Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Artikel Penelitian

Perbandingan Performa Algoritma Metode Bagging dan Boosting pada Prediksi Konsentrasi PM_{10} di Jakarta Utara

Elita Rizkiani Putri [a,*], *Dede Brahma Arianto* [b]

[a] *Program Studi Kesehatan Lingkungan, Fakultas Kesehatan Masyarakat, Universitas Indonesia, Depok, 16424, Indonesia*
[a] *Magister Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 14 November 2023
Revisi Akhir: 06 Februari 2024
Diterbitkan *Online*: 17 Mei 2024

KATA KUNCI

PM_{10},
Random Forest,
Catboost,
XGBoost

KORESPONDENSI

E-mail: elita.rizkiani@ui.ac.id*

A B S T R A C T

Jakarta Utara merupakan salah satu wilayah di DKI Jakarta yang mengalami peningkatan hari dengan kualitas udara berkategori tidak sehat, yakni 21 hari pada tahun 2017 menjadi 117 hari di 2018, tetapi kemudian menurun menjadi 45 hari pada tahun 2019. Kategori tidak sehat tersebut dipengaruhi oleh polusi udara. Salah satu polutan yang ada di udara adalah PM_{10}. Saat ini, kualitas udara dapat diprediksi menggunakan pendekatan algoritma *machine learning*. Contoh metode *machine learning* yang terkenal adalah Metode Bagging dan Boosting yang ada di Metode Ensemble. Contoh algoritma dengan Metode Bagging adalah Random Forest, sedangkan pada Metode Boosting adalah Catboost dan XGBoost. Penelitian ini bertujuan membandingkan performa algoritma Metode Bagging berupa Random Forest dan algoritma Metode Boosting berupa Catboost dan XGBoost dalam memprediksi konsentrasi PM_{10} di Jakarta Utara. Data yang digunakan adalah data harian tahun 2017—2019 untuk faktor meteorologis dan polutan lainnya di wilayah tersebut. Faktor meteorologis digunakan karena faktor ini dapat memengaruhi konsentrasi dan pembentukan polutan. Sementara itu, faktor polutan digunakan karena beberapa penelitian sebelumnya menggunakan faktor ini dalam memprediksi konsentrasi PM_{10}. Penelitian ini dilakukan dengan studi literatur, pemerolehan data, pra-pemprosesan data, dan pemodelan data. Beberapa metrik evaluasi juga digunakan untuk melihat evaluasi dari pemodelan. Berdasarkan hasil pemodelan, algoritma Random Forest menghasilkan akurasi *data testing* yang lebih tinggi (R^2 = 0,6424) dibandingkan XGBoost (R^2 = 0,6340) dan Catboost (R^2 = 0,6294).

1. PENDAHULUAN

Polusi udara merupakan sekumpulan zat-zat berbahaya yang bersumber dari kegiatan manusia atau alam [1]. Kemunculan polusi udara terjadi saat konsentrasi pencemar di udara melampaui batas yang seharusnya dan memengaruhi kehidupan makhluk hidup [2]. Kondisi ini dapat menyebabkan permasalahan kesehatan. Dalam jangka pendek, polusi dapat menyebabkan penyakit yang berkaitan dengan paru-paru, sedangkan dalam jangka panjang dapat menyebabkan kanker dan kematian pada bayi [2]. Kualitas udara dipengaruhi oleh sumber emisi, pembentukan polutan sekunder, dan faktor meteorologi [3]. Di wilayah perkotaan, sumber emisi tertinggi dihasilkan dari transportasi, industri, pembangkit listrik, dan permukiman akibat perkembangan urbanisasi [4].

Particulate matter merupakan salah satu agen kimia berbentuk partikel yang tersusun atas partikel padat dan cairan *droplet* yang kerap berkontribusi dan berperan terhadap indikator polusi udara [5], [6]. Polutan ini diemisikan dari pabrik, pembangkit listrik, tempat insinerasi, tempat konstruksi, api, dan debu kendaraan [2]. Berdasarkan ukurannya, partikel ini terbagi menjadi $PM_{2.5}$ yang berdiameter ≤ 2.5 µm dan PM_{10} yang berdiameter ≤ 10 µm. PM_{10} sebagai polutan konsentrasinya dipengaruhi oleh faktor meteorologi, yang mana faktor ini sangat berperan pada kualitas udara di kota maupun pedesaan dan berpengaruh pada pembentukan dan besaran konsentrasi polutan tersebut [7], [8]. Penelitian sebelumnya menunjukkan bahwa faktor meteorologi seperti suhu, kelembapan, dan kecepatan angin memengaruhi konsentrasi PM_{10} [8], [9].

https://doi.org/10.25077/TEKNOSI.v10i1.2024.74-80

Jakarta merupakan wilayah padat penduduk dengan beragam sumber emisi. Jika diurutkan dari yang terbesar, emisi wilayah ini bersumber dari transportasi, industri, pembangkit listrik, dan permukiman [4]. Tahun 2015, penelitian sebelumnya menyebutkan, sebanyak 6 miliar gram emisi PM_{10} sebagian besar disumbang oleh transportasi (43%) dan industri (46%) [4]. Jakarta Utara merupakan wilayah yang lahannya digunakan untuk industri dan pelabuhan [10]. Tahun 2017—2018, Jakarta Utara merupakan salah satu wilayah di DKI Jakarta yang mengalami penurunan kualitas udara. Hal ini ditandai dengan peningkatan hari kualitas udara berkategori tidak sehat, yakni 21 hari pada tahun 2017 menjadi 117 hari di 2018 [11]. Namun, berdasarkan data Dinas Lingkungan Hidup DKI Jakarta, jumlahnya berkurang di tahun 2019, yakni menjadi 45 hari [12].

Saat ini, kualitas udara dapat diprediksi menggunakan algoritma *machine learning*, caranya dengan membuat pemodelan suatu konteks masalah sesuai data yang dimiliki, kemudian hasil prediksi yang sensitif dapat dihasilkan [2]. Prosesnya, variabel independen atau fitur berperan sebagai *input* untuk melihat hubungan terhadap variabel dependen atau target yang akan diprediksi [13]. Dalam prediksi kualitas udara, beberapa algoritma dalam kelompok Metode Ensemble karena ketenaran dan penerapannya dapat digunakan [13]. Sementara itu, dua kelompok Metode Ensemble yang terkenal adalah *bagging* dan *boosting* [14]. Algoritma yang menerapkan konsep *bagging* adalah Random Forest, sedangkan algoritma yang menerapkan konsep *boosting* contohnya Catboost dan XGBoost [13]–[15]. Penelitian dengan 3 algoritma tersebut telah dilakukan sebelumnya oleh [16] terhadap konsentrasi $PM_{2.5}$ dan O_3 dengan data meteorologi (suhu, titik embun, kelembapan relatif, tekanan permukaan, radiasi matahari, curah hujan, kecepatan angin, dan arah angin) dan polutan ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, dan O_3). Hasilnya menunjukkan R^2 *score data testing* sebesar 0,912 pada XGBoost, sebesar 0,923 pada Catboost, dan sebesar 0,869 pada Random Forest. Penelitian prediksi konsentrasi PM_{10} oleh [17] di kota Kandy dan Battaramulla di Sri Lanka menggunakan XGBoost dan Catboost menunjukkan R^2 *score* kedua algoritma sebesar >0,98. Sementara itu, penelitian oleh [13] di Stuttgart, Jerman yang juga memprediksi konsentrasi PM_{10} (menggunakan input temporal, meteorologi, lalu lintas, dan polutan dari stasiun pengukuran) menghasilkan R^2 *score* sebesar 0,8 dengan algoritma Random Forest. Penelitian lainnya berupa prediksi $PM_{2.5}$ di Taiwan yang dilakukan oleh [18] dengan algoritma Random Forest menggunakan data meteorologi dan polutan menunjukkan R^2 *score data testing* sebesar 0,8699.

Seperti contoh penelitian yang disebutkan, Metode Bagging dengan algoritma Random Forest dan Metode Boosting dengan algoritma Catboost dan XGBoost dapat digunakan dalam prediksi kualitas udara. Sementara itu, penelitian sebelumnya menyebutkan bahwa prediksi konsentrasi PM_{10} dapat menggunakan faktor polutan sebagai *input* [19]. Selain itu, faktor meteorologi juga berpengaruh terhadap konsentrasi atau pembentukan PM_{10}. Oleh karenanya, penelitian ini bertujuan membandingkan performa algoritma Random Forest pada Metode Bagging dengan algoritma Catboost dan XGBoost pada Metode Boosting untuk memprediksi konsentrasi PM_{10} di Jakarta Utara berdasarkan faktor polutan dan meteorologi.

2. METODE

2.1. Tinjauan Pustaka

2.1.1. Faktor Meteorologi Terhadap Konsentrasi PM_{10}

Faktor meteorologi merupakan hal yang memengaruhi kualitas udara, baik karena satu atau bersamaan dengan faktor meteorologi lainnya [20]. Pengaruhnya terkait dengan pembentukan dan perubahan konsentrasi polutan [7]. Menurut [21], kondisi meteorologi memengaruhi massa polutan udara dengan jumlah besar dengan cara menyebarkan, mengencerkan, dan mengumpulkannya. Hasil penelitian [9] menyatakan kelembapan relatif, kecepatan angin, dan suhu berbanding terbalik dengan konsentrasi PM_{10}, baik di musim kemarau maupun musim hujan. Dengan kata lain, semakin tinggi kelembapan relatif, kecepatan angin, dan suhu, semakin rendah konsentrasi PM_{10}, begitupun sebaliknya. Kondisi ini terjadi akibat kecepatan angin memengaruhi penyebaran dan konsentrasi polutan tersebut [22]. Sementara itu, kelembapan memengaruhi penguapan konsentrasi PM_{10} karena penambahan partikel uap air mendorong deposisi kering, yakni pengendapan polutan ke permukaan bumi [23], [24]. Berkebalikan dengan sebelumnya, penelitian lain menyebutkan bahwa suhu berkorelasi positif terhadap konsentrasi PM_{10} [25], [26]. Sebagai contoh, penelitian oleh [26] menemukan penambahan suhu beriringan dengan penambahan konsentrasi PM_{10}, tetapi hal ini berlaku jangka pendek dan disebabkan polutan dihasilkan dari pemanas rumah atau bahan bakar.

2.1.2. Polutan Sebagai Input Pemodelan

Penelitian prediksi dan analisis konsentrasi $PM_{2.5}$ dan PM_{10} menggunakan 4 algoritma menyebutkan adanya peningkatan akurasi prediksi saat faktor polutan udara (SO_2, CO, O_3, dan NO_2) dan meteorologi diikutsertakan dalam pemodelan [19]. Selain itu, prediksi konsentrasi PM_{10} dengan menerapkan 7 algoritma oleh [27] menggunakan data polutan dan faktor meteorologi (tanpa mempertimbangkan faktor musim) dapat bekerja dengan baik.

2.1.3. Pemodelan

Metode Ensemble adalah metode yang berguna untuk mencegah kelemahan pada sebuah model dengan cara mengintegrasikannya [28]. Bagging dan Boosting merupakan dua metode dalam kelompok Metode Ensemble. Bagging merupakan metode untuk mereduksi *overfitting* yang bekerja dengan memisahkan dan melatih setiap model, kemudian merata-ratakan *output* semua model untuk mengurangi varians [28]. Algoritma yang mengimplementasikan konsep ini adalah Random Forest. Sementara itu, Boosting merupakan metode pembuatan model yang dilakukan secara berulang, yakni dengan urutan membangun model dasar, memperkirakan perbedaan nilai prediksi dengan nilai asli, kemudian

membangun model selanjutnya berdasarkan perbedaan nilai tersebut hingga perbedaan tersebut memiliki nilai *error* paling kecil [28]. Algoritma yang mengimplementasikan konsep ini adalah XGBoost dan Catboost.

Saat praktiknya, algoritma yang disebutkan sebelumnya menggunakan *hyperparameter*. *Hyperparameter* digunakan untuk mengatur proses *machine learning* dan menentukan parameter akhir dari model [29]. *Hyperparameter* yang berubah akan berpengaruh terhadap performa algoritma yang digunakan [29]. Meskipun demikian, hal ini dapat diatasi dengan Grid Search, yakni pendekatan konfigurasi model dengan memilih nilai-nilai *hyperparameter* yang ingin dieksplorasi, kemudian nilai-nilai tersebut dikombinasikan saat melatih dan menguji model untuk mengetahui nilai mana yang terbaik untuk model yang diuji [30]. Pencarian *hyperparameter* terbaik ini juga diiringi dengan tahap penting berupa *cross-validation*, yakni metode statistik yang memperkirakan akurasi dari algoritma *machine learning* [31]. Salah satu metode yang bisa digunakan adalah K-Fold, yang mana parameter K merupakan jumlah lipatan yang dibagi menjadi sekumpulan data tertentu [31].

Catboost merupakan algoritma berbasis GBDT (*gradient-boosting decision tree*) yang dapat digunakan untuk klasifikasi, regresi, dan prediksi [17]. Kemampuan algoritma ini dipengaruhi oleh pengaturan *hyperparameter*-nya [15]. Kelebihan algoritma ini adalah mampu mengurangi bias akibat kebocoran data saat *data training*, meningkatkan tingkat keakuratan model, mengurangi *overfitting*, dan meningkatkan kecepatan prediksi [32].

XGBoost atau Extreme Gradient Boosting merupakan algoritma yang juga berbasis GBDT. Cara kerja algoritma ini ialah dengan menngomputasikan fitur atau variabel independen yang penting, membaginya menjadi kelompok kecil sesuai nilainya, kemudian kelompok kecil terus dibagi sampai model yang dibuat tidak menghasilkan peningkatan atau kriteria yang ditentukan sudah tercapai, tujuannya untuk memprediksi target atau variabel dependen yang numerik dan kontinu pada setiap iterasinya berdasarkan variabel independen yang dimiliki [33]. Performa model ini meningkat karena selama iterasinya ditambahkan *decision tree* dengan proses Boosting untuk mengoreksi *error* dari *decision tree* yang sebelumnya dibuat [33].

Random Forest merupakan algoritma dengan konsep *bagging* dan berdasar *decision tree* atau pohon keputusan. Pohon-pohon keputusan yang ada pada algoritma ini mengambil sampel dari *data training*, sedangkan data sisanya dipakai untuk mengetahui kesalahan pada *decision tree* [34]. Akurasi pada algoritma ini dipengaruhi oleh [35]:

a. Banyaknya pohon prediktor;
b. Kedalaman setiap pohon;
c. *Bootstrap* untuk mereduksi variansi pohon prediktor pada *data training*; dan
d. Kriteria, yakni menentukan *square error* karena mempengaruhi nilai *mean square error* (MSE).

Untuk mengevaluasi pemodelan yang dilakukan, metrik evaluasi digunakan, di antaranya R Square (R^2), Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), dan MAPE. R^2 adalah pengukuran bagaimana model dapat menjelaskan variasi pada variabel dependen, MAE adalah pengukuran untuk melihat besar perbedaan antara nilai sebenarnya dengan nilai prediksi, MAPE merupakan persentase deviasi dari nilai yang diprediksi terhadap nilai asli, dan RMSE adalah akar dari Mean Square Error, yakni akar dari rata-rata kuadrat deviasi dari nilai sebenarnya terhadap nilai prediksi [36], [37]. Semakin kecil nilai MAE, MSE, MAPE, dan RMSE, semakin baik model yang digunakan, sedangkan semakin besar nilai R^2, semakin baik model yang digunakan. Dalam penelitian ini, evaluasi metrik yang menjadi acuan adalah R^2, sebab berdasarkan [38], R^2 sebaiknya digunakan untuk evaluasi analisis regresi di berbagai bidang keilmuan. Nilai berkisar R^2 ialah antara 0 hingga 1. R^2 yang mendekati 0 menandakan variabel independen kurang menjelaskan variabel dependen, sedangkan R^2 yang mendekati 1 berarti variabel independen semakin mampu menjelaskan perubahan variabel dependen [39].

2.2. Data

Variabel independen atau fitur pada penelitian ini berupa faktor meteorologis, di antaranya suhu rata-rata, kelembapan rata-rata, dan kecepatan angin rata-rata. Variabel independen lainnya yang dipertimbangkan adalah konsentrasi polutan, di antaranya SO_2, CO, O_3, dan NO_2. Sementara itu, variabel dependen atau target penelitian ini adalah konsentrasi PM_{10}. Variabel-variabel yang dipilih ini berdasarkan studi literatur dan ketersediaan data yang dapat diakses melalui laman resmi pemerintah.

2.2.1. Pemerolehan Data

Data yang digunakan pada penelitian ini didasarkan atas variabel yang berhubungan dengan konsentrasi PM_{10} berdasarkan studi literatur dan ketersediaan data yang ada di laman resmi pemerintah. Sumber data berupa data sekunder dalam bentuk data harian dari tahun 2017–2019 yang dapat diakses secara *online* melalui laman Badan Klimatologi Meteorologi dan Geofisika (BMKG) (tautan: [40]) untuk faktor meteorologis dan Open Data Jakarta (tautan: [12], [41], [42]) untuk konsentrasi polutan. Data faktor meteorologis dipilih berdasarkan Stasiun Meteorologi Maritim Tanjung Priok, sedangkan data konsentrasi polutan dipilih berdasarkan Stasiun Pemantauan Kualitas Udara (SPKU) Kelapa Gading.

Pemilahan dan pengelompokkan data iklim dan konsentrasi polutan untuk analisis dilakukan pada Microsoft Excel, sedangkan penggabungan dua data tersebut dilakukan pada Jupyter Notebook sekaligus sebagai media analisis dan pemodelan. Jumlah dataset setelah penggabungan menjadi 1095 baris dengan 11 kolom. Library pendukung yang digunakan di antaranya Numpy, Pandas, Matplotlib, Seaborn, Scipy, dan Scikit-learn.

https://doi.org/10.25077/TEKNOSI.v10i1.2024.74-81

2.2.2. *Pra-Pemprosesan Data*

Kegiatan pada tahap ini berupa pengecekan dan penanganan *missing values*, pembersihan kolom yang tidak diperlukan, dan *exploratory data analysis*. Pengecekan *missing values* dilakukan untuk mengetahui banyaknya nilai kosong pada kolom, sedangkan penanganannya dilakukan dengan *drop* baris. Sementara itu, pembersihan kolom-kolom dataset yang tidak diperlukan dilakukan untuk mempermudah analisis data. Jumlah data setelah dua tahap tersebut dilakukan adalah 960 baris dengan 8 kolom. Tabel 1. menunjukkan detail kolom setelah dataset dibersihkan.

Tabel 1. Kolom Dataset.

Nama Kolom	Keterangan
pm10	Konsentrasi PM_{10}
so2	Konsentrasi SO_2
co	Konsentrasi CO
o3	Konsentrasi O_3
no2	Konsentrasi NO_2
Tavg	Suhu rata-rata
RH_avg	Kelembapan rata-rata
ff_avg	Kecepatan angin rata-rata
tanggal	Tanggal pengukuran

Kemudian, *exploratory data analysis* dilakukan dengan analisis univariat dan analisis bivariat. Analisis univariat dilakukan untuk mengetahui distribusi data, sedangkan analisis bivariat dilakukan untuk menguji normalitas setiap variabel dan melihat korelasi. Menurut [43] uji normalitas dapat dilakukan dengan Shapiro-Wilk jika sampel <50, sedangkan untuk sampel >50 menggunakan Kolmogorov Smirnov. Uji normalitas yang digunakan pada penelitian ini menggunakan Kolmogorov Smirnov karena jumlah sampel atau data sebanyak 960 baris. Korelasi variabel dilakukan dengan Korelasi Spearman, sebab seluruh variabel berdistribusi tidak normal. Korelasi Spearman adalah uji statistik nonparametrik untuk melihat hubungan monotonik antara dua atau satu variabel yang sudah dilakukan pemeringkatan terhadap variabel yang diukur [44], [45]. Hal ini menyebabkan pemeringkatan pada data dilakukan sebelumnya, sebab korelasi data yang distribusinya tidak normal, seperti data yang memiliki nilai ekstrem atau *outliers*, perlu dilakukan pemeringkatan data selain dari nilai aslinya. Korelasi ini menghasilkan nilai ρ yang diperoleh dengan Persamaan (1) [46].

$$\rho = \frac{6\sum d_i^2}{n(n^2-1)} \quad (1)$$

Setelah korelasi dilakukan, grafik tren data dibuat. Kemudian, variabel-variabel yang sudah diproses masuk ke tahap pemodelan, yakni variabel dependen atau target (konsentrasi PM_{10}) dan variabel independen atau fitur (SO_2, CO, O_3, suhu rata-rata, kelembapan rata-rata, dan kecepatan angin rata-rata).

Sebelum masuk tahap pemodelan, pembagian dataset dilakukan. Kolom 'tanggal' dikecualikan dari tahap pembagian dataset ini. Data pemodelan sebanyak 960 baris dengan 8 kolom dibagi menjadi 80% untuk *training* dan 20% untuk *testing*. Angka 80 dan 20 dipilih karena angka-angka tersebut merupakan rasio yang umum digunakan untuk pembagian data *testing* dan *training* [47]. Saat pembagian, 'random_state = 42' digunakan agar pemilihan data acak selalu tetap.

2.2.3. *Pemodelan Data*

Sebelum masing-masing algoritma pemodelan diterapkan, pencarian *hyperparameter* terbaik untuk setiap model dilakukan dengan bantuan GridSearchCV dan *cross validation* dengan bantuan KFold dari *library* Scikit-learn. Nilai *cross validation* yang digunakan untuk pencarian *hyperparameter* setiap model adalah 5. Tabel 2. menunjukkan pilihan nilai untuk mencari *hyperparameter* terbaik. Khusus Random Forest, penyetelan *hyperparameter* ditambahkan 'random_state' agar pengacakan data dan hasil metrik evaluasi selalu konstan saat sintaks dijalankan berulang.

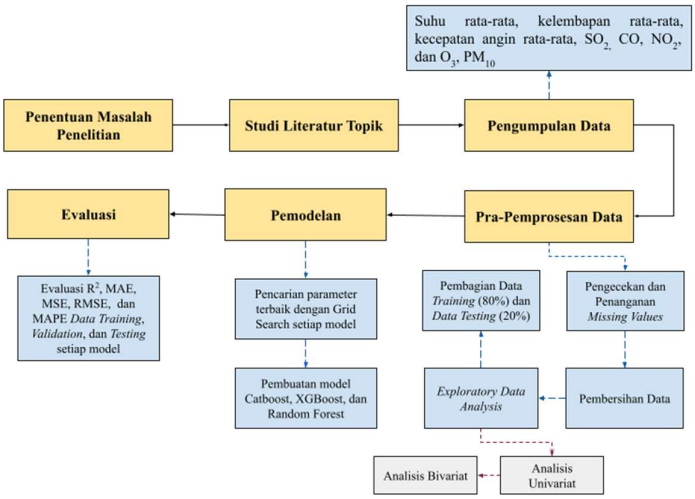
Tabel 2. Parameter Algoritma Catboost, XGBoost, dan Random Forest.

Algoritma	*Hyperparameter*	Nilai
Catboost	depth	4, 6, 8, 10
	learning_rate	0.01, 0.05, 0.1, 0.2
	iterations	100, 200, 300, 500
	min_child_samples	2, 5, 10
	l2_leaf_reg	1, 2, 3
XGBoost	max_depth	4, 6, 8, 10
	learning_rate	0.01, 0.05, 0.1, 0.2
	n_estimators	100, 200, 300, 500
	min_child_weight	2, 5, 10
	reg_lambda	1, 2, 3
Random Forest	max_depth	4, 6, 8, 10
	max_features	'log2', 'sqrt'
	n_estimators	100, 200, 300, 500
	min_samples_split	2, 5, 10
	min_samples_leaf'	1, 2, 4
	'random_state'	0, 42, 80, 90

Setelahnya, nilai *hyperparameter* terbaik yang diperoleh digunakan untuk masing-masing algoritma.

2.3. Alur Penelitian

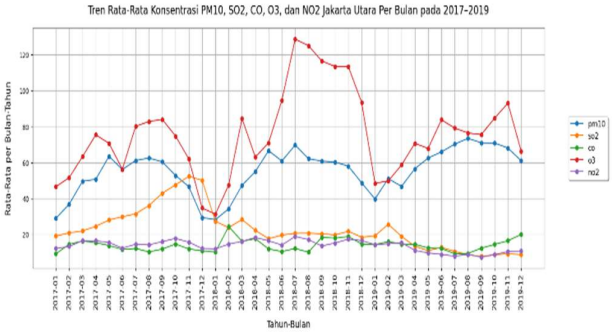
Secara keseluruhan, penelitian dilakukan dengan alur atau tahapan seperti ditunjukkan pada Gambar 1.



Gambar 1. Alur Penelitian "Perbandingan Performa Algoritma Bagging dan Boosting pada Prediksi Konsentrasi PM_{10} di Jakarta Utara".

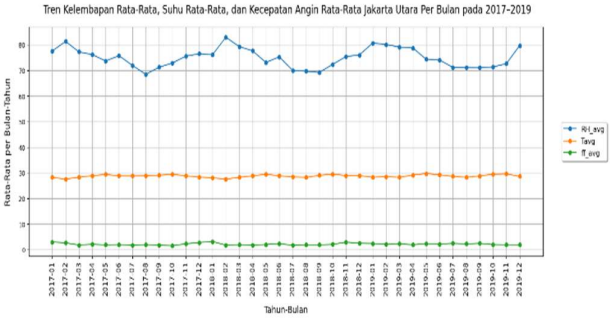
3. HASIL

Berdasarkan hasil analisis konsentrasi polutan, tren konsentrasi PM_{10} di Jakarta Utara cenderung mengalami kenaikan dari tahun 2017–2018. Berkebalikan dengan PM_{10}, tren SO_2 cenderung menurun dari 2017—2019. Sementara itu, tren kenaikan konsentrasi O_3 sangat terlihat pada April–Juli 2018. Detail tren konsentrasi polutan dapat dilihat pada Gambar 2.



Gambar 2. Tren Kelembapan Rata-Rata, Suhu Rata-Rata, dan Kecepatan Angin Rata-Rata Jakarta Utara Per Bulan pada 2017–2019.

Berdasarkan analisis faktor meteorologi, tren kelembapan rata-rata di Jakarta Utara memiliki pola yang sama pada tahun 2017–2019. Pola yang cenderung sama juga terlihat pada suhu rata-rata dan kecepatan angin rata-rata. Detail tren faktor meteorologi dapat dilihat pada Gambar 3.



Gambar 3. Tren Rata-Rata Konsentrasi PM_{10}, SO_2, CO, O_3, dan NO_2 di Jakarta Utara Per Bulan pada 2017–2019.

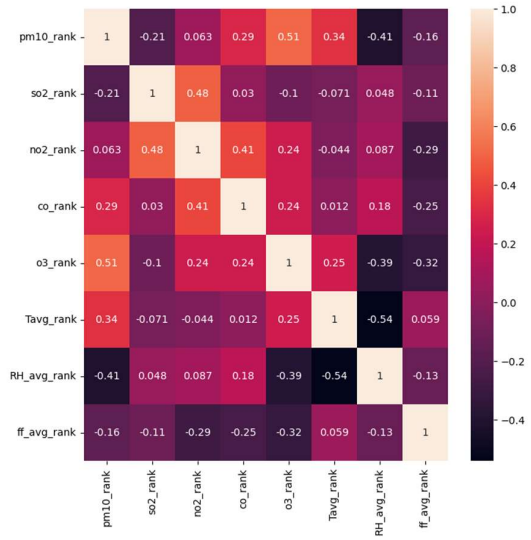
Selain tren data, analisis univariat dilakukan terhadap variabel penelitian. Hasil analisis univariat terdapat pada Tabel 3.

https://doi.org/10.25077/TEKNOSI.v10i1.2024.74-81

Tabel 3. Hasil Analisis Univariat.

	Mean	Median	Min	Max	Standar Deviasi
Konsentrasi PM_{10}	55,41	59	13	98	16,35
Konsentrasi SO_2	22,30	20	6	55	12,52
Konsentrasi CO	13,97	13	3	71	7,11
Konsentrasi O_3	75,31	68	16	234	36,68
Konsentrasi NO_2	13,60	13	2	36	5,31
Suhu rata-rata	28,74	28,85	24,9	30,6	0,91
Kelembapan rata-rata	75,01	75	58	94	5,63
Kecepatan angin rata-rata	2,13	2	0	7	0,91

Analisis bivariat dilakukan dengan korelasi. Berdasarkan uji korelasi yang dilakukan, data berkorelasi secara signifikan (p-value <0,05). Korelasi antara variabel independen dan dependen yang paling kuat adalah antara konsentrasi O_3 dengan PM_{10}, sedangkan korelasi terendah ada pada konsentrasi NO_2 dengan konsentrasi PM_{10}. Korelasi positif terjadi antara konsentrasi PM_{10} dengan variabel independen NO_2, CO, O_3, dan suhu rata-rata, artinya semakin tinggi besaran variabel independen, semakin tinggi konsentrasi PM_{10}. Korelasi negatif terjadi antara konsentrasi PM_{10} dengan variabel independen SO_2, kelembapan rata-rata, dan kecepatan angin rata-rata, artinya semakin tinggi besaran variabel dependen, semakin rendah konsentrasi PM_{10}. Besaran korelasi dapat dilihat pada Gambar 4.



Gambar 4. *Heatmap* Korelasi Spearman Antarvariabel.

Sebelum pemodelan, pembagian data dilakukan, dengan jumlah baris data terdapat pada Tabel 3.

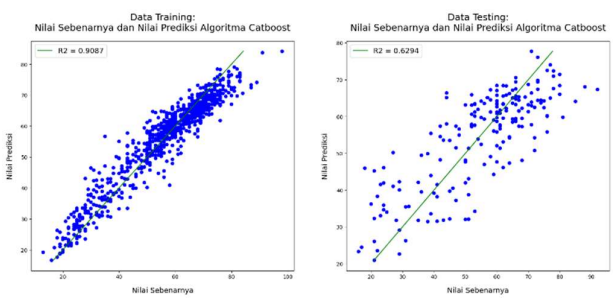
Tabel 4. Hasil Pembagian *Data Training* dan *Data Testing*.

Data	Jumlah Baris	Kolom
Data Training	768	7
Data Testing	192	7

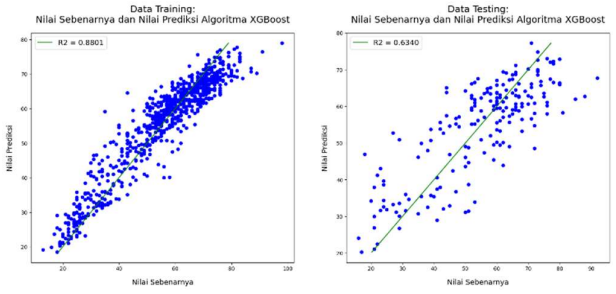
Setelah pembagian data, pencarian *hyperparameter* terbaik untuk setiap algoritma dilakukan dengan GridSearchCV. Tabel 4. menunjukkan nilai *hyperparameter* yang diperoleh tersebut.

Tabel 5. Hasil *Hyperparameter* Terbaik dari GridSearchCV Untuk Setiap Algoritma.

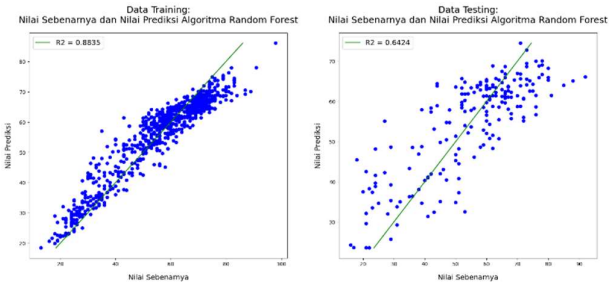
Algoritma	*Hyperparameter*	Nilai
Catboost	depth	8
	learning_rate	0.05
	iterations	300
	min_child_samples	2
	l2_leaf_reg	3
XGBoost	max_depth	6
	learning_rate	0.05
	n_estimators	100
	min_child_weight	5
	reg_lambda	3
Random Forest	max_depth	10
	max_features	'sqrt'
	n_estimators	200
	min_samples_split	2
	min_samples_leaf'	1
	'random_state'	0



Gambar 5. Nilai Sebenarnya dan Nilai Prediksi Algoritma Catboost pada *Data Training* dan *Data Testing*.



Gambar 6. Nilai Sebenarnya dan Nilai Prediksi Algoritma XGBoost pada *Data Training* dan *Data Testing*.



Gambar 7. Nilai Sebenarnya dan Nilai Prediksi Algoritma Random Forest pada *Data Training* dan *Data Testing*.

Tabel 6. Evaluasi Metrik Setiap Algoritma pada *Data Training* dan *Data Testing*.

Data	Metrik	Catboost	XGBoost	Random Forest
Data Training	**R^2**	0,9087	0,8801	0,8835
	MAE	3,7903	4,3647	4,3821
	MSE	24,1702	31,7518	30,8479
	RMSE	4,9163	5,6349	5,5541
	MAPE	0,0785	0,0893	0,0894
Data Testing	**R^2**	0,6294	0,6340	0,6424
	MAE	8,1758	8,0333	7,8751
	MSE	102,0678	100,7869	98,4887
	RMSE	10,1029	10,0393	9,9241
	MAPE	0,1830	0,1754	0,1787

Berdasarkan Tabel 6. dan Gambar 5—7., skor tertinggi untuk *data training* dihasilkan pada algoritma Catboost (R^2 = 0,9087), kemudian diikuti dengan Random Forest (R^2 = 0,8835) dan XGBoost (R^2 = 0,8801). Namun, pada *data testing*, skor tertinggi diperoleh dari Random Forest (R^2 = 0,6424) dibandingkan XGBoost (R^2 = 0,6340) dan Catboost (R^2 = 0,6294). Selain itu, algoritma pada Random Forest memiliki MAE, MSE, dan RMSE yang lebih kecil pada *data testing* dibandingkan algoritma lain, meskipun MAPE-nya lebih tinggi dari Catboost dan XGBoost.

4. PEMBAHASAN

Berdasarkan korelasi yang telah dilakukan, PM_{10} berkorelasi positif dengan suhu rata-rata. Hal ini serupa dengan penelitian [48] pada 2016—2018 terhadap konsentrasi $PM_{2.5}$, yakni semakin tinggi suhu, semakin tinggi konsentrasi $PM_{2.5}$, tetapi berkebalikan dengan penelitian oleh [9]. Sementara itu, faktor meteorologi seperti kelembapan dan kecepatan angin rata-rata berkorelasi negatif dan sejalan dengan penelitian oleh [9].

Sementara itu, saat pemodelan Random Forest, *running code* secara berulang menimbulkan perubahan metrik evaluasi. Kondisi ini berkebalikan dengan Catboost dan XGBoost, yang mana metrik evaluasi tetap konstan walaupun *code* dijalankan berulang. Untuk menangani hal tersebut, *hyperparameter* 'random_state' digunakan pada pemodelan dengan Random Forest agar pemilihan data acak selalu konstan. 'random_state' dapat memengaruhi metrik evaluasi melalui angka yang digunakan, sehingga angka yang berbeda akan menghasilkan metrik evaluasi yang berbeda.

Hasil pemodelan menunjukkan *data training* memiliki hasil yang sangat baik, tetapi kemudian menurun pada *data testing*. Performa *data testing* terbaik dihasilkan pada algoritma Random Forest, sedangkan performa terendah dihasilkan pada algoritma Catboost. Penelitian serupa dengan 4 skenario yang dilakukan oleh [13] menghasilkan R^2 *score* yang lebih baik saat sebuah skenario mempertimbangkan faktor meteorologi, temporal, lalu lintas, dan konsentrasi polutan yang diukur stasiun-stasiun pengukuran. Sementara itu, penelitian oleh [19] juga menghasilkan prediksi yang baik dengan R^2 *score* >0,70 pada model Multiple Linear Regression, Support Vector Regression, Random Forest, dan Gradient Boosting dengan mempertimbangkan faktor temporal, meteorologi, dan polutan lainnya. Oleh karena itu, akurasi pemodelan pada penelitian ini dapat mempertimbangkan faktor temporal. Faktor spasial atau wilayah juga dapat dipertimbangkan untuk mengetahui adanya distribusi polusi [49]. Hal ini mengingat PM_{10} di Jakarta sebagian besar berasal dari lalu lintas dan Jakarta Utara merupakan wilayah yang lahannya digunakan untuk kegiatan industri dan pelabuhan.

5. KESIMPULAN

Algoritma Random Forest menghasilkan skor tertinggi pada *data testing* (R^2 = 0,6424) diikuti dengan XGBoost (R^2 = 0,6340) dan Catboost (R^2 = 0,6294). Selain itu, algoritma pada Random Forest memiliki MAE, MSE, dan RMSE yang lebih kecil pada *data testing* dibandingkan algoritma lain, meskipun MAPE-nya lebih tinggi dari dua algoritma yang diperbandingkan. Meskipun tertinggi pada *data testing*, skor Random Forest merupakan yang kedua tertinggi pada

https://doi.org/10.25077/TEKNOSI.v10i1.2024.74-81

data training (R^2 = 0,8835), sedangkan Catboost tertinggi (R^2 = 0,9087). Hasil R^2 *score* pemodelan yang telah dilakukan menunjukkan adanya penurunan pada *data testing*. Hal tersebut menunjukkan pada penelitian ini algoritma Random Forest memiliki hasil yang lebih baik dalam memprediksi nilai konsentrasi.

Untuk penelitian selanjutnya, hal yang dapat dipertimbangkan saat proses pemodelan karena bisa memengaruhi akurasi di antaranya berupa:

1. Pertimbangan fitur atau variabel independen berupa temporal. Selain itu fitur spasial dapat ditambahkan untuk mengetahui sumber polutan di wilayah tersebut (transportasi dan industri).
2. Pertimbangan pemilihan angka pada *hyperparameter* 'random_state', sebab memengaruhi performa model dan prediksi yang dihasilkan.
3. Adanya penurunan R^2 *score* pada *data testing* dapat ditinjau lebih lanjut.

DAFTAR PUSTAKA

[1] National Institute of Environmental Health Sciences, "Air Pollution and Your Health," National Institute of Environmental Health Sciences. Accessed: Oct. 10, 2023. [Online]. Available: https://www.niehs.nih.gov/health/topics/agents/air-pollution/index.cfm

[2] A. Bozdağ, Y. Dokuz, and Ö. B. Gökçek, "Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey," *Environ. Pollut.*, vol. 263, 2020, doi: 10.1016/j.envpol.2020.114635.

[3] A. Biswal, V. Singh, L. Malik, G. Tiwari, K. Ravindra, and S. Mor, "Spatially resolved hourly traffic emission over megacity Delhi using advanced traffic flow data," *Earth Syst. Sci. Data*, vol. 15, no. 2, pp. 661–680, Feb. 2023, doi: 10.5194/ESSD-15-661-2023.

[4] P. Lestari, M. Khafid Arrohman, S. Damayanti, and Z. Klimont, "Emissions and spatial distribution of air pollutants from anthropogenic sources in Jakarta," 2022, doi: 10.1016/j.apr.2022.101521.

[5] World Health Organization, "Ambient (outdoor) air pollution," World Health Organization. Accessed: Oct. 10, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[6] US EPA, "Particulate Matter (PM) Basics." Accessed: Oct. 23, 2023. [Online]. Available: https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

[7] Y. Liu, Y. Zhou, and J. Lu, "Exploring the relationship between air pollution and meteorological conditions in China under environmental governance," *Sci. Reports 2020 101*, vol. 10, no. 1, pp. 1–11, Sep. 2020, doi: 10.1038/s41598-020-71338-7.

[8] D. Sirithian and P. Thanatrakolsri, "Relationships between Meteorological and Particulate Matter Concentrations (PM2.5 and PM10) during the Haze Period in Urban and Rural Areas, Northern Thailand," *Air, Soil Water Res.*, vol. 15, 2022, doi: 10.1177/11786221221117264.

[9] N. A. Dung *et al.*, "Effect of Meteorological Factors on PM10 Concentration in Hanoi, Vietnam," *J. Geosci. Environ. Prot.*, vol. 7, no. 11, pp. 138–150, Nov. 2019, doi: 10.4236/GEP.2019.711010.

[10] G. Syuhada *et al.*, "Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia," *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, 2023, doi: 10.3390/ijerph20042916.

[11] BPS Provinsi DKI Jakarta, "Jumlah Hari berdasarkan Kategori Indeks Standar Pencemar Udara Menurut Lokasi Pengukuran di Provinsi DKI Jakarta 2018," BPS Provinsi DKI Jakarta. Accessed: Oct. 23, 2023. [Online]. Available: https://jakarta.bps.go.id/indicator/153/378/1/jumlah-hari-berdasarkan-kategori-indeks-standar-pencemar-udara-menurut-lokasi-pengukuran-di-provinsi-dki-jakarta.html

[12] Dinas Lingkungan Hidup DKI Jakarta, "Indeks Standar Pencemaran Udara (ISPU) Tahun 2019," Open Data Jakarta. Accessed: Oct. 04, 2023. [Online]. Available: https://data.jakarta.go.id/dataset/data-indeks-standar-pencemar-udara-ispu-di-provinsi-dki-jakarta-tahun-2019

[13] A. Samad, S. Garuda, U. Vogt, and B. Yang, "Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations," *Atmos. Environ.*, vol. 310, no. July, p. 119987, 2023, doi: 10.1016/j.atmosenv.2023.119987.

[14] W. N. Shaziayani, A. Z. Ul-Saufie, S. Mutalib, N. Mohamad Noor, and N. S. Zainordin, "Classification Prediction of PM10 Concentration Using a Tree-Based Machine Learning Approach," *Atmosphere (Basel).*, vol. 13, no. 4, pp. 1–11, 2022, doi: 10.3390/atmos13040538.

[15] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, pp. 1–45, Dec. 2020, doi: 10.1186/S40537-020-00369-8/FIGURES/9.

[16] S. Wang, Y. Ren, and B. Xia, "PM2.5 and O3 Concentration Estimation Based on Interpretable Machine Learning," *Atmos. Pollut. Res.*, vol. 14, no. 9, p. 101866, 2023, doi: 10.1016/j.apr.2023.101866.

[17] L. Mampitiya *et al.*, "Machine Learning Techniques to Predict the Air Quality Using Meteorological Data in Two Urban Areas in Sri Lanka," *Environ. 2023, Vol. 10, Page 141*, vol. 10, no. 8, p. 141, Aug. 2023, doi: 10.3390/ENVIRONMENTS10080141.

[18] Doreswamy, K. S. Harishkumar, Y. Km, and I. Gad, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.

[19] A. Barthwal, D. Acharya, and D. Lohani, "Prediction and analysis of particulate matter (PM2.5 and PM10) concentrations using machine learning techniques," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 1323–1338, 2023, doi: 10.1007/s12652-021-03051-w.

[20] Z. Deqing, S. Tang, R. Ci, and D. Qiong, "Analysis of the Air Pollution Index and Meteorological Factors and Risk Assessment for Tibet," *J. Phys. Conf. Ser.*, vol. 1838, 2021, doi: 10.1088/1742-6596/1838/1/012047.

[21] H. Yang, Q. Peng, J. Zhou, G. Song, and X. Gong, "The unidirectional causality influence of factors on PM2.5 in Shenyang city of China," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020, doi: 10.1038/s41598-020-65391-5.

[22] A. L. Clements *et al.*, "Source identification of coarse particles in the Desert Southwest, USA using Positive Matrix Factorization," *Atmos. Pollut. Res.*, vol. 8, no. 5, pp. 873–884, Sep. 2017, doi: 10.1016/J.APR.2017.02.003.

[23] T. Handhayani, "An integrated analysis of air pollution and meteorological conditions in Jakarta," *Sci. Rep.*, vol. 13,

no. 1, pp. 1–11, 2023, doi: 10.1038/s41598-023-32817-9.
[24] Y. Wu *et al.*, "Comparison of dry and wet deposition of particulate matter in near-surface waters during summer," *PLoS One*, vol. 13, no. 6, pp. 1–15, 2018, doi: 10.1371/journal.pone.0199241.
[25] Z. Husnina, K. Wangdi, T. Puspita, S. M. Praveena, and Z. Ni, "Profiling Temporal Pattern of Particulate Matter (PM10) and Meteorological Parameters in Jakarta Province during 2020-2021," *J. Kesehat. Lingkung.*, vol. 15, no. 1, pp. 16–26, 2023, doi: 10.20473/jkl.v15i1.2023.16-26.
[26] S. Kirešová and M. Guzan, "Determining the Correlation between Particulate Matter PM10 and Meteorological Factors," *Eng*, vol. 3, no. 3, pp. 343–363, 2022, doi: 10.3390/eng3030025.
[27] J. Kujawska, M. Kulisz, P. Oleszczuk, and W. Cel, "Machine Learning Methods to Forecast the Concentration of PM10 in Lublin, Poland," *Energies*, vol. 15, no. 17, pp. 1–23, 2022, doi: 10.3390/en15176428.
[28] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. New York: Springer Publishing Company, 2012. doi: 10.1007/978-1-4419-9236-7.
[29] Y. A. Ali, E. M. Awwad, M. Al-Razgan, and A. Maarouf, "Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity," *Process. 2023, Vol. 11, Page 349*, vol. 11, no. 2, p. 349, Jan. 2023, doi: 10.3390/PR11020349.
[30] D. M. Belete and M. D. Huchaiah, "Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results," *Int. J. Comput. Appl.*, vol. 44, no. 9, pp. 875–886, Sep. 2022, doi: 10.1080/1206212X.2021.1974663.
[31] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis," *Informatics 2021, Vol. 8, Page 79*, vol. 8, no. 4, p. 79, Nov. 2021, doi: 10.3390/INFORMATICS8040079.
[32] Z. Guo, X. Wang, and L. Ge, "Classification prediction model of indoor PM2.5 concentration using CatBoost algorithm," *Front. Built Environ.*, vol. 9, p. 1207193, Jul. 2023, doi: 10.3389/FBUIL.2023.1207193/BIBTEX.
[33] I. Ayus, N. Natarajan, and D. Gupta, "Comparison of machine learning and deep learning techniques for the prediction of air pollution: a case study from China," *Asian J. Atmos. Environ.*, vol. 17, no. 1, pp. 1–22, Dec. 2023, doi: 10.1007/S44273-023-00005-W/FIGURES/14.
[34] M. Méndez, M. G. Merayo, and M. Núñez, "Machine learning algorithms to forecast air quality: a survey," *Artif. Intell. Rev. 2023 569*, vol. 56, no. 9, pp. 10031–10066, Feb. 2023, doi: 10.1007/S10462-023-10424-4.
[35] T. Plocoste and S. Laventure, "Forecasting PM10 Concentrations in the Caribbean Area Using Machine Learning Models," *Atmosphere (Basel).*, vol. 14, no. 1, pp. 1–13, 2023, doi: 10.3390/atmos14010134.
[36] A. Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," *Interdiscip. J. Information, Knowledge, Manag.*, vol. 14, pp. 45–76, Sep. 2018, doi: 10.28945/4184.
[37] J. Kaliappan, K. Srinivasan, S. Mian Qaisar, K. Sundararajan, C. Y. Chang, and C. Suganthan, "Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate," *Front. Public Heal.*, vol. 9, no. September, pp. 1–12, 2021, doi: 10.3389/fpubh.2021.729795.
[38] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.
[39] I. N. Achmad and R. S. Witiastuti, "Underpricing, Institutional Ownership and Liquidity Stock of IPO Companies in Indonesia," *Manag. Anal. J.*, vol. 7, no. 3, pp. 281–291, 2018, [Online]. Available: http://maj.unnes.ac.id
[40] Badan Meteorologi Klimatologi dan Geofisika, "Pusat Database BMKG," Badan Meteorologi Klimatologi dan Geofisika. Accessed: Oct. 04, 2023. [Online]. Available: https://dataonline.bmkg.go.id/home
[41] Dinas Lingkungan Hidup DKI Jakarta, "Indeks Standar Pencemaran Udara (ISPU) Tahun 2017," Open Data Jakarta. Accessed: Oct. 04, 2023. [Online]. Available: https://data.jakarta.go.id/dataset/indeks-standar-pencemaran-udara-ispu-tahun-2017
[42] Dinas Lingkungan Hidup DKI Jakarta, "Indeks Standar Pencemaran Udara (ISPU) Tahun 2018," Open Data Jakarta. Accessed: Oct. 04, 2023. [Online]. Available: https://data.jakarta.go.id/dataset/indeks-standar-pencemar-udara-di-provinsi-dki-jakarta-tahun-2018
[43] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Ann. Card. Anaesth.*, vol. 22, no. 1, pp. 67–72, 2019, doi: 10.4103/aca.ACA_157_18.
[44] J. W. Heo *et al.*, "Smoking is associated with pneumonia development in lung cancer patients," *BMC Pulm. Med.*, vol. 20, no. 1, pp. 1–8, May 2020, doi: 10.1186/S12890-020-1160-8/TABLES/3.
[45] C. Xiao, J. Ye, R. M. Esteves, and C. Rong, "Using Spearman's correlation coefficients for exploratory data analysis on big dataset," *Concurr. Comput. Pract. Exp.*, vol. 28, no. 14, pp. 3866–3878, Sep. 2016, doi: 10.1002/CPE.3745.
[46] M. Lobo and R. D. Guntur, "Spearman's rank correlation analysis on public perception toward health partnership projects between Indonesia and Australia in East Nusa Tenggara Province," *J. Phys. Conf. Ser.*, vol. 1116, no. 2, 2018, doi: 10.1088/1742-6596/1116/2/022020.
[47] V. R. Joseph, "Optimal Ratio for Data Splitting," *Stat. Anal. Data Min.*, vol. 15, no. 4, pp. 531–538, 2022, doi: 10.1002/sam.11583.
[48] W. L. Kusuma, W. Chih-Da, Z. Yu-Ting, H. H. Hapsari, and J. L. Muhamad, "PM2.5 Pollutant in Asia—A Comparison of Metropolis Cities in Indonesia and Taiwan," *Int. J. Environ. Res. Public Health*, vol. 16, no. 24, pp. 1–12, 2019, doi: 10.3390/ijerph16244924.
[49] K. I. Solihah, D. N. Martono, and B. Haryanto, "Analysis of Spatial Distribution of PM2.5and Human Behavior on Air Pollution in Jakarta," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 940, no. 1, 2021, doi: 10.1088/1755-1315/940/1/012018.

NOMENKLATUR

ρ	Nilai Korelasi Spearman
d	Margin setiap nilai pasangan
n	Nilai pasangan pada peringkat Spearman

BIODATA PENULIS

Elita Rizkiani Putri
Elita Rizkiani Putri merupakan mahasiswa tingkat akhir program studi Kesehatan Lingkungan, Fakultas Kesehatan Masyarakat, Universitas Indonesia.

Dede Brahma Arianto
Dede Brahma Arianto merupakan praktisi di bidang data yang telah menempuh pendidikan Magister Informatika, Fakultas Teknik Industri, Universitas Islam Indonesia.