JURNAL ILMIAH SAINS & TEKNOLOGI
UNIVERSITAS BANTEN JAYA

Jurnal Ilmiah Sains dan Teknologi Vol. 9, No. 1, Februari 2025
P–ISSN: 1907-1205
E-ISSN: 2622-6391

PENGGUNAAN *MACHINE LEARNING* UNTUK MEMPREDIKSI PAPARAN PM2.5 DAN DAMPAKNYA TERHADAP KESEHATAN

[1]Lita Widiastuti, [2]Wiga Maulana Baihaqi
Universitas Amikom Purwokerto
Email: [1]mitalita97@gmail.com, [2]wiga@amikompurwokerto.ac.id

Abstract

The decline in air quality in Indonesia, caused by the use of fossil fuels, causes air pollution with high concentrations of PM2.5 and poses a risk to public health. This research aims to apply machine learning methods to analyze factors that influence air quality on health, as well as predict exposure to PM2.5 and welfare costs due to premature mortality caused by air pollution. The research method used is CRISP-DM, which includes data collection, data understanding, data preparation, modeling, evaluation and application. The data used comes from OECD public sources for the period 1994–2022. Based on forecasting results in research, it shows that exposure to PM2.5 will continue to increase in the next 8 years, which will contribute to increased welfare costs due to health impacts that can cause respiratory problems and premature death. The ARIMA model was used to predict PM2.5 exposure, while ARIMAX+ETS was used to predict welfare costs, with increased accuracy after hyper parameter tuning. To overcome this, policies such as the adoption of environmentally friendly technology, clean energy transition, waste management and green city development are recommended so that it is hoped that they can reduce the impact of air pollution and reduce welfare costs in the future.

Keywords: *Air quality, Machine Learning, prediction, PM2.5 exposure, CRISP-DM*

PENDAHULUAN

Udara adalah elemen penting untuk kelangsungan hidup suatu makhluk, terutama manusia. Namun kenyataannya, kualitas udara bersih di Indonesia saat ini semakin menurun. Salah satu penyebab utama pencemaran udara di negara Indonesia yaitu pemakaian bahan bakar fosil, yang umumnya digunakan untuk kendaraan bermotor dan kebutuhan industri. Bahan bakar fosil mengandung zat dan partikel yang mencemari udara, menciptakan polusi yang berdampak negatif terhadap lingkungan dan kehidupan makhluk hidup (Salsabila, Wardah Nibras & Sudarti, 2023). Menurut laporan kualitas udara dunia IQAir 2023 yang dirilis pada Maret 2024, Indonesia berada di posisi ke-14 sebagai salah satu negara dengan tingkat polusi udara tertinggi di dunia. Konsentrasi PM2.5 (*Particulate matter*) di Indonesia tercatat sebesar 37,1 µg per meter kubik (Sumiyati, 2024). Kondisi ini menunjukkan bahwa polusi udara di Indonesia telah mencapai tingkat yang mengkhawatirkan.

Polusi udara menjadi ancaman besar bagi kesehatan manusia dan makhluk hidup, tidak hanya di Indonesia tetapi juga di seluruh dunia. Berdasarkan data WHO, polusi udara menyebabkan sekitar 7 juta kematian setiap tahun secara global. Selain berdampak buruk pada kesehatan manusia, polusi udara juga berkontribusi pada terbentuknya kabut asap dan hujan asam, menghancurkan ekosistem hutan, serta mencemari lingkungan secara keseluruhan. Meski kualitas udara di Indonesia sedemikian buruknya, tetapi masih banyak orang yang mengabaikan bahaya polusi udara bagi kesehatan (Kemenkes, 2024). Polusi tersebut juga tidak hanya berdampak negatif pada kesehatan masyarakat, tetapi juga mempengaruhi perekonomian dan tingkat produktivitas suatu negara. Oleh karena itu, memahami sejauh mana penggunaan transportasi

berkelanjutan dapat berkontribusi dalam meningkatkan kualitas udara dan kesejahteraan masyarakat menjadi hal yang sangat penting (Rahmawati & Pratama, 2023).

Dampak dari polusi udara tidak hanya terlihat melalui angka atau data statistik, tetapi secara langsung mempengaruhi kesehatan masyarakat. Selain itu, penurunan kualitas udara juga membawa konsekuensi pada aktivitas sosial dan ekonomi, udara yang tercemar dapat menurunkan tingkat produktivitas akibat masalah kesehatan, sekaligus meningkatkan beban biaya pelayanan kesehatan (Rahmawati & Pratama, 2023). Salah satu upaya untuk memahami dampak polusi udara secara menyeluruh terkait hubungan antara kualitas udara dan kesehatan masyarakat semakin menjadi perhatian utama. Analisis terhadap faktor-faktor yang mempengaruhi kualitas udara dan dampak paparan jangka panjang terhadap kesehatan, perlu ditinjau untuk memberikan landasan kebijakan yang tepat. Dalam konteks ini, penerapan machine learning menjadi relevan untuk menganalisis data dan menemukan pola tertentu, sehingga dapat mendukung pemahaman yang lebih mendalam tentang hubungan antara kualitas udara dan dampaknya terhadap kesehatan masyarakat.

Saat ini, terdapat berbagai teknik *machine learnin*g yang tersedia, masing-masing dengan karakteristik uniknya. Dalam menerapkan *machine learning* untuk menyelesaikan suatu masalah, penting untuk memilih dan menggunakan teknik yang sesuai, sehingga model prediktif yang dihasilkan dapat memiliki kinerja optimal dan relevan dengan permasalahan yang dihadapi (Anwar & Permana, 2021). Keunggulan utamanya adalah kemampuannya dalam mengolah data yang kompleks serta mengenali pola-pola yang mungkin sulit terdeteksi menggunakan metode tradisional (Siti Nurjanah et al., 2024). Sudah banyak penelitian yang telah menerapkan penggunaan *machine learning* dalam memprediksi suatu hal, seperti penelitian yang dilakukan oleh (Siti Nurjanah et al., 2024) yang berjudul Prediksi Kecepatan Angin untuk Mengetahui Potensi Sumber Energi Alternatif menggunakan Model Regresi Lasso: Studi Kasus Kota Makassar pada Tahun 2024 yang bertujuan untuk meramalkan kecepatan angin di Kota Makassar dengan menerapkan *machine learning* terutama menggunakan model regresi Lasso dan hasilnya menunjukkan model yang digunakan akurat dalam memprediksi kecepatan angin di Kota Makassar. Hal ini ditunjukkan oleh nilai MSE yang rendah (0.334) dan nilai R2 yang tinggi (0.97). Sehingga prediksi kecepatan angin tertinggi di tahun 2024 mencapai 10.76 meter per detik. Kecepatan angin ini berpotensi menghasilkan listrik hingga 1597 watt. Hasil penelitian ini memberikan petunjuk yang kuat bahwa Kota Makassar memiliki potensi sangat baik untuk memanfaatkan energi angin sebagai sumber listrik alternatif.

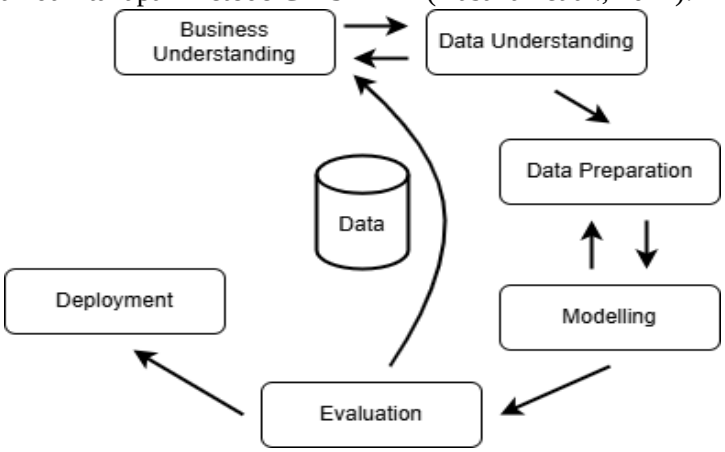
Dengan adanya penelitian sebelumnya dalam mengatasi permasalahan yang terjadi maka penelitian ini juga bertujuan untuk menerapkan *machine learning* dalam menganalisis faktor-faktor yang mempengaruhi kualitas udara terhadap kesehatan, serta memberikan prediksi terkait paparan PM2.5 dan biaya kesejahteraan akibat mortalitas dini yang disebabkan oleh polusi udara khususnya di Indonesia. Dengan pendekatan ini, penelitian diharapkan mampu mengidentifikasi faktor-faktor yang berdampak pada kesehatan masyarakat, memberikan prediksi paparan pm2,5 dimasa yang akan datang, serta menawarkan rekomendasi untuk mendukung kebijakan lingkungan yang lebih efektif dan berkelanjutan.

METODE PENELITIAN

Pada penelitian ini menggunakan pendekatan metode deskriptif kuantitatif. Penelitian kuantitatif bertujuan untuk menggali suatu kondisi atau situasi dengan cara terstruktur, memecahkan masalah berdasarkan data dan fakta yang tersedia. Data kuantitatif, yang berbentuk angka serta berasal dari fakta yang ada (Haryanto et al., 2023). Berikut adalah proses yang akan dilakukan:

1. Pengumpulan Data
   Pada penelitian ini data yang digunakan berasal dari data publik yang didapatkan dari https://data-explorer.oecd.org/.
2. Metode CRISP-DM

Metode CRISP-DM, yaitu pemodelan proses pengembangan data yang digunakan untuk mendapatkan hasil penelitian (Syahfutris et al., 2023). Metode CRISP-DM ini adalah kerangka kerja umum yang digunakan untuk memecahkan masalah dalam bisnis dan penelitian yang mencakup enam tahapan utama, yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* serta *Deployment* dan berikut adalah penjelasan dari tahapan metode CRISP-DM (Hasanah et al., 2021):



Gambar 1. Metode CRISP-DM

1) *Business Understanding* (Pemahaman Bisnis)
   Tahap ini melibatkan pemahaman kebutuhan dan tujuan dari perspektif bisnis, yang kemudian diterjemahkan menjadi definisi masalah dalam konteks data mining. Selanjutnya, rencana dan strategi disusun untuk mencapai tujuan tersebut.
2) *Data Understanding* (Pemahaman Data)
   Pada tahap ini, proses dimulai dengan pengumpulan data, diikuti oleh deskripsi dan evaluasi kualitas data yang tersedia.
3) *Data Preparation* (Persiapan Data)
   Tahap ini bertujuan untuk membangun dataset akhir dari data mentah. Prosesnya mencakup pengecekan data yang digunakan agar siap digunakan pada tahap pemodelan.
4) *Modeling* (Pemodelan)
   Tahap ini melibatkan penggunaan machine learning untuk menentukan teknik, alat bantu, dan algoritma.
5) *Evaluation* (Pengujian)
   Pada tahap ini, performa pola yang dihasilkan oleh algoritma dievaluasi untuk menilai tingkat keakuratannya.
6) *Deployment* (Penerapan)
   Tahapan terakhir mencakup penyusunan laporan dan pembuatan artikel jurnal berdasarkan model yang dihasilkan.

HASIL DAN PEMBAHASAN

Pada bagian pembahasan ini, akan diulas hasil-hasil yang diperoleh dari penerapan metode penelitian yang telah dijelaskan sebelumnya. Pembahasan ini bertujuan untuk menginterpretasikan temuan-temuan utama, menghubungkannya dengan tujuan penelitian, serta memberikan wawasan terkait implikasi dan relevansi hasil dalam konteks masalah yang diteliti. Berikut adalah penjelasannya:

1. *Business Understanding*
   Penerapan machine learning dalam penelitian ini bertujuan untuk menganalisis faktor-faktor yang mempengaruhi kualitas udara terhadap kesehatan. Selain itu, penelitian ini juga bertujuan untuk memprediksi terkait paparan PM2.5 dan biaya kesejahteraan akibat mortalitas dini yang disebabkan oleh polusi udara khususnya di Indonesia.
2. *Data Understanding*
   Penelitian ini menggunakan data yang diambil dari website https://data-explorer.oecd.org/. Data tersebut merupakan data mengenai *green growth* yang berisi berisi indikator-indikator terpilih untuk memantau kemajuan menuju pertumbuhan hijau untuk mendukung pembuatan kebijakan dan memberikan informasi kepada masyarakat luas. Dimana pada penelitian ini mengambil tahun dari 1994-2022 dengan negara yang digunakan yaitu Indonesia, Malaysia, Brunei Darussalam, Singapura, dan Thailand, namun pada penelitian ini akan lebih berfokus untuk memprediksi negara Indonesia. Kolom-kolom yang digunakan ada 27, dan berikut adalah penjelasan dari masing-masing kolom:

Tabel 1. Deskripsi Kolom yang Digunakan

No	Nama Kolom	Deskripsi Penjelasan
1	REF_AREA	Menunjukkan area geografis atau negara.
2	TIME_PERIOD	Periode waktu (biasanya tahun) saat data diambil.
3	Demand-based GHG intensity energy-related GHG per capita	Intensitas emisi gas rumah kaca terkait energi per kapita berdasarkan permintaan konsumsi.
4	Production-based CO2 emissions	Total emisi karbon dioksida yang dihasilkan dari produksi dalam suatu wilayah.
5	Production-based GHG emissions	Total emisi gas rumah kaca (GHG) yang dihasilkan dari kegiatan produksi dalam suatu wilayah geografis tertentu.
6	Population exposure to PM2.5	Persentase populasi yang terpapar polusi partikel halus (PM2.5) di udara, yang dapat berdampak negatif pada kesehatan.
7	Mortality from exposure to ambient PM2.5	Jumlah kematian yang diakibatkan oleh paparan PM2.5 di udara luar ruangan.
8	Mortality from exposure to residential radon	Jumlah kematian akibat paparan gas radon di dalam rumah.
9	Mortality from exposure to lead	Jumlah kematian yang diakibatkan oleh paparan timbal.
10	Nitrogen balance	Keseimbangan nitrogen, mengindikasikan dampak aktivitas manusia terhadap lingkungan.
11	Female population	Jumlah populasi perempuan.
12	Population, ages 0-14	Jumlah populasi dalam kelompok usia 0-14 tahun.
13	Population, ages 15-64	Jumlah populasi dalam kelompok usia 15-64 tahun.
14	Net migration	Perbedaan antara jumlah orang yang masuk ke suatu wilayah dan yang keluar dalam periode tertentu.
15	Population density	Kepadatan penduduk per unit area.
16	Welfare costs of premature mortalities from exposure to lead	Biaya kesejahteraan akibat kematian dini yang diakibatkan oleh paparan timbal.
17	Welfare costs of premature mortalities from exposure to ambient PM2.5	Biaya kesejahteraan akibat kematian dini dari paparan PM2.5.

18	Welfare costs of premature mortalities from exposure to residential radon	Biaya kesejahteraan akibat kematian dini akibat paparan radon.
19	Life expectancy at birth	Harapan hidup rata-rata saat lahir.
20	Total fertility rate	Jumlah rata-rata anak yang dilahirkan oleh seorang perempuan selama masa suburnya.
21	Energy intensity per capita	Jumlah energi yang digunakan per kapita.
22	Renewable energy supply	Pasokan energi terbarukan dalam suatu wilayah.
23	Energy productivity, GDP per unit of TES	Produktivitas energi, yaitu jumlah PDB yang dihasilkan per unit total energi yang digunakan.
24	Total energy supply	Jumlah total energi yang tersedia di suatu wilayah.
25	Real GDP per capita	Produk domestik bruto nyata per kapita, disesuaikan dengan inflasi.
26	Purchasing power parity	Paritas daya beli, menunjukkan perbandingan daya beli antar negara berdasarkan harga barang/jasa.
27	Real GDP	Produk domestik bruto nyata suatu wilayah, disesuaikan dengan inflasi.

3. *Data Preparation*

Pada tahap ini dilakukan pengecekan *missing values*, *duplicate data*, *outlier data*, dan *imbalance data*. Berikut adalah penjelasan dari masing-masing pengecekan:

1) *Missing values*

Missing values adalah ketidakadaan data pada suatu entri atau observasi. Dalam konteks data science, missing value sangat krusial dalam tahap pembersihan data (data wragling) sebelum analisis dan prediksi dilakukan. Data wargling adalah proses untuk mempersiapkan data, dengan menghapus informasi yang tidak relevan, sehingga data menjadi siap untuk dianalisis (Abrar et al., 2023). Pada data yang digunakan masih terdapat kolom yang memiliki nilai kosong, dimana lebih jelasnya bisa dilihat pada gambar 2 di bawah ini:

	0
REF_AREA	0
TIME_PERIOD	0
Demand-based GHG intensity energy-related GHG per capita	20
Female population	0
Welfare costs of premature mortalities from exposure to lead	20
Production-based CO2 emissions	10
Welfare costs of premature mortalities from exposure to ambient PM2.5	20
Nitrogen balance	123
Purchasing power parity	0
Life expectancy at birth	0
Population exposure to PM2.5	40
Mortality from exposure to ambient PM2.5	20
Total fertility rate	0
Production-based GHG intensity, energy-related GHG per capita	20
Energy intensity per capita	5
Population density	0
Production-based GHG emissions	20
Population, ages 0-14	0
Net migration	0
Renewable energy supply	22
Mortality from exposure to residential radon	20
Real GDP per capita	0
Welfare costs of premature mortalities from exposure to residential radon	20
Population, ages 15-64	0
Mortality from exposure to lead	20
Energy productivity, GDP per unit of TES	5
Total energy supply	5
Real GDP	0

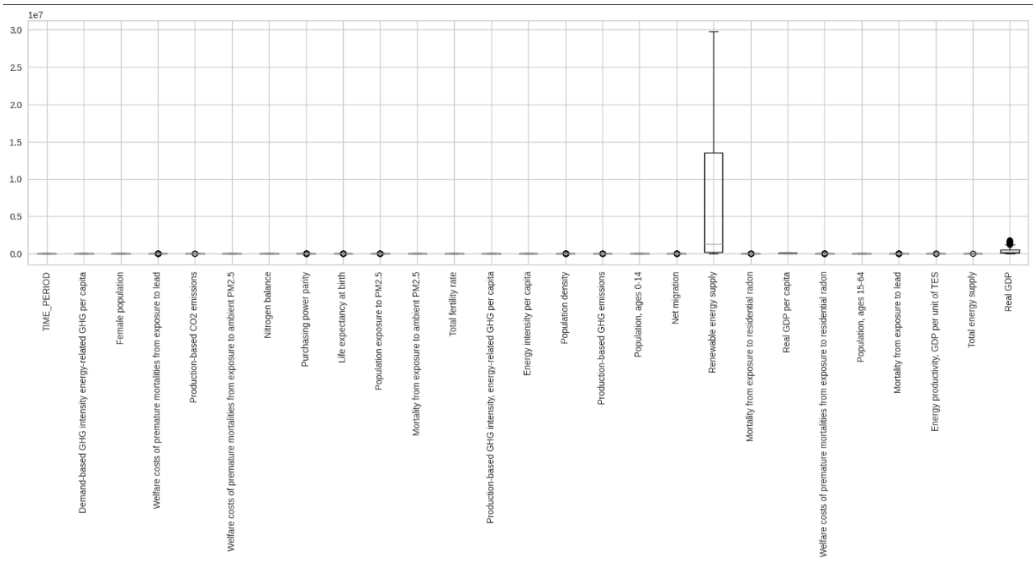
Gambar 2. Jumlah Missing Values

2) *Duplicate data*

Duplicate data adalah data yang identik atau sekumpulan data yang mengandung objek data yang sama (duplikat). Sealam proses pengolahan, seringkali terjadi tumpang tindih antara data (Abrar et al., 2023). Setelah dilakukan pengecekan pada data yang digunakan tidak terdapat duplikat data yang ditemukan.

3) *Outlier data*

Outlier data adalah titik data yang memiliki nilai yang jauh berbeda dibandingkan dengan nilai-nilai dalam populasi tertentu. Meskipun definisi ini terlihat sederhana, penentuan titik data yang dianggap outlier sebenarnya cukup subjektif dan bergantung pada penelitian serta jumlah data yang tersedia. Salah satu cara untuk mengidentifikasi data outlier adalah dengan menggunakan boxplot (Abrar et al., 2023). Setelah dilakukan pengecekan pada data ini masih ada yang memiliki outlier, akan tetapi hal tersebut tidak akan ditangani karena setiap negara memiliki perhitungan tersendiri dan penting untuk digunakan. Hasil pengecekan outliernya dapat dilihat pada gambar 3 di bawah ini:



Gambar 3. Outlier Data

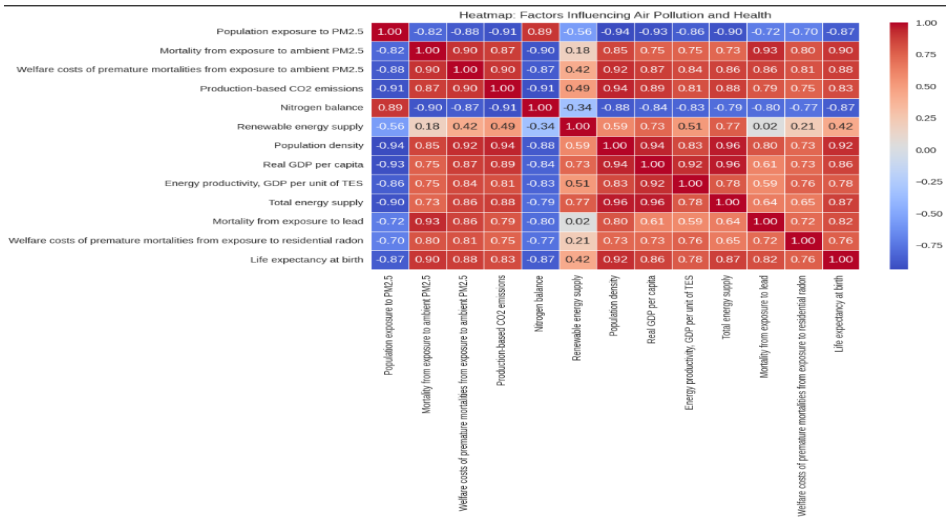
4) *Imbalance data*

Imbalance data merupakan masalah yang sering ditemui dalam klasifikasi machine learning, di mana distribusi antara kelas-kelas tidak seimbang atau tidak proporsional(Abrar et al., 2023). Pada tahap ini dilakukan pengecekan tipe data, dimana dari 27 kolom yang digunakan semuanya bertipe data float kecuali REF_AREA beripe data object dan TIME_PERIOD bertipe data int64. Lalu selanjutnya dicek untuk imbalance data pada REF_AREA dan semua jumlah datanya sama pada masing-masing negara dan seimbang satu sama lain.

5) Exploratory Data Analysis

Exploratory Data Analysis (EDA) adalah proses analisis data yang bertujuan untuk menggali, memahami, dan mengevaluasi karakteristik, pola, serta hubungan yang terdapat di dalam sebuah dataset secara mendalam dan intuitif. Tujuan utama dari EDA adalah memberikan wawasan awal yang mendalam tentang data, sehingga dapat dijadikan dasar yang kuat sebelum melanjutkan ke tahap analisis yang lebih kompleks, seperti penerapan model statistik atau algoritma machine learning (Merdiansah & Ali Ridha, 2024). Pada penelitian ini untuk memperoleh wawasan yang lebih dalam mengenai data menggunakan heatmap. Dimana nilai korelasi berkisar antara -1 hingga

1. Dikatakan memiliki hubungan positif kuat antara dua variabel apabila nilai mendekati 1, dikatakan hubungan negatif kuat antara dua variabel apabila nilai mendekati -1. Untuk melihat korelasi hubungan antar kolom yang memiliki pengaruh besar pada penelitian ini dapat dilihat pada gambar 4 yang ada di bawah ini:



Gambar 4. Heatmap Faktor yang Mempengaruhi Polusi Udara

4. *Modeling* dan *Evaluation*

Pada modeling ini dilakukan 2 model yaitu model regresi untuk pengisian nilai kosong dan forecasting untuk melakukan prediksi pada Population exposure to PM2.5 dan Welfare costs of premature mortalities from exposure to ambient PM2.5. Berikut adalah penjelasan masing-masing dari setiap modeling:

a) Model Regresi

Pada model regresi pada penelitian ini menggunakan library PyCaret, dimana library tersebut mampu melakukan berbagai proses, seperti penyiapan data, pembuatan model, perbandingan model, penyesuaian parameter, dan penyederhanaan model (Anwar & Permana, 2021). Ada 5 negara yang digunakan dalam pengisian nilai kosong ini namun semua tahapan dalam pengisiannya sama yang membedakan hanya negaranya saja, sehingga tidak semua dijelaskan dan penulis hanya menuliskan untuk satu negara yaitu Indonesia, dan berikut adalah tahapan yang dilakukan:

1. Pengujian Multikolinearitas dan VIF

   Pengujian ini digunakan untuk mengetahui adanya hubungan antara beberapa atau semua variabel pada model regresi. Dilanjutkan uji VIF (*Variance Inflation Factor*) dengan ketentuan nilai VIF masing-masing variabel tidak boleh > 10. Model tersebut diindikasikan memiliki gejala multikolinearitas apabila nilai VIF lebih besar dari 10 (Amalia Hufil Fadhila & Haryanti, 2020). Dari hasil tersebut variabel yang berkorelasi dengan variabel target yang telah ditentukan yang akan dipilih. Dimana pada penelitian ini memiliki 2 target kolom yang digunakan yaitu Welfare costs of premature mortalities from exposure to ambient PM2.5 dan Population exposure to PM2.5. Namun untuk pengujiannya tetap dilakukan secara terpisah.

2. Pengisian Nilai Kosong

   Untuk mengisi data yang hilang pada penelitian ini menggunakan interpolasi, yaitu sebuah teknik matematika yang bertujuan memperkirakan nilai di antara data yang sudah diketahui. Metode ini memungkinkan kita mengisi kekosongan dalam data

atau memperkirakan nilai di antara titik-titik yang telah diamati berdasarkan pola atau tren yang ada. Interpolasi digunakan untuk memprediksi nilai yang hilang dengan memanfaatkan informasi dari data di sekitarnya (Widianti & Pratama, 2024).

3. Melakukan scalling data

Pada tahap ini yang bertujuan untuk menormalisasi nilai-nilai dalam dataset, mengubahnya dari rentang yang besar menjadi skala antara 0 dan 1. Normalisasi ini berperan penting dalam algoritma machine learning karena dapat mengurangi perbedaan skala antar variabel, sehingga model dapat mempelajari pola data dengan lebih efektif dan efisien (Widianti & Pratama, 2024). Pada penelitian ini memiliki 2 target kolom yang digunakan yaitu target pertama Population exposure to PM2.5 dan target keduanya yaitu Welfare costs of premature mortalities from exposure to ambient PM2.5. Dimana hasil scalling datanya dapat dilihat pada gambar 5 dan 6 dibawah ini:

	Description	Value
0	Session id	123
1	Target	Population exposure to PM2.5
2	Target type	Regression
3	Original data shape	(22, 5)
4	Transformed data shape	(22, 5)
5	Transformed train set shape	(17, 5)
6	Transformed test set shape	(5, 5)
7	Numeric features	4
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Normalize	True
13	Normalize method	zscore
14	Fold Generator	KFold
15	Fold Number	5
16	CPU Jobs	-1
17	Use GPU	False
18	Log Experiment	False
19	Experiment Name	reg-default-name
20	USI	9bbd

Gambar 5. Hasil Scalling Data Target Ke 1

	Description	Value
0	Session id	123
1	Target	Welfare costs of premature mortalities from exposure to ambient PM2.5
2	Target type	Regression
3	Original data shape	(26, 5)
4	Transformed data shape	(26, 5)
5	Transformed train set shape	(20, 5)
6	Transformed test set shape	(6, 5)
7	Numeric features	4
8	Rows with missing values	3.8%
9	Preprocess	True
10	Imputation type	simple
11	Numeric imputation	mean
12	Categorical imputation	mode
13	Normalize	True
14	Normalize method	zscore
15	Fold Generator	KFold
16	Fold Number	5
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	reg-default-name
21	USI	bfc5

Gambar 6. Hasil Scalling Data Target Ke 2

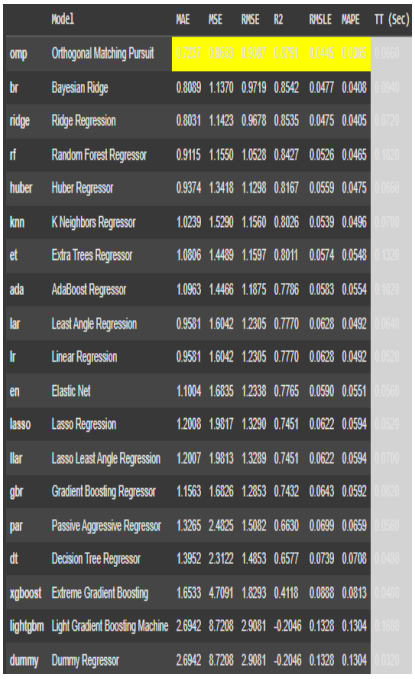
4. Membandingkan model terbaik

Setelah dilakukan scalling dilanjutkan membandingkan model untuk menemukan model terbaik. Dimana hasil perbandingan model terbaiknya dapat dilihat pada tabel 2 berikut:

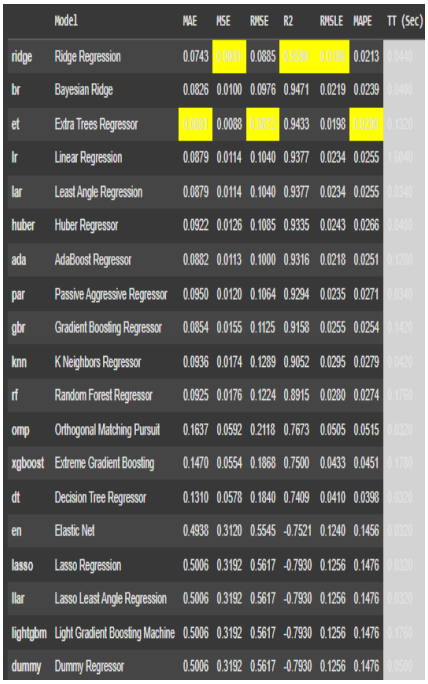
Tabel 2. Hasil Model Terbaik dari Masing-Masing Target

Target	population exposure to PM2.5	welfare costs of premature mortalities from exposure to ambient PM2.5
Model Terbaik	Orthogonal Matching Pursuit	Ridge Regression
MAE	0.7237	0.0743
MSE	0.8683	0.0081
RMSE	0.9087	0.0885
R2	0.8791	0.9590
RMSLE	0.0445	0.0196
MAPE	0.0365	0.0213

Untuk hasil perbandingannya lebih detailnya dapat dilihat pada gambar 7 dan 8 di bawah ini:



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
omp	Orthogonal Matching Pursuit	0.7237	0.8683	0.9087	0.8791	0.0445	0.0365	0.0660
br	Bayesian Ridge	0.8089	1.1370	0.9719	0.8542	0.0477	0.0408	0.0640
ridge	Ridge Regression	0.8031	1.1423	0.9678	0.8535	0.0475	0.0405	0.0720
rf	Random Forest Regressor	0.9115	1.1550	1.0528	0.8427	0.0526	0.0465	0.1620
huber	Huber Regressor	0.9374	1.3418	1.1298	0.8167	0.0559	0.0475	0.0660
knn	K Neighbors Regressor	1.0239	1.5290	1.1560	0.8026	0.0539	0.0496	0.0700
et	Extra Trees Regressor	1.0806	1.4409	1.1597	0.8011	0.0574	0.0548	0.1320
ada	AdaBoost Regressor	1.0963	1.4466	1.1875	0.7786	0.0583	0.0554	0.0820
lar	Least Angle Regression	0.9581	1.6042	1.2305	0.7770	0.0628	0.0492	0.0640
lr	Linear Regression	0.9581	1.6042	1.2305	0.7770	0.0628	0.0492	0.0620
en	Elastic Net	1.1004	1.6835	1.2338	0.7765	0.0590	0.0551	0.0560
lasso	Lasso Regression	1.2008	1.9817	1.3290	0.7451	0.0622	0.0594	0.0620
llar	Lasso Least Angle Regression	1.2007	1.9813	1.3289	0.7451	0.0622	0.0594	0.0700
gbr	Gradient Boosting Regressor	1.1563	1.6826	1.2853	0.7432	0.0643	0.0592	0.0620
par	Passive Aggressive Regressor	1.3265	2.4825	1.5082	0.6630	0.0699	0.0659	0.0660
dt	Decision Tree Regressor	1.3952	2.3122	1.4853	0.6577	0.0739	0.0708	0.0600
xgboost	Extreme Gradient Boosting	1.6533	4.7091	1.8293	0.4118	0.0888	0.0813	0.0680
lightgbm	Light Gradient Boosting Machine	2.6842	8.7208	2.9081	-0.2046	0.1328	0.1304	0.1600
dummy	Dummy Regressor	2.6842	8.7208	2.9081	-0.2046	0.1328	0.1304	0.0520



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
ridge	Ridge Regression	0.0743	0.0081	0.0885	0.9590	0.0196	0.0213	0.0440
br	Bayesian Ridge	0.0826	0.0100	0.0976	0.9471	0.0219	0.0239	0.0400
et	Extra Trees Regressor	0.0691	0.0088	0.0922	0.9433	0.0198	0.0200	0.1020
lr	Linear Regression	0.0879	0.0114	0.1040	0.9377	0.0234	0.0255	0.0440
lar	Least Angle Regression	0.0879	0.0114	0.1040	0.9377	0.0234	0.0255	0.0340
huber	Huber Regressor	0.0922	0.0126	0.1085	0.9335	0.0243	0.0266	0.0400
ada	AdaBoost Regressor	0.0882	0.0113	0.1000	0.9316	0.0218	0.0251	0.1280
par	Passive Aggressive Regressor	0.0950	0.0120	0.1064	0.9294	0.0235	0.0271	0.0340
gbr	Gradient Boosting Regressor	0.0854	0.0155	0.1125	0.9158	0.0255	0.0254	0.1420
knn	K Neighbors Regressor	0.0936	0.0174	0.1289	0.9052	0.0295	0.0279	0.0420
rf	Random Forest Regressor	0.0925	0.0176	0.1224	0.8915	0.0280	0.0274	0.1760
omp	Orthogonal Matching Pursuit	0.1637	0.0592	0.2118	0.7673	0.0505	0.0515	0.0320
xgboost	Extreme Gradient Boosting	0.1470	0.0554	0.1868	0.7500	0.0433	0.0451	0.1780
dt	Decision Tree Regressor	0.1310	0.0578	0.1840	0.7409	0.0410	0.0398	0.0320
en	Elastic Net	0.4938	0.3120	0.5545	-0.7521	0.1240	0.1456	0.0320
lasso	Lasso Regression	0.5006	0.3192	0.5617	-0.7930	0.1256	0.1476	0.0320
llar	Lasso Least Angle Regression	0.5006	0.3192	0.5617	-0.7930	0.1256	0.1476	0.0320
lightgbm	Light Gradient Boosting Machine	0.5006	0.3192	0.5617	-0.7930	0.1256	0.1476	0.1760
dummy	Dummy Regressor	0.5006	0.3192	0.5617	-0.7930	0.1256	0.1476	0.0560

Gambar 7. Hasil Model Terbaik Target ke 1 **Gambar 8.** Hasil Model Terbaik Target ke 2

5. Melakukan Hyperparameter Tuning
   Setelah membandingkan model selanjutnya dilakukan tuning. Dimana hasil setelah dituningnya dapat dilihat pada gambar di bawah ini:

Tabel 3. Hasil Perbandingan Setelah Hyperparameter Tuning

	Population exposure to PM2.5		**Welfare costs of premature mortalities from exposure to ambient PM2.5**	
	Before Tuning	***After Tuning***	***Before Tuning***	***After Tuning***
MAE	0.7237	0.5160	0.0743	0.0482
MSE	0.8683	0.4661	0.0081	0.0033
RMSE	0.9087	0.6827	0.0885	0.0576
R2	0.8791	0.8901	0.9590	0.9857
RMSLE	0.0445	0.0284	0.0196	0.0126
MAPE	0.0365	0.0227	0.0213	0.0136

b) Model Forecasting
   Pada model ini akan dilakukan prediksi untuk Population exposure to PM2.5 dan Welfare costs of premature mortalities from exposure to ambient PM2.5 untuk 8 tahun mendatang. Berikut adalah hasil dari prediksi dari masing-masing target:
   1. Population exposure to PM2.5
      Melakukan perbandingan dengan beberapa model dan model terbaiknya yaitu ARIMA, yang hasilnya dapat dilihat pada tabel 4 berikut:

Tabel 4. Perbandingan Model pada Target Population exposure to PM2.5

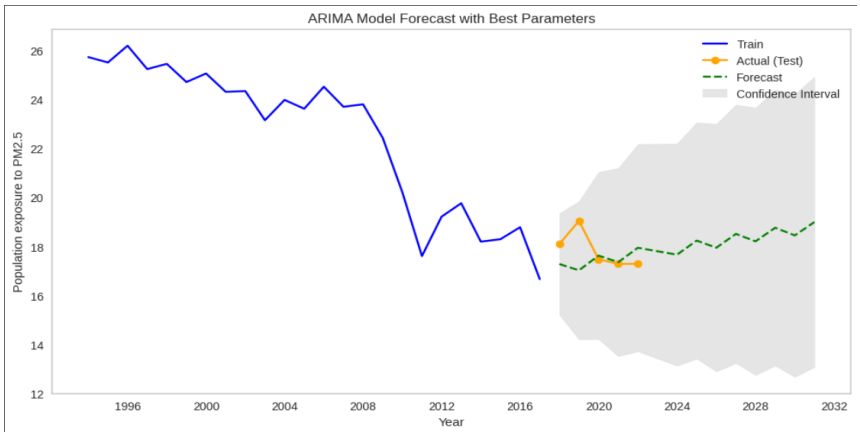
Model	MAE	MSE	MAPE
ARIMA	0.917837	1.449610	5.00%
ARIMAX	1.590548	2.654509	8.88%
ETS	1.198628	2.087193	6.47%
ARIMAX+ETS	2.905496	13.927375	15.696136%

Setelah melakukan perbandingan model dan mendapatkan hasil terbaiknya lalu dilakukan hyperparameter tuning, berikut adalah hasil yang didapatkan:

Tabel 5. Hasil Hyperparameter Tuning Model Terbaik pada Target

Evaluation	*Before Tuning*	*After Tuning*	*Difference*
MAE	0.918094	0.742105	0.175989
MSE	1.449878	1.041266	0.408612
MAPE	5.00%	4.04%	0.96%

Berikut adalah hasil prediksinya, yang menunjukkan bahwa akan terjadi peningkatan Population exposure to PM2.5 dari tahun ke tahun selama 8 tahun mendatang.



Gambar 9. Hasil Forecasting Population exposure to PM2.5

2. Welfare costs of premature mortalities from exposure to ambient PM2.5
   Melakukan perbandingan dengan beberapa model dan model terbaiknya yaitu ARIMAX+ETS, yang hasilnya dapat dilihat pada tabel 6 berikut:

Tabel 6. Hasil Perbandingan Model pada Target

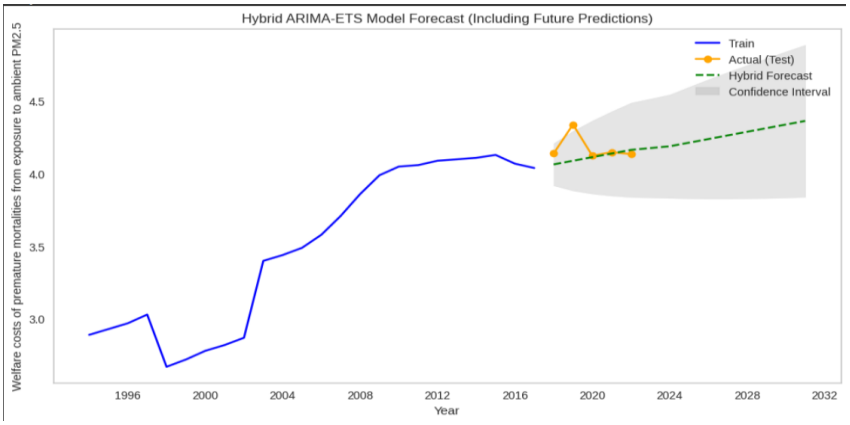
Model	MAE	MSE	MAPE
ARIMA	0.148507	0.028509	3.52%
ARIMAX	0.107348	0.020218	2.53%
ETS	0.111937	0.015741	2.66%
ARIMAX+ETS	0.014312	0.07698	1.80%

Setelah melakukan perbandingan model dan mendapatkan hasil terbaiknya lalu dilakukan hyperparameter tuning, berikut adalah hasil yang didapatkan:

Tabel 7. Hasil Hyperparameter Tuning dari Model Terbaik

Evaluation	*Before Tuning*	*After Tuning*	*Difference*
MAE	0.076980	0.074786	0.002194
MSE	0.014312	0.013845	0.000467
MAPE	1.803041%	1.750984%	0.52057

Berikut adalah hasil prediksinya, yang menunjukkan bahwa akan terjadi peningkatan welfare costs of premature mortalities from exposure to ambient PM2.5 dari tahun ke tahun selama 8 tahun mendatang.



Gambar 10. Hasil Forecasting Welfare costs of premature mortalities

Jadi dapat ditarik kesimpulan bahwa semakin meningkat population exposure to PM2.5 maka welfare costs of premature mortalities from exposure to ambient PM2.5 juga akan mengalami peningkatan.

5. *Deployment*

   Setelah dilakukan modeling didapatkan population exposure to PM2.5 diprediksi akan mengalami peningkatan untuk 8 tahun mendatang. Pencemaran udara oleh *Partikulat Matter* (PM), terutama PM2.5 dengan ukuran partikel di bawah 2,5 mikrometer, berpotensi membahayakan kesehatan global karena dapat masuk ke saluran pernapasan manusia dan memicu gangguan pernapasan, penyakit kardiovaskular, hingga meningkatkan risiko kematian dini (Yuwanda et al., 2024). Berikut ada beberapa rekomendasi kebijakan yang dapat diterapkan untuk mengurangi peningkatan population exposure to PM2.5 diantaranya (WHO, 2024):

   a. Mengadopsi teknologi ramah lingkungan untuk mengurangi emisi, meningkatkan pengelolaan limbah kota dan pertanian, serta memanfaatkan gas metana dari tempat pembuangan sampah untuk digunakan sebagai biogas.
   b. Memastikan akses masyarakat terhadap solusi energi bersih yang terjangkau untuk keperluan memasak, pemanasan, dan penerangan.
   c. Beralih ke pembangkit listrik bersih, memprioritaskan transportasi ramah lingkungan, serta mengganti kendaraan diesel berat dengan kendaraan rendah emisi.
   d. Meningkatkan efisiensi energi bangunan dan menciptakan kota hijau, padat, dan hemat energi.
   e. Menggunakan bahan bakar rendah emisi, energi terbarukan tanpa pembakaran, dan teknologi seperti kogenerasi serta tenaga surya atap.
   f. Menerapkan pengurangan limbah, daur ulang, dan teknologi pengelolaan limbah biologis, dengan pembakaran hanya sebagai opsi terakhir menggunakan kontrol emisi ketat.

g. Mengintegrasikan pembangunan rendah karbon dalam layanan kesehatan untuk menciptakan sistem yang tangguh, hemat biaya, dan ramah lingkungan.

KESIMPULAN

Penelitian ini menggunakan *machine learning* untuk memprediksi paparan PM2.5 dan dampaknya terhadap kesehatan masyarakat di Indonesia. Proses analisis mencakup pemahaman bisnis, persiapan data, hingga pemodelan menggunakan regresi untuk pengisian data yang hilang dan forecasting untuk prediksi jangka panjang. Berdasarkan hasil penelitian menunjukkan bahwa peramalan untuk 8 tahun ke depan, paparan PM2.5 diperkirakan terus meningkat, yang akan berkontribusi langsung pada kenaikan biaya kesejahteraan dari dampak kesehatan, seperti gangguan pernapasan dan kematian dini. Pada penelitian ini model terbaik untuk memprediksi paparan PM2.5 adalah ARIMA, sedangkan untuk biaya kesejahteraan adalah ARIMAX+ETS, masing-masing menunjukkan peningkatan akurasi setelah hyperparameter tuning. Mengacu pada hasil ini, diperlukan implementasi rekomendasi kebijakan seperti adopsi teknologi ramah lingkungan, transisi ke sumber energi bersih, pengelolaan limbah yang lebih baik, serta pembangunan kota hijau yang hemat energi. Langkah-langkah ini diharapkan dapat menekan polusi udara dan dampak kesehatannya, sekaligus mengurangi beban ekonomi akibat peningkatan biaya kesejahteraan di masa depan.

SARAN

Berdasarkan hasil penelitian yang telah dilakukan, maka ada beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya:

1. Penggunaan algoritma selain yang telah digunakan pada penelitian ini.
2. Memprediksi untuk jangka yang lebih panjang lagi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak penyelenggara MSIB yang telah membuat program ini, serta pihak penyelenggara program MBKM Universitas Amikom Purwokerto, khusunya Fakultas Ilmu Komputer yang telah mendukung adanya program ini. Tidak lupa, terima kasih juga kepada dosen pembimbing MBKM penulis yang telah membimbing selama program MBKM ini dan teman-teman yang telah terlibat dan membantu serta memberi dukungan kepada penulis sehingga penelitian ini dapat terselesaikan.

DAFTAR PUSTAKA

Abrar, I. N., Abdullah, A., & Sucipto, S. (2023). Liver Disease Classification Using the Elbow Method to Determine Optimal K in the K-Nearest Neighbor (K-NN) Algorithm. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *12*(2), 218–228. https://doi.org/10.32736/sisfokom.v12i2.1643

Amalia Hufil Fadhila, & Haryanti, P. (2020). Pengaruh Profitabilitas, Islamic Governance Score, Dan Ukuran Bank Terhadap Pengungkapan Islamic Sosial Reporting (Isr) Pada Bank Umum Syariah Di Indonesia. *Malia (Terakreditasi)*, *11*(2), 187–206. https://doi.org/10.35891/ml.v11i2.1872

Anwar, M. T., & Permana, D. R. A. (2021). Perbandingan Performa Model Data Mining untuk Prediksi Dropout Mahasiwa. *Jurnal Teknologi Dan Manajemen*, *19*(2), 33–40. https://doi.org/10.52330/jtm.v19i2.34

Haryanto, C., Rahaningsih, N., & Muhammad Basysyar, F. (2023). Komparasi Algoritma Machine Learning Dalam Memprediksi Harga Rumah. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 533–539. https://doi.org/10.36040/jati.v7i1.6343

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Kemenkes. (2024). Bahaya Polusi Udara bagi Kesehatan: Dampak, Penyebab dan Pencegahannya. *Ayosehat.Kemkes.Go.Id*. https://ayosehat.kemkes.go.id/bahaya-polusi-udara-bagi-kesehatan

Merdiansah, R., & Ali Ridha, A. (2024). Sentiment Analysis of Indonesian X Users Regarding Electric Vehicles Using IndoBERT. *Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI*, *7*(1), 221–228.

Rahmawati, S., & Pratama, I. N. (2023). Pengaruh Penggunaan Transportasi Berkelanjutan Terhadap Kualitas Udara Dan Kesejahteraan Masyarakat. *Journal of Enviromental Policy and Technology*, *1*(2), 90–99.

Salsabila, Wardah Nibras, Y., & Sudarti. (2023). Analisis Perkembangan Penanggulangan Pencemaran Udara Yang Disebabkan Oleh Bahan Bakar Fosil. *Jurnal Pendidikan, Sains Dan Teknologi*, *2*(4), 1010–1014. https://doi.org/10.47233/jpst.v2i4.1331

Siti Nurjanah, Yoan Purbolingga, Dila Marta Putri, Asde Rahmawati, Fahrizal Fahrizal, & Bastul Wajhi Akramunnas. (2024). Prediksi Kecepatan Angin untuk Mengetahui Potensi Sumber Energi Alternatif menggunakan Model Regresi Lasso: Studi Kasus Kota Makassar pada Tahun 2024. *Jurnal Penelitian Rumpun Ilmu Teknik*, *3*(1), 278–288. https://doi.org/10.55606/juprit.v3i1.3501

Sumiyati. (2024). Indonesia Peringkat 14, Negara dengan Tingkat Polusi Udara Tertinggi di Dunia. *VIVA.Co.Id*. https://www.viva.co.id/gaya-hidup/kesehatan-intim/1752796-indonesia-peringkat-14-negara-dengan-tingkat-polusi-udara-tertinggi-di-dunia

Syahfutris, Y., Rasyid Munthe, I., Zuhri Harahap, S., Sains dan Teknologi, F., & Sistem, P. (2023). Analisis Machine Learning Algoritma Regresi Linear Untuk Memprediksi Saham Di Bank Bri Di Bursa Saham Indonesia. *Jurnal Tekinkom (Teknik Informasi Dan Komputer)*, *6*(1), 81–87. https://doi.org/10.37600/tekinkom.v6i1.747

WHO. (2024). Ambient (outdoor) air pollution. *Www.Who.Int*. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Widianti, A., & Pratama, I. (2024). Penanganan Missing Values Dan Prediksi Data Timbunan Sampah Berbasis Machine Learning. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *9*(2), 242–251. https://doi.org/10.36341/rabit.v9i2.4789

Yuwanda, A., Budiatama, A., & Yusuf, D. E. (2024). Edukasi Mengenai Dampak Buruk Polusi Partikulat Matter (PM) 2,5 Terhadap Gangguan Kognitif pada Siswa Sekolah SMK Kesehatan Bhakti Insani. *Pengabdian Kepada Masyarakat)*, *1*(1), 7–11. https://doi.org/10.70608/thm8wf06