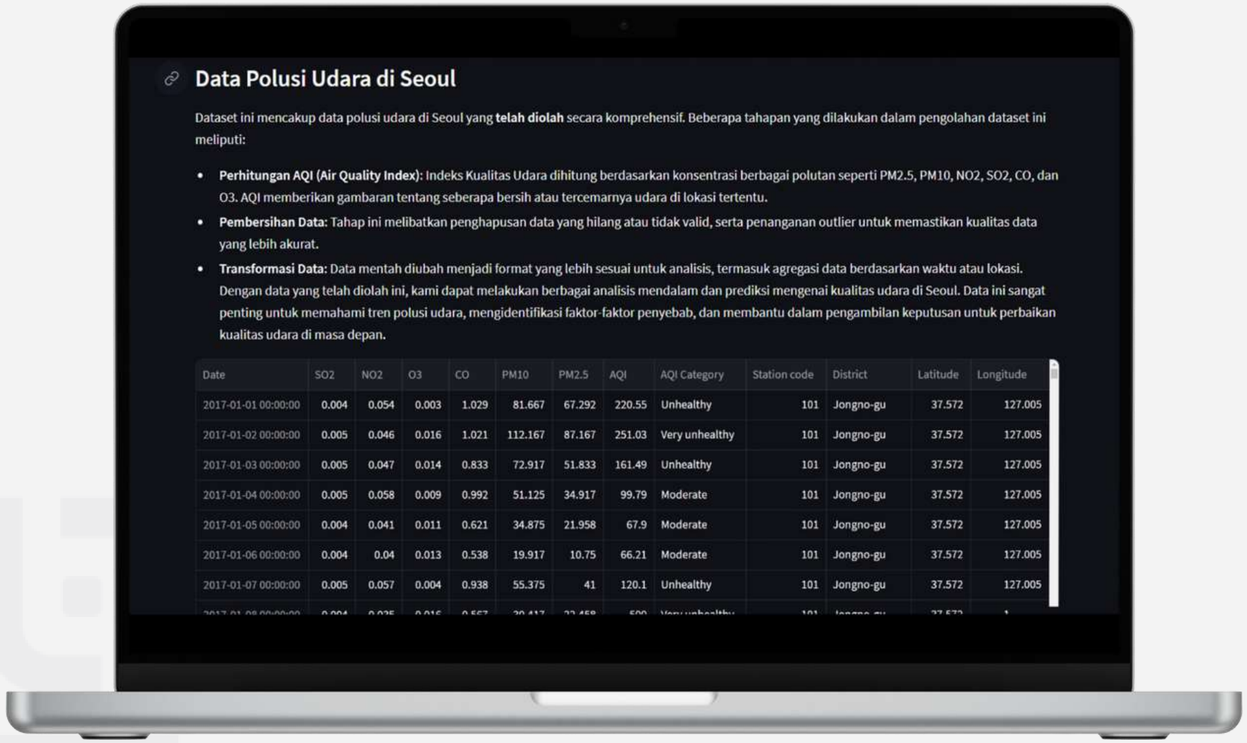
Nowon-gu Mengalami Peningkatan Kualitas Udara

- Distrik Nowon-gu memiliki rata-rata AQI terendah pada tahun 2018 dengan nilai 93,53.
- Nowon-gu mengalami penurunan AQI terbesar dari 2017 hingga 2019, dengan penurunan sebesar -16.79 poin AQI.
- Hal ini menunjukkan upaya untuk meningkatkan kualitas udara di Nowon-gu membuahkan hasil.

Gwangjin-gu Mengalami Penurunan Kualitas Udara

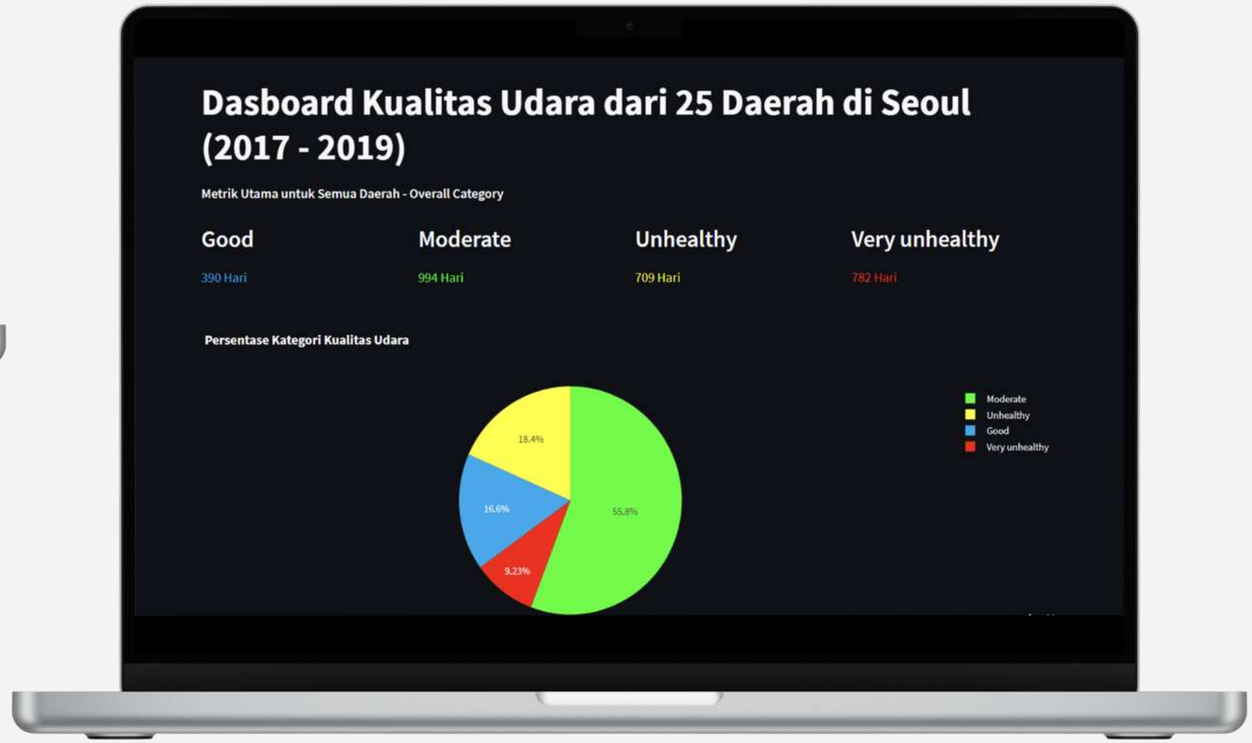
- Distrik Gwangjin-gu mengalami peningkatan AQI terbesar dari 2017 hingga 2019, dengan peningkatan sebesar 40.30 poin AQI.
- Hal ini menunjukkan kualitas udara di Gwangjin-gu memburuk dan perlu mendapat perhatian serius.

Dashboard Streamlit (1/2)

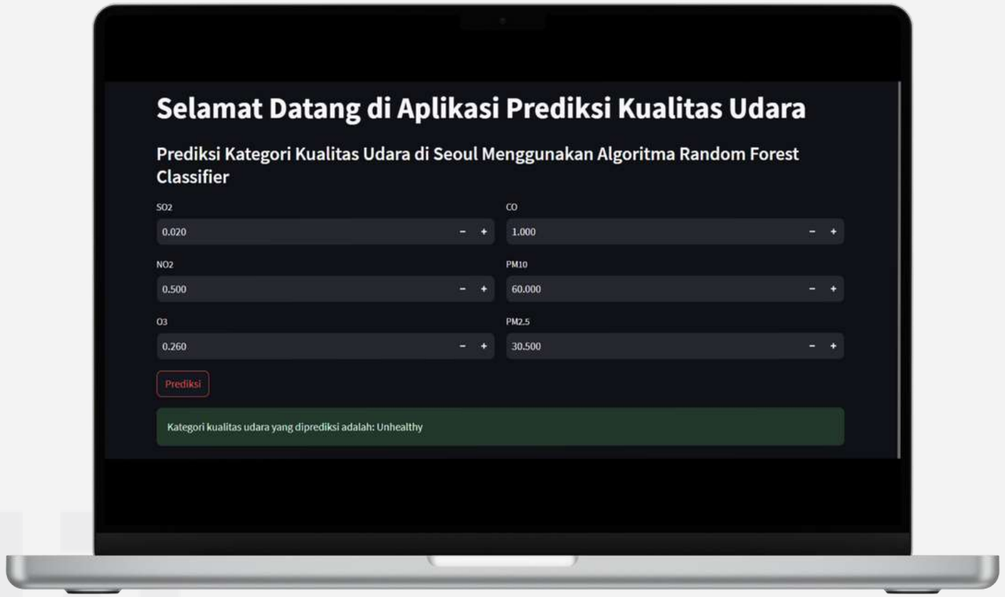


Menu Home

Menu EDA

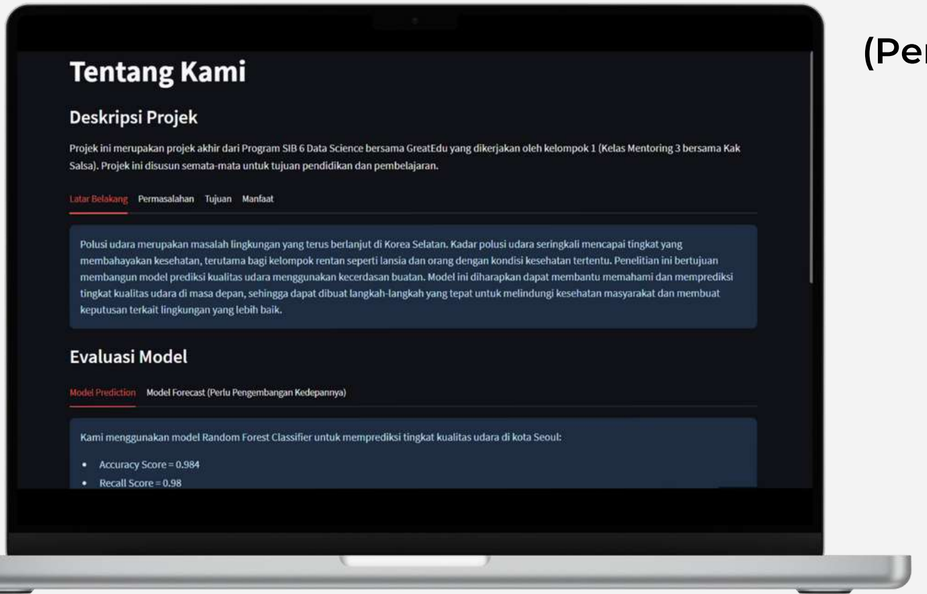


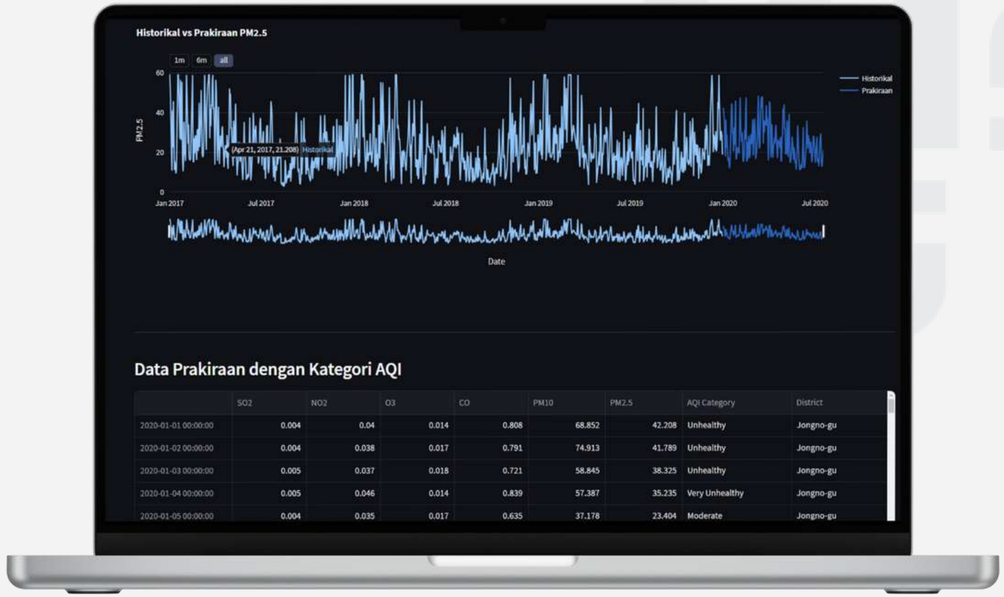
Dashboard Streamlit (2/2)



Menu Prediction

Menu About





Menu Forecast
(Perlu Pengembangan Lanjutan)

7

CONCLUSION AND RECOMMENDATION

Conclusion

1. Dengan menggunakan data polusi udara dan **model machine learning** telah didapatkan hasil prediksi kualitas udara dengan **akurasi 98%**. Dengan demikian kita telah dapat memprediksi kualitas udara dengan data terbaru dengan hasil yang akurat.
2. **Faktor terpenting** dalam penentuan kualitas udara adalah **PM2.5** karena dapat menyebabkan dampak serius pada kesehatan pernapasan dan kardiovaskular.
3. Berdasarkan analisa geografis, daerah yang rentan memiliki **polusi udara tinggi** adalah daerah yang berada di sekitar kawasan **industri**, **lalu-lintas padat** dan **aktivitas manusia lainnya**.

Recommendation

Pengembangan Aplikasi Mobile

Integrasi dengan Sistem Kesehatan

Integrasi dengan Sistem Navigasi

Promosi Transportasi Ramah Lingkungan

Pengembangan Infrastruktur Hijau

REFERENSI

Google Colab : https://colab.research.google.com/drive/1mQHPx1Z-horXgvc6vfRgpT9Vpxm8lAu_?usp=sharing

Streamlit : https://greatedu-final-project---air-quality-seoul-analysis-dan-predic.streamlit.app/

Github : https://github.com/filbertleo88/GreatEdu-Final-Project---Air-Quality-Seoul-Analysis-dan-Prediction

Streamlit

Air pollution is terrible for our children. Every single scientist, every single doctor will tell you the same thing: Air pollution damages our children's brains, their hearts, and their lungs.

Julianne Moore

TERIMA KASIH

ADA PERTANYAAN?