Action - UNESA

Prediksi Kualitas Udara dengan Algoritma Gradient Boosting Classifier untuk *Imbalance Data*

by **Double Layer**

Politeknik Statistika STIS

Muh **Farhan**

Dutatama Rosewika Taufiq Hadihardaya

Politeknik Statistika STIS

Framework dan Metode

Framework

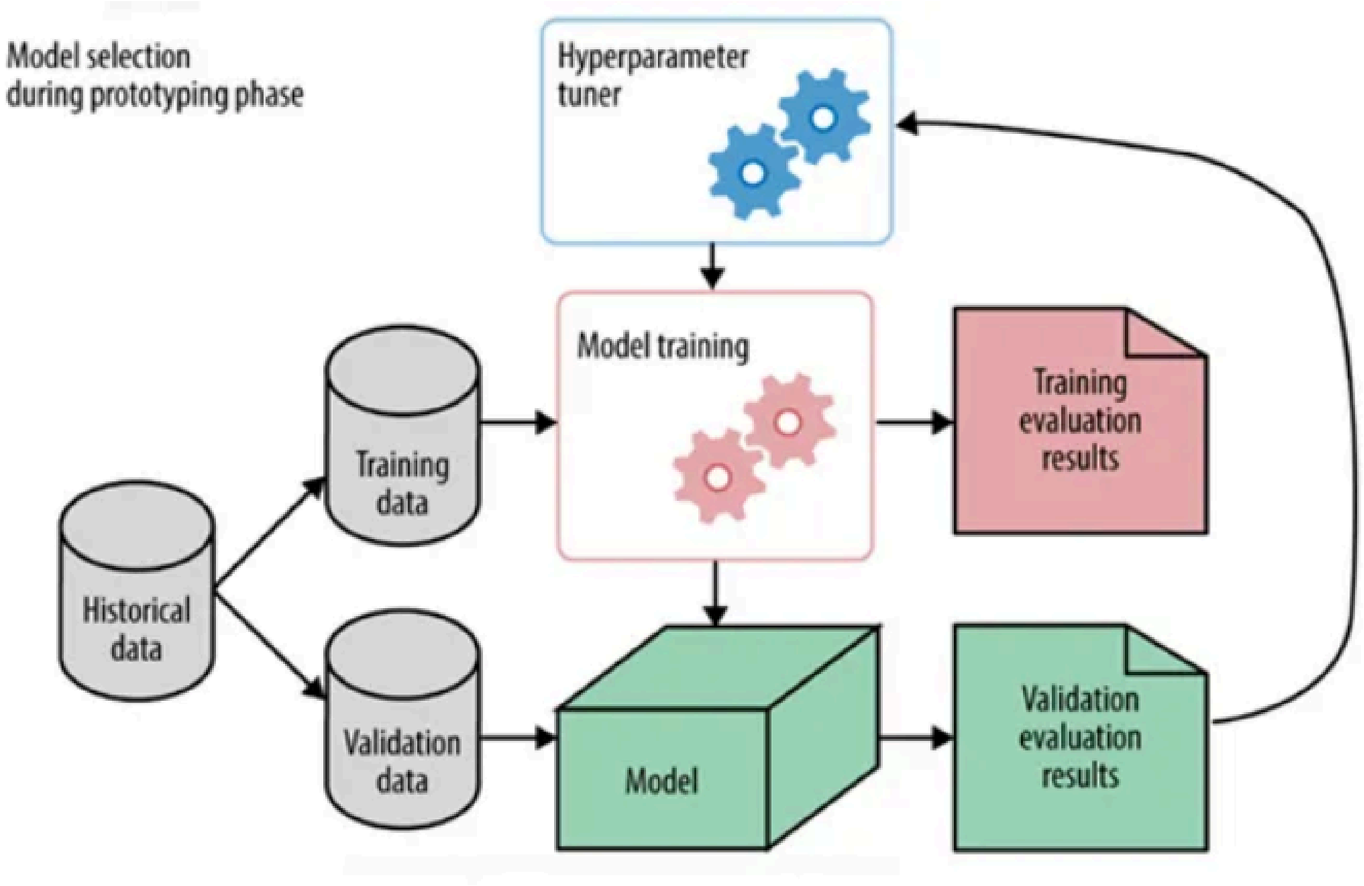
optuna/**optuna**

A hyperparameter optimization framework

244 Contributors | 18k Used by | 334 Discussions | 11k Stars | 1k Forks

Digunanakn dalam proses ***hyperparameter optimization*** secara otomatis dengan metode ***sequential optimization*** yang berfokus pada pencarian *hyperparameter optimal menggunakan pendekatan berbasis* ***Bayesian Optimization*** *atau* ***Tree-structured Parzen Estimator (TPE)***

Framework


Model selection
during prototyping phase
Hyperparameter
tuner
Model training
Training
data
Historical
data
Validation
data
Model
Training
evaluation
results
Validation
evaluation
results

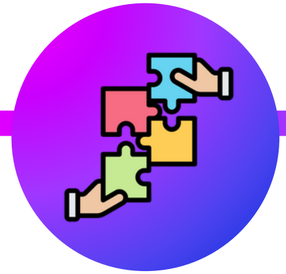
Metode

Preprocessing

Dilakukan dengan mengecek karakteristik data, penyesuaian isian data, penghapusan fitur-fitur yang kurang relevan, dan imputasi *missing value*

Analisis Data Eksploratif

Dilakukan dengan memvisualisasikan karakteristik data dalam rangka mengeksplorasi karakteristik data secara mendalam



Model Building

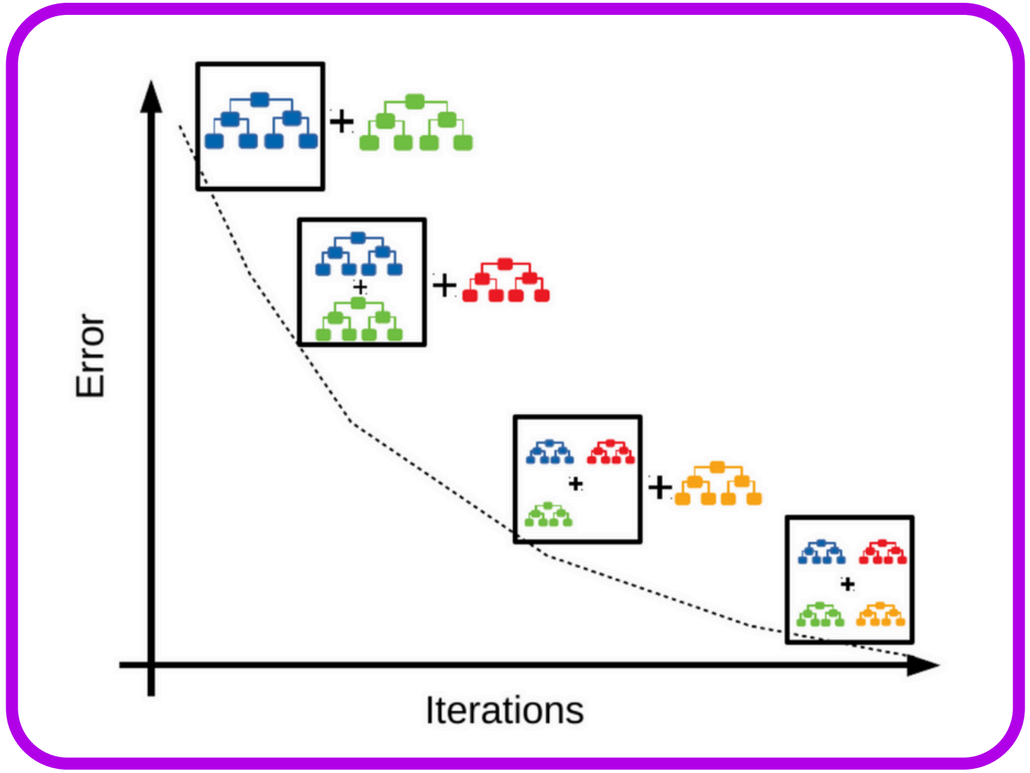
Dilakukan dengan menggunakan framework Optuna dan model Gradient Bossting Classifier

Model Evaluation

Dilakukan dengan membuat metrik-metrik evaluasi yang sesuai untuk tugas klasifikasi seperti *confusion matrix* dan tabel performa model

Metode

Algoritma Gradient Boosting Classifier



- **Unggul** dalam tugas klasifikasi
- **Efektif** untuk *imbalance data*
- **Fleksibel** dalam pemilihan *hypermarameter*
- Mampu m**engurangi risiko *overfitting*** pada data *training*

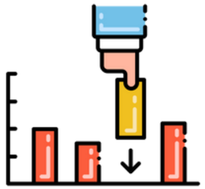
Metode

Preprocessing

Pemilihan fitur
yang relevan

Penyesuaian format
dan isian data

Encoding variabel
kategorikal menggunakan
target encoder



Imputasi *missing data*
menggunakan algoritma
missForest

Metode

Pembagian Dataset

Validation Frame

10% dari *class "Safety"* (pada keseluruhan data training)

50% dari *class "Dangerous"* (pada keseluruhan data training)

Training Frame

70% data *training*

30% data *testing*

dengan **stratifikasi** berdasarkan class pada pada variabel target

Metode

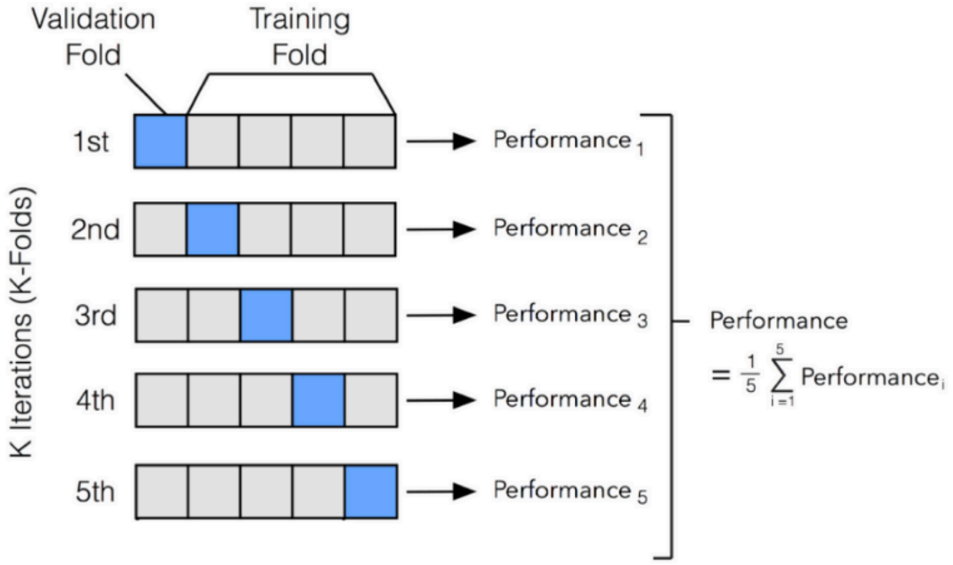
Evaluasi Model

Metrik Evaluasi yang Digunakan

$$Precision = \frac{TP}{FP + TP}$$

$$Recall = \frac{TP}{FN + TP}$$

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall}$$



Metode

Hyperparameter Tuning

Kombinasi Hyperparameter yang Digunakan

- n estimators
- learning rate
- max depth
- subsamples
- max features
- min samples split
- min samples leaf

	precision	recall	f1
Model 1	1.000000	0.935484	0.966667
Model 2	0.937500	0.967742	0.952381
Model 3	1.000000	0.935484	0.966667
Model 4	1.000000	1.000000	1.000000
Model 5	0.688889	1.000000	0.815789
...	...	...	...
Model 146	1.000000	1.000000	1.000000
Model 147	0.688889	1.000000	0.815789
Model 148	0.935484	0.935484	0.935484
Model 149	0.939394	1.000000	0.968750
Model 150	0.885714	1.000000	0.939394

Analisis Data Eksploratif

Informasi Data Training

Target: Air_quality_category

Feature	Tipe Data	Missing Value	Feature	Tipe Data	Missing Value
DataId	Kategorik	-	**pm10_tempcov**	Numerik	45%
WHO_Region	Kategorik	-	**pm25_tempcov**	Numerik	63%
iso3	Kategorik	-	**no2_tempcov**	Numerik	43%
country_name	Kategorik	-	**type_of_stations**	Kategorik	-
city	Kategorik	-	**number_of_stations**	Numerik	-
year	Numerik	-	**population**	Numerik	40%
pm10_concentration	Numerik	27%	**latitude**	Numerik	-
pm25_concentration	Numerik	50%	**longitude**	Numerik	-
no2_concentration	Numerik	34%	**who_ms**	Kategorik	-

25 999 Observasi

Unit Observasi: kota tiap tahun

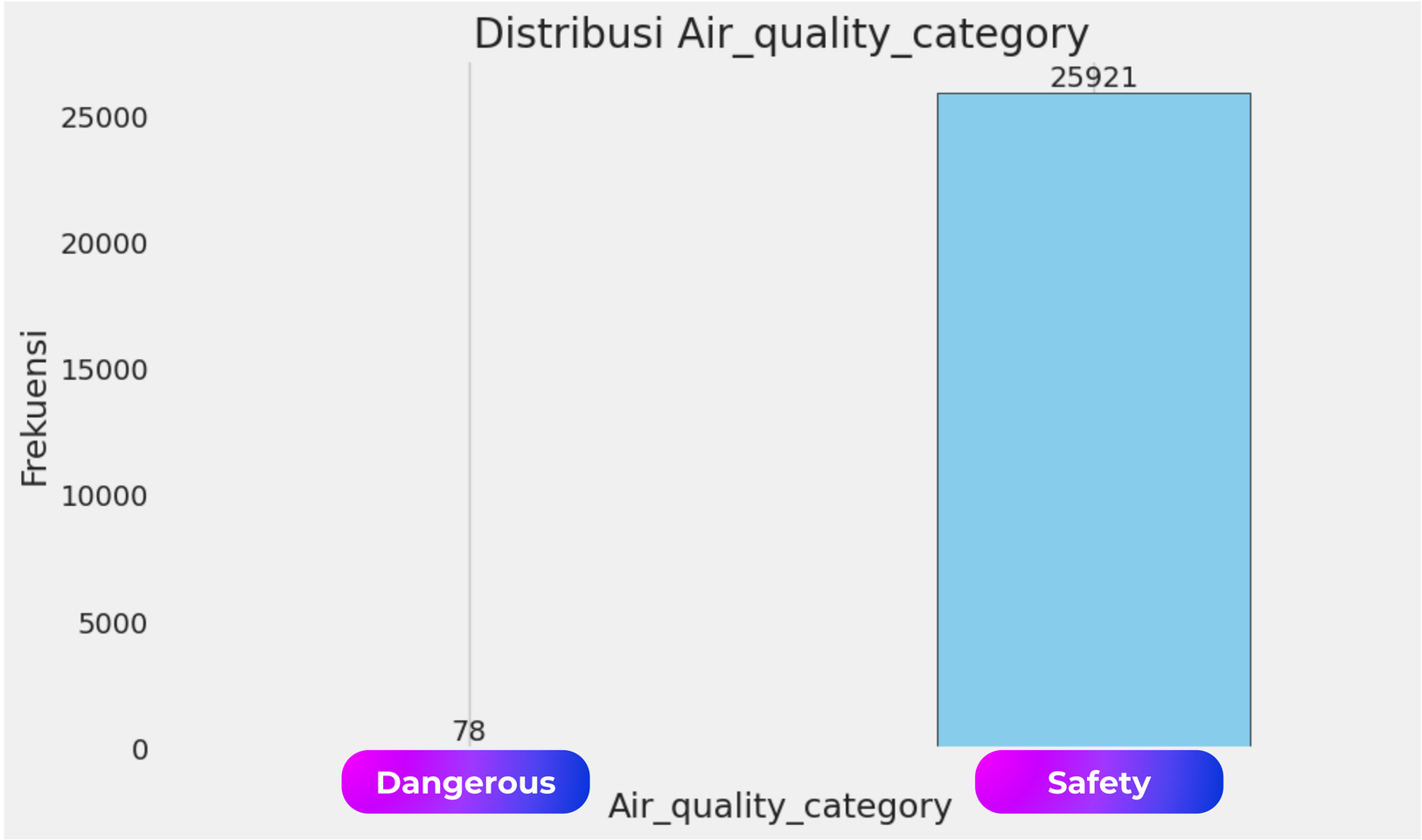
Informasi Data Testing

Feature	Tipe Data	Missing Value	Feature	Tipe Data	Missing Value
DataId	Kategorik	-	**pm10_tempcov**	Numerik	41%
WHO_Region	Kategorik	-	**pm25_tempcov**	Numerik	51%
iso3	Kategorik	-	**no2_tempcov**	Numerik	38%
country_name	Kategorik	-	**type_of_stations**	Kategorik	-
city	Kategorik	-	**number_of_stations**	Numerik	-
year	Numerik	-	**population**	Numerik	52%
pm10_concentration	Numerik	32%	**latitude**	Numerik	-
pm25_concentration	Numerik	38%	**longitude**	Numerik	-
no2_concentration	Numerik	31%	**who_ms**	Kategorik	-

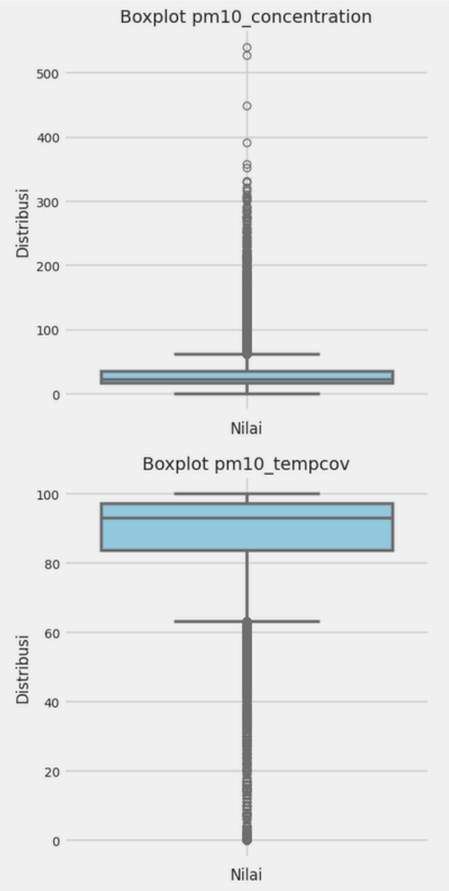
14 005 Observasi

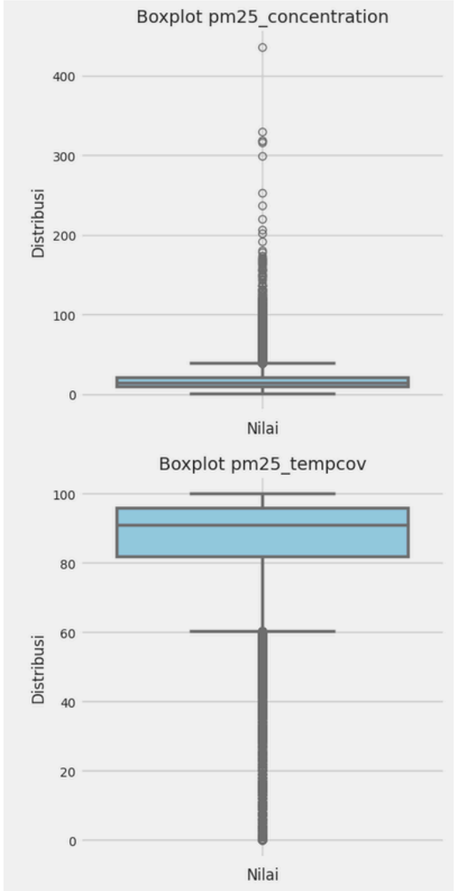
Unit Observasi: kota tiap tahun

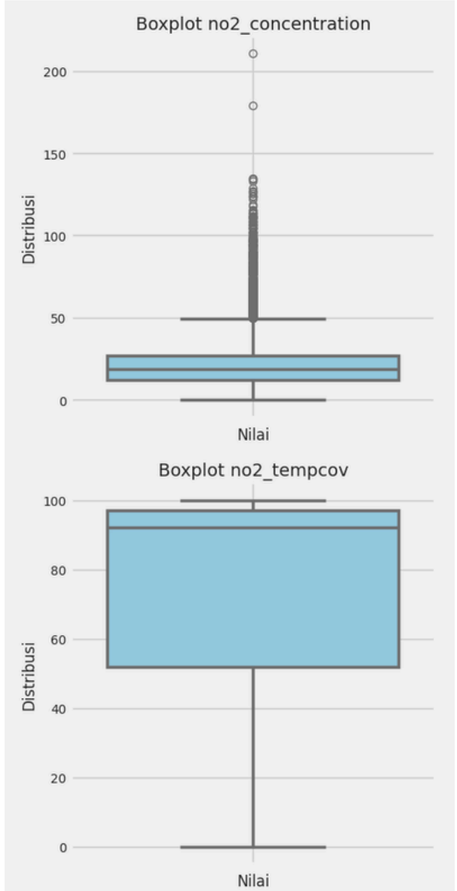
Distribusi Variabel Target

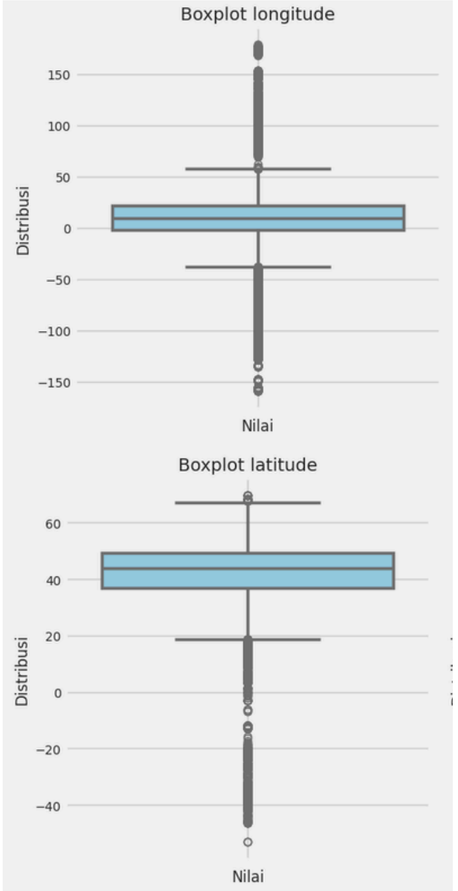


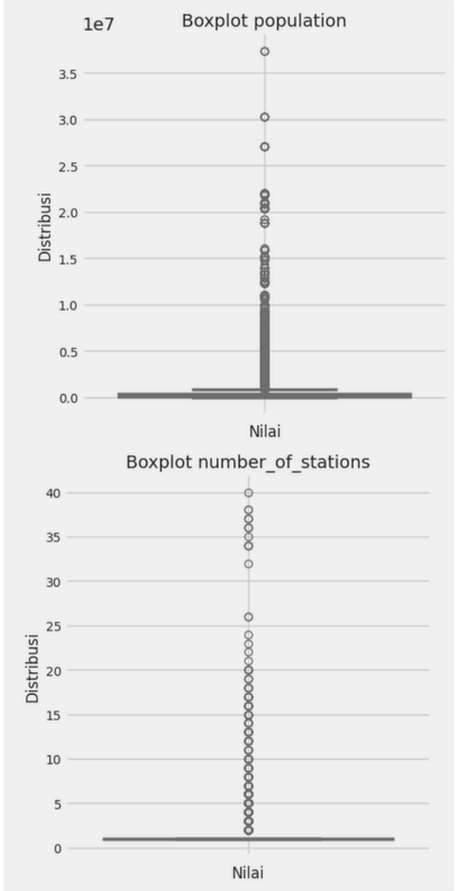
Distribusi Data Fitur







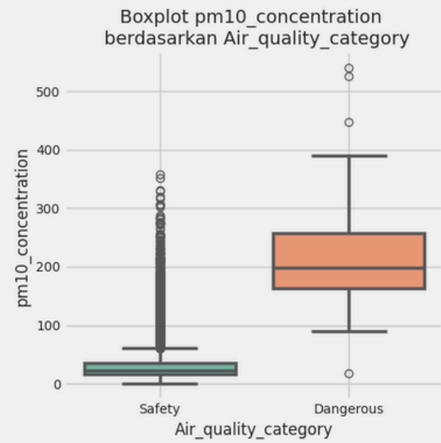


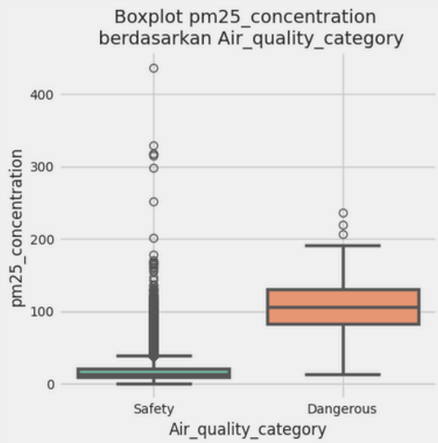


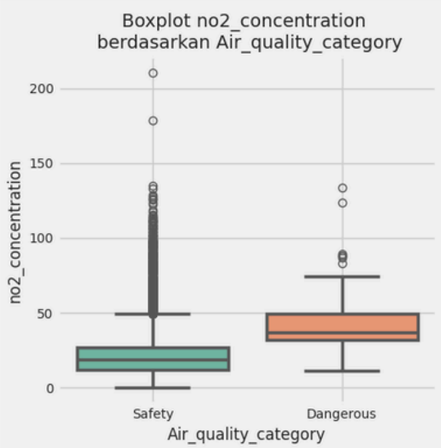
Deteksi Outliers

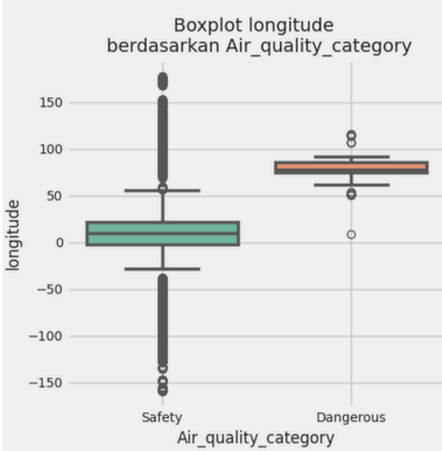
Feature	Tipe Data	Outlier	Feature	Tipe Data	Outlier
DataId	Kategorik	-	**pm10_tempcov**	Numerik	8%
WHO_Region	Kategorik	-	**pm25_tempcov**	Numerik	9%
iso3	Kategorik	-	**no2_tempcov**	Numerik	-
country_name	Kategorik	-	**type_of_stations**	Kategorik	-
city	Kategorik	-	**number_of_stations**	Numerik	12%
year	Numerik	-	**population**	Numerik	13%
pm10_concentration	Numerik	12%	**latitude**	Numerik	7%
pm25_concentration	Numerik	10%	**longitude**	Numerik	33%
no2_concentration	Numerik	3%	**who_ms**	Kategorik	-

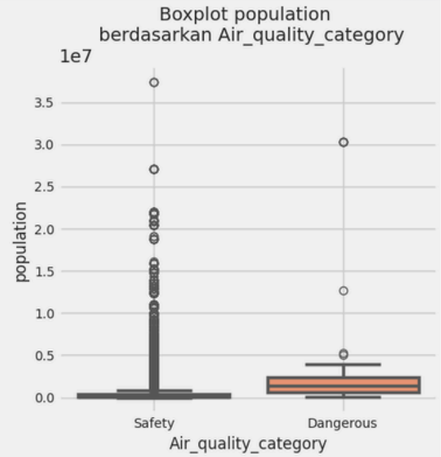
Distribusi Data Berdasarkan Kategori Kualitas Udara

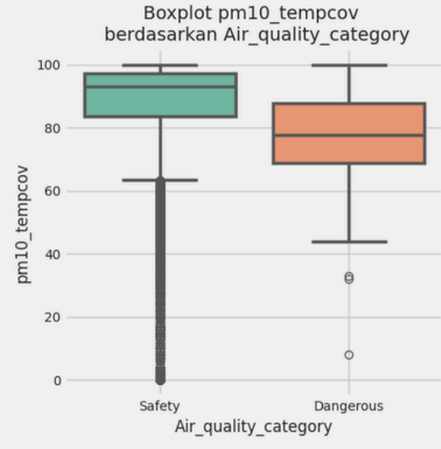


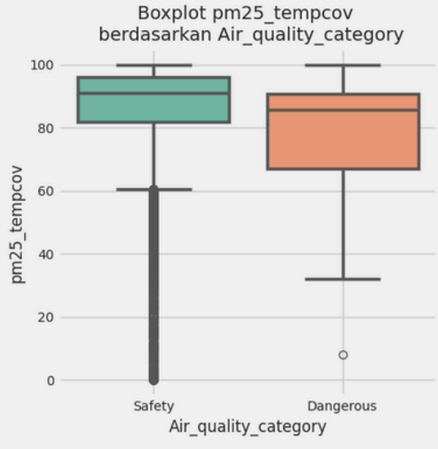


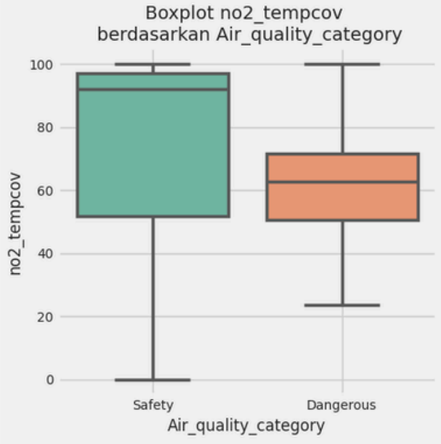


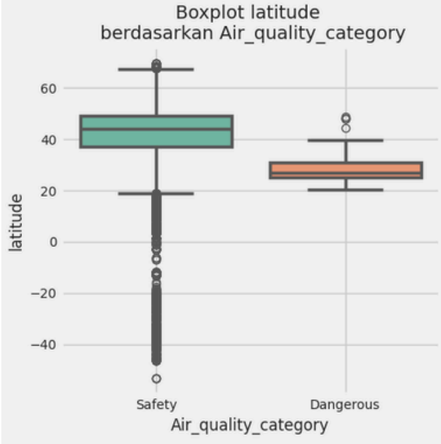


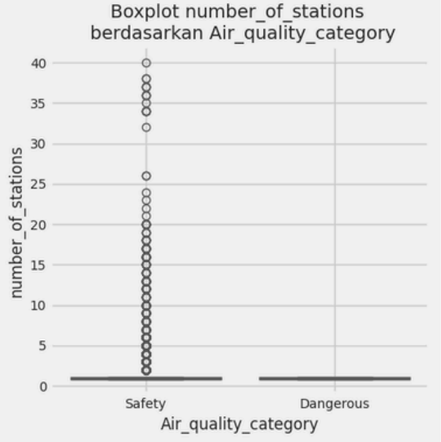




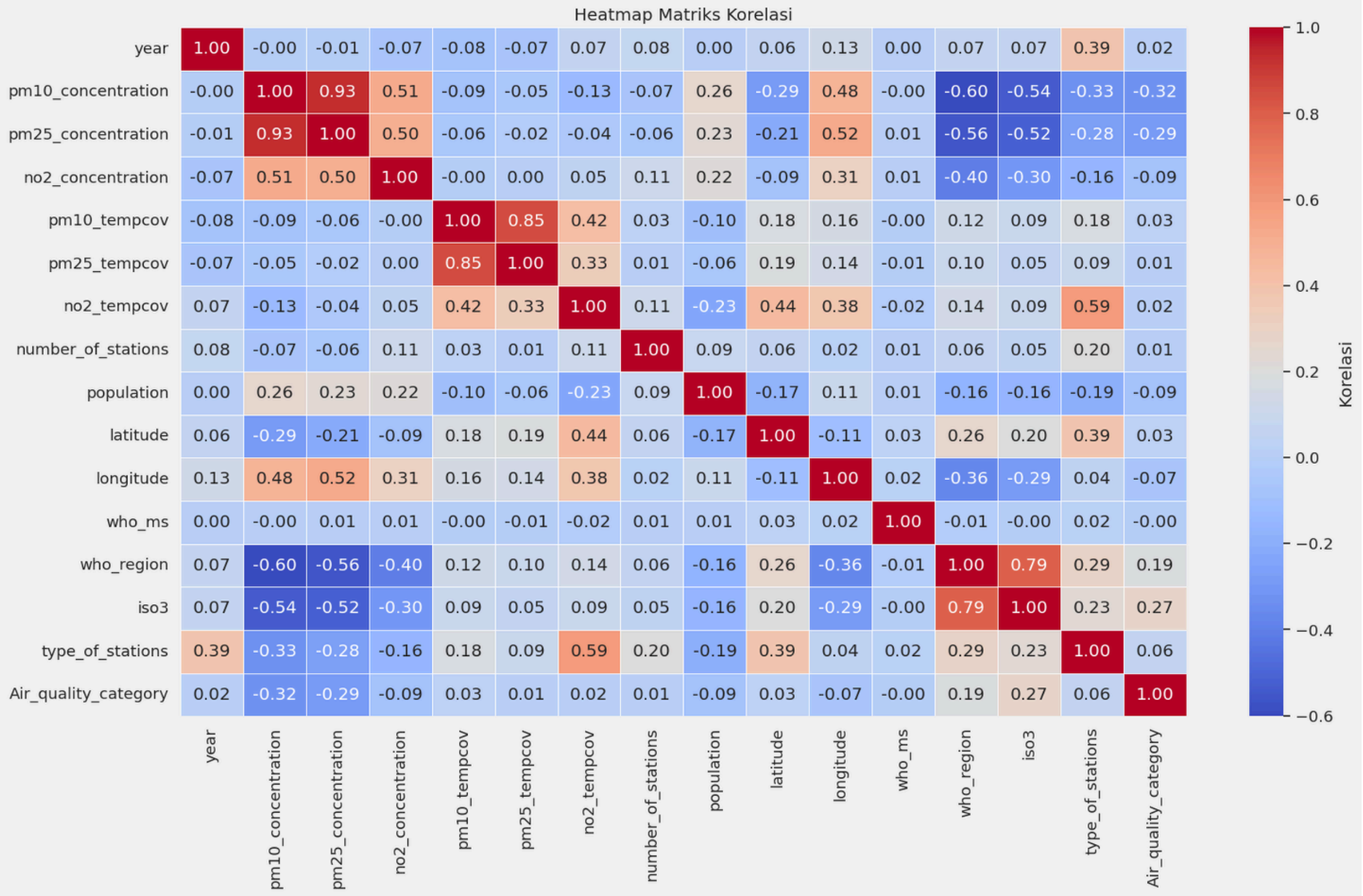




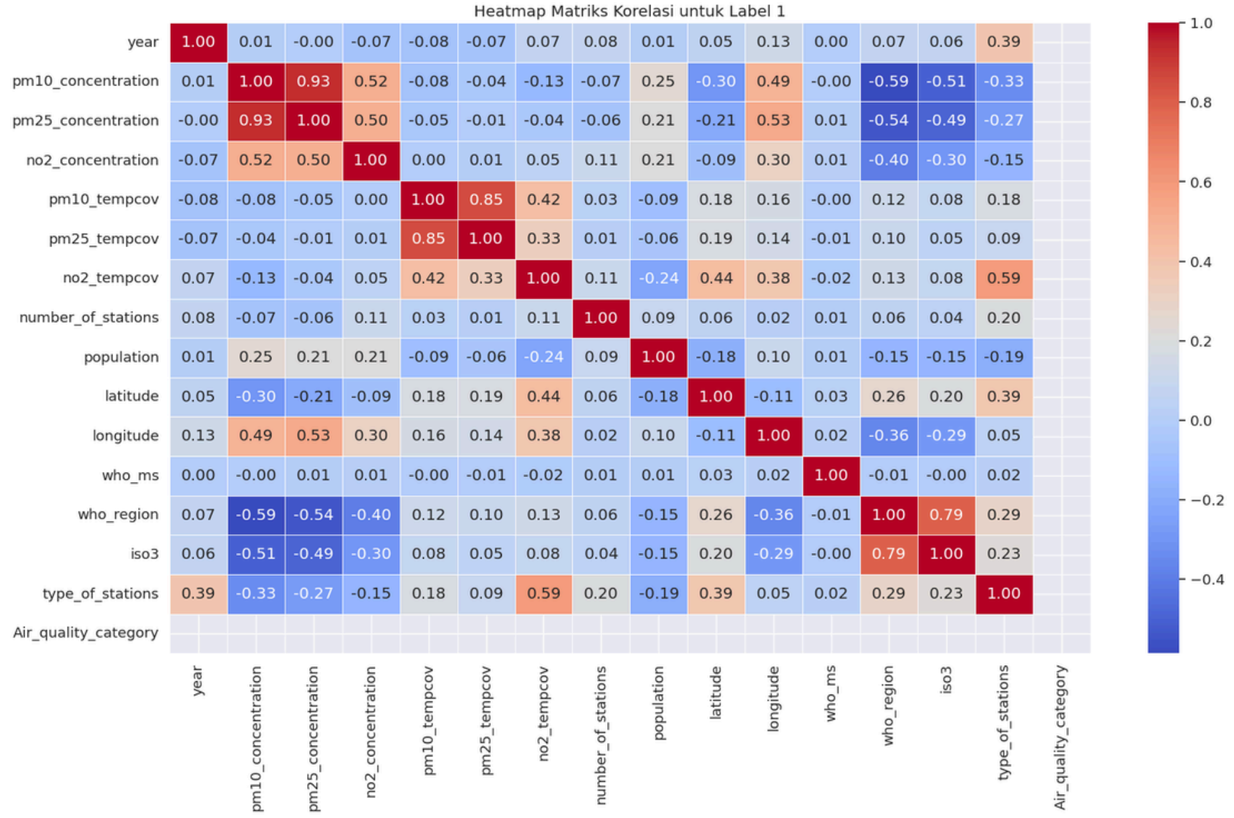




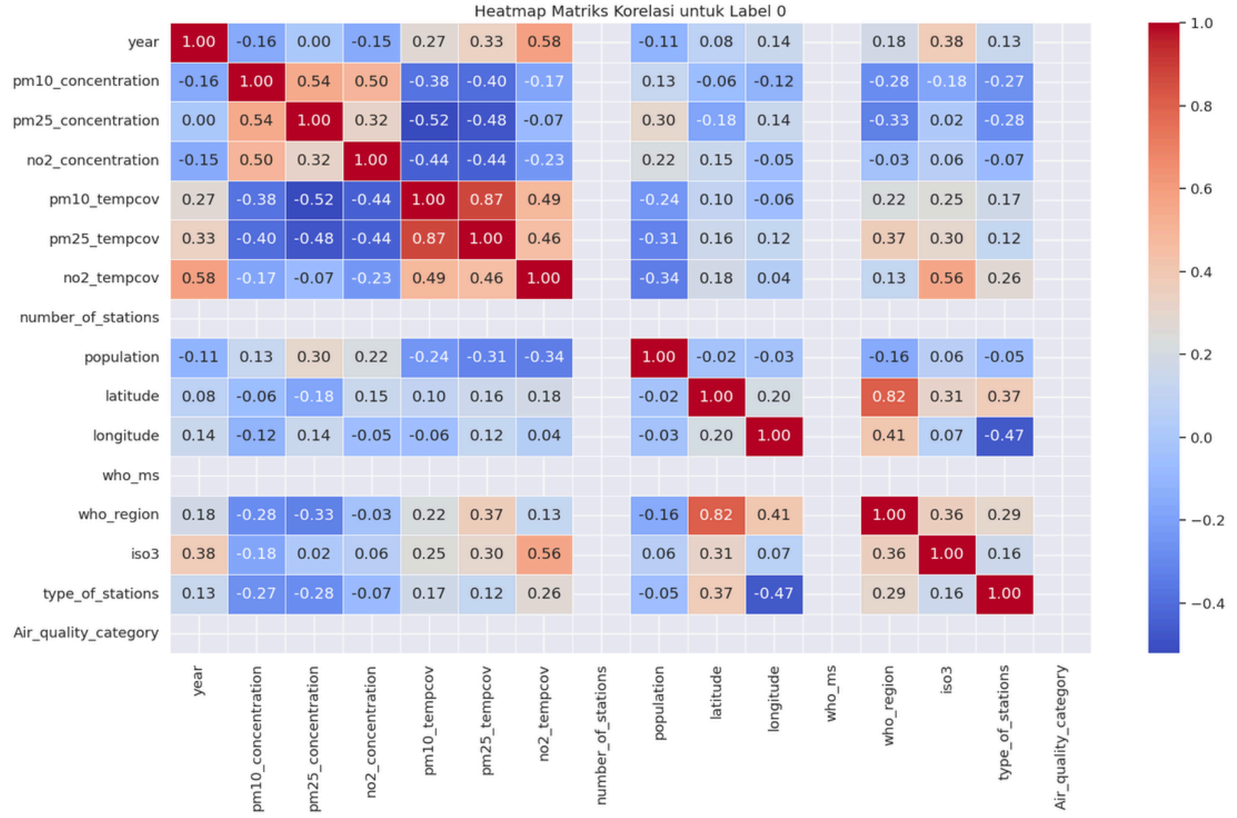
Heatmap Korelasi



Heatmap Korelasi Berdasarkan Kategori Kualitas Udara



Safety



Dangerous

Hasil

Confusion Matrix

Predicted

Actual		**Positive**	**Negative**
	Positive	2524	68
	Negative	0	31

Berdasarkan data validation

Metrik Evaluasi Model

Precision	Recall	F1-Score
1.0	0.97	0.98

Feature Importance

Feature	Importance
pm10_concentration	**33.65**
pm25_concentration	**31.98**
iso3	10.99
longitude	5.29
who_region	4.43

Kesimpulan dan Rekomendasi

Kesimpulan

Performa

Model Gradient Boosting Classifier yang dibangun mampu mengklasifikasikan kualitas udara sebagai aman ("Safety") atau berbahaya ("Dangerous") dengan tingkat akurasi atau ketepatan yang baik (**F1-Score bernilai 0.98**)

Variabel

Model yang dibangun berhasil megnidentifikasi vairabel yang berkontribusi tinggi pada tingkat kualitas udara adalah **konsentrasi PM10 dan konsentrasi PM25**

False Negative

Model yang dibangun menunjukkan tingkat **false negative yang relatif rendah**, yaitu sebesar **2.59%**

Rekomendasi

Integrasi IoT

Model yang dibangun dapat diintegrasikan dengan perangkat-perangkat IoT atau pltaform digital untuk memberikan informasi langsung kepada masyarakat tentang kondisi udara di daerah mereka

Pengontrolan Aktivitas Industri

Model ini dapat menjadi dasar dalam membuat kebijakan berbasis data seperti pengurangan aktivitas industri pada hari-hari dengan kualitas udara buruk atau mengontrol pembakaran terbuka.

Pembantuan Pencegahan Gangguan Kesehatan

Informasi akhir yang dapat dihasilkan oleh model ini mengenai kondisi kualitas udara dapat menjadi pegangan awal pemerintah untuk menetapkan kebijakan sebagai upaya pencegahan gangguan kesehatan akibat kualitas udara yang buruk

Pemantauan *realtime*

Pemantauan kualitas udara secara *real-time* juga menjadi sangat memungkinkan untuk dilakukan menggunakan model yang dibangun dengan integrasi pada beberapa alat sensor udara

Terima Kasih!

by **Double Layer**

Politeknik Statistika STIS