LAPORAN FINAL PROJECT

PEMBELAJARAN MESIN H 2024

"Analisis Time Series dan Forecasting Tingkat Konsentrasi PM2.5 terhadap Data Pemantauan Kualitas Udara di Athena, Yunani"

Disusun oleh:

1. Fadhl Akmal Madany 5025221028
2. Keanu Fortuna Taufan 5025221040
3. Zelvan Abdi Wijaya 5025221125

DEPARTEMEN TEKNIK INFORMATIKA

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

DAFTAR ISI

1. PENDAHULUAN

1.1 Latar Belakang

Polusi udara merupakan masalah kesehatan lingkungan besar yang mempengaruhi masyarakat secara global. The Global Burden of Disease mengestimasi bahwa segala jenis polusi udara bertanggung jawab terhadap kematian 12.2% pria dan 11.3% wanita secara global pada tahun 2019 [1]. Kasus dari Ella Kissi-Debrah, orang pertama di dunia yang secara legal dinyatakan wafat karena polusi udara, adalah contoh nyata dari bagaimana polusi udara mempengaruhi kehidupan manusia [2].

Tiga minggu sebelum ulang tahunnya yang ketujuh, Ella menderita infeksi pada paru-paru yang seiring dengan berjalannya waktu berkembang menjadi batuk yang tak kunjung reda. Dokter mendiagnosis Ella dengan penyakit asma. Tak lama setelah diagnosis tersebut, Ella batuk terus menerus hingga pingsan, kondisi yang disebabkan karena kurangnya oksigen masuk ke otak. Tetangganya berhasil melakukan pertolongan pertama pada Ella saat kejadian itu. Ella dilarikan ke rumah sakit dan dipulangkan keesokan harinya. Mengira semuanya baik-baik saja, gejala Ella kambuh seminggu kemudian, kali ini membuatnya koma selama tiga hari. Penyebab dari asma yang diderita Ella saat itu masih belum diketahui. Dalam jangka waktu dua tahun, Rosamund membawa putrinya ke berbagai pemeriksaan, mulai dari pola tidur dan uji bakteri hingga pemeriksaan epilepsi dan fibrosis sistik. Pemeriksaan-pemeriksaan tersebut hanya membawa satu kesimpulan: bahwa Ella sensitif terhadap alergen.

Ella menjalani perawatan intensif dengan dokter paru-paru pada tahun 2012, hingga gejala batuk yang dideritanya kambuh, membuatnya harus keluar masuk rumah sakit secara rutin, terhitung lebih dari tiga puluh kali sejak diagnosis asma yang ia terima. Pada bulan Februari 2013, tiga minggu setelah ulang tahunnya yang kesembilan, Ella mengalami serangan asma yang fatal, membuatnya menghembuskan nafas terakhir. Ella didiagnosis mengalami gangguan pernapasan akut sebagai penyebab kematiannya. Pihak medis tidak pernah menyebutkan polusi udara sebagai penyebab potensial asma yang dideritanya. Mengikuti saran dari seorang penduduk lokal, Rosamund, ibu kandung Ella, menyelidiki kadar polutan udara pada malam putrinya wafat. Rosamund menemukan bahwa tak jauh dari tempat mereka tinggal, terdapat jalan raya padat kendaraan dengan kadar nitrogen dioksida di atas batas legal [3].

Dokter spesialis pernapasan dari University of Southampton menyelidiki rekam medis Ella dan menemukan kesimpulan bahwa terdapat keparahan dari asma yang diderita Ella terkait dengan tingkat polusi udara di tempat mereka tinggal [4]. Pada tahun 2019, kasus Ella sampai menuju Pengadilan Tinggi. Setelah kajian mengenai langkah-langkah yang dilakukan otoritas setempat, pemerintah, hingga Walikota London pada saat itu mengenai kontrol polusi udara, didapat kejelasan bahwa Ella menghabiskan masa hidupnya di daerah dengan kadar emisi nitrogen dioksida pada udara di atas batas yang ditentukan secara nasional dan Uni Eropa, serta kadar PM di atas pedoman WHO. Pada akhir tahun 2020, pengadilan secara legal memutuskan bahwa polusi udara menjadi salah satu penyebab kematian Ella [5].

Polusi seringkali dikaitkan pada perubahan iklim (i.e. permasalahan di masa depan). Opini publik mengenai perubahan iklim sendiri beragam, di mana tidak sedikit masyarakat yang tidak mengakui adanya perubahan iklim, atau perubahan iklim yang disebabkan oleh aktivitas manusia. Meskipun begitu, polusi udara berkontribusi terhadap berbagai masalah pernapasan seperti asma, penyakit paru obstruktif kronis, kanker paru-paru, hingga stroke. Meskipun hubungan antara polusi udara dan penyakit pernapasan tidaklah sederhana, studi epidemiologi baru-baru ini menemukan asosiasi tingkat polusi udara yang disebabkan oleh lalu lintas terhadap peningkatan kasus asma dan penyakit pernapasan lainnya [6].

Kasus Ella, dan masih banyak lainnya, sudah selayaknya menjadi pengingat bagi umat manusia bahwa, terlepas dari segala "kontroversi" mengenai aktivisme polusi udara dan perubahan iklim, aktivitas manusia yang mempengaruhi tingkat kualitas udara memiliki dampak langsung terhadap keberlangsungan umat manusia. Laporan State of Global Air memperkirakan bahwa polusi udara berkontribusi terhadap sekitar 6.67 juta kematian secara global pada tahun 2019. Polusi udara juga diperkirakan menurunan rerata ekspektansi hidup manusia secara global hingga 1 tahun 8 bulan, kematian sekitar setengah juta bayi secara global setiap tahunnya, dan meningkatkan risiko kesehatan bagi masyarakat rentan seperti populasi berumur [7].

Bumi, planet yang diteorikan terbentuk dari ledakan kosmos, tidak perlu "diselamatkan"; ia telah selamat dari hujan meteor hingga zaman es, dan akan terus selamat tanpa umat manusia. Manusia, di sisi lain, perlu diselamatkan. Maka cukuplah alasan sederhana ini menjadi sebab bagi umat manusia untuk lebih memperhatikan lingkungan, sedikit demi sedikit.

Langkah pertama dalam menangani polusi udara adalah memperoleh pengetahuan mengenai kualitas udara di lingkungan sekitar. Air Quality Monitoring Station (AQMS) adalah instrumen yang dibuat untuk melaksanakan tugas ini. Dengan ketersediaan data AQMS yang dapat diakses secara publik, penulis berinisiatif untuk melakukan analisis dari data yang tersedia. Penulis melakukan analisis deret waktu pada data kualitas udara yang dikumpulkan oleh Air Quality Monitoring Station (AQMS) di Kota Athena, Yunani. Melalui analisis ini, diharapkan terdapat insight mengenai bagaimana masing-masing metrik kualitas udara berkaitan satu sama lain, dan bagaimana tren kualitas udara dalam selang waktu pengamatan.

Dengan berkembangnya riset pada agoritma pembelajaran, prediksi lebih lanjut dapat dilakukan untuk menemukan suatu model yang dapat memprediksi kadar polutan PM2.5 dalam suatu waktu berdasarkan konsentrasi gas berbahaya yang dideteksi. Particulate Matter (PM2.5) adalah partikel udara yang berukuran lebih kecil dari atau sama dengan 2.5 µm. Secara tradisional, konsentrasi PM2.5 diukur menggunakan metode penyinaran sinar Beta (Beta Attenuation Monitoring) dengan satuan mikrogram per meter kubik (µm/m3). Dari semua polutan dalam udara, PM2.5 polusi menjadi ancaman kesehatan terbesar. Karena ukurannya yang kecil, PM2.5 dapat bertahan di udara untuk waktu yang lama dan dapat masuk ke dalam tubuh hingga ke dalam aliran darah saat dihirup [6].

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, terdapat beberapa rumusan masalah dalam laporan ini:

1. Bagaimana kualitas udara di Kota Athena berdasarkan data yang dikumpulkan oleh Air Quality Monitoring Station (AQMS)?
2. Bagaimana pola dan tren kualitas udara, khususnya konsentrasi PM2.5, di Kota Athena, Yunani berdasarkan data Air Quality Monitoring Station (AQMS)?
3. Bagaimana performa algoritma regresi seperti Linear Regression, Random Forest, *eXtreme Gradient Boosting* (XGBoost), dan *Long-Short Term Memory* (LSTM) dalam memprediksi kadar polutan PM2.5 di Kota Athena, Yunani?
4. Bagaimana pengaruh *hyperparameter tuning*, inklusi data historis, dan metode *bagging* meningkatkan performa model dalam memprediksi konsentrasi PM2.5 di Kota Athena, Yunani?

1.3 Batasan Masalah

Beberapa batasan masalah perlu ditetapkan agar penelitian lebih terfokus dan tidak memperluas pokok bahasan, antara lain:

1. Penelitian ini dilakukan menggunakan data deret waktu kualitas udara yang dikumpulkan oleh Air Quality Monitoring Station (AQMS) di Kota Athena, Yunani.
2. Analisis time series dan forecasting akan dilakukan berdasarkan data historis yang tersedia dalam rentang waktu tertentu dengan fokus pada PM 2.5.
3. Penelitian ini hanya akan membandingkan empat algoritma utama: Linear Regression, Random Forest, *eXtreme Gradient Boosting* (XGBoost), dan *Long-Short Term Memory* (LSTM). Algoritma lain yang mungkin relevan tidak akan disertakan dalam analisis ini.
4. Performa model prediktif akan dievaluasi berdasarkan metrik *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), koefisien determinasi (R^2) dan *Symmetric Mean Absolute Percentage Error* (SMAPE).
5. Evaluasi akan dibatasi pada kemampuan model dalam memprediksi konsentrasi PM 2.5 di masa depan.

1.4 Tujuan

Tujuan dari penelitian ini adalah:

1. Mengidentifikasi dan menganalisis tren historis serta pola musiman dari konsentrasi PM 2.5 di Athena berdasarkan analisis data deret waktu AQMS..
2. Mengevaluasi kinerja model prediksi konsentrasi PM2.5 menggunakan algoritma *Linear Regression*, *Random Forest*, *eXtreme Gradient Boosting* (XGBoost), dan *Long-Short Term Memory* (LSTM) dalam memprediksi konsentrasi PM2.5 di Kota Athena, Yunani.

3. Menentukan model yang memberikan performa terbaik berdasarkan metrik evaluasi MAE, RMSE, R^2, dan SMAPE dalam memprediksi konsentrasi PM 2.5 di Kota Athena, Yunani.
4. Membandingkan pengaruh optimasi hyperparameter, inklusi data historis, dan *bagging* terhadap kinerja model prediksi konsentrasi PM2.5 di Kota Athena, Yunani.

1.5 Manfaat

Adapun beberapa manfaat yang dapat diperoleh dari penelitian ini:

1. Memberikan informasi mengenai tren dan hubungan antar metrik kualitas udara di Kota Athena, Yunani, yang dapat membantu dalam memahami kondisi kualitas udara di kota tersebut.
2. Memberikan informasi tentang kinerja berbagai model prediksi konsentrasi PM2.5, yang dapat membantu dalam memilih model yang tepat untuk memprediksi kualitas udara.
3. Memberikan informasi tentang pengaruh optimasi model terhadap kinerja prediksi, yang dapat membantu dalam meningkatkan performa prediksi konsentrasi PM2.5.
4. Menyediakan data dan analisis yang dapat digunakan oleh pembuat kebijakan dan otoritas terkait untuk merancang strategi mitigasi yang lebih baik dan tepat sasaran dalam mengurangi polusi udara di Athena.
5. Menambah literatur ilmiah dalam bidang analisis data deret waktu dan *forecasting* kualitas udara, serta penggunaan algoritma *Linear Regression*, *Random Forest*, *eXtreme Gradient Boosting* (XGBoost), dan *Long-Short Term Memory* (LSTM).

2. Studi Literatur

2.1 Sampling

Pengambilan sampel (sampling) adalah teknik untuk memilih subset dalam suatu populasi yang menjadi fokus dalam suatu penelitian. Dalam banyak kasus, penggunaan seluruh data populasi sulit bahkan mustahil. Sampling dari suatu populasi seringkali menjadi solusi yang praktis dan memungkinkan data untuk didapat dengan lebih cepat dan hemat biaya. Sampel digunakan untuk menarik kesimpulan karakteristik populasi. Teknik sampel yang baik dapat diukur dari seberapa beragam sampel tersebut dan seberapa baik sampel tersebut mewakili keseluruhan populasi [8].

2.2 Linear Regression

Linear regression adalah teknik regresi sederhana dan paling sering ditemukan dalam pembelajaran mesin. Dalam banyak kasus, linear regression menghasilkan kinerja yang setara dengan metode-metode mutakhir dengan keuntungan berupa model yang mudah diinterpretasikan [9]. Jika diberikan fitur $\hat{X}$ dan target $\hat{Y}$ dalam suatu observasi $(X_1^T, Y_1)^T, \ldots, (X_n^T, Y_n)^T \in \mathbb{R}^d \times \mathbb{R}$, tujuan dari linear regression adalah memodelkan suatu estimasi melalui persamaan linear:

$$\hat{Y} = \hat{X}^T \hat{\beta} + \hat{\varepsilon} \tag{2.1}$$

Dengan $\hat{\beta} \in \mathbb{R}^d$ adalah suatu estimator [10]. Estimator $\hat{\beta}$ dipilih sedemikian sehingga hasil estimasi tidak terpaut jauh dengan observasi (i.e. meminimalisasi error) mengikuti persamaan:

$$\hat{\beta} \coloneqq \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \theta)^2 \tag{2.2}$$

Di mana arg min merepresentasikan $\theta \in \mathbb{R}^d$ ketika minimum tercapai [10]. Pada metode ordinary least squares (metode yang dipakai dalam eksperimen ini), estimator $\hat{\beta}$ diperoleh melalui:

$$\hat{\beta} = \left(\hat{X}^T \hat{X}\right)^{-1} \hat{X}^T \hat{Y} \tag{2.3}$$

Metode ordinary least square menghasikan linear unbiased estimator optimal dalam banyak kasus untuk jumlah sampel $n \geq d$ [10]. Estimator $\hat{\beta}$ juga dapat diperoleh melalui metode-metode lainnya, seperti least square method [9]. Sementara linear regression dipilih sebagai baseline perbandingan terhadap model-model lain yang lebih kompleks, penggunaan algoritma berbasis linear regression tidak jarang ditemui pada penelitian time series forecasting [11], [12], [13].

2.3 Random Forest

Random forest secara sederhana adalah algoritma ensemble learning yang bekerja dengan membangun banyak decision tree. Hasil inference dari kumpulan tree tersebut menjadi basis dari prediksi random forest. Pada kasus klasifikasi, pendekatan yang umum adalah dengan melakukan majority vote pada tiap tree dalam random forest. Dalam kasus regresi, digunakan variasi dari decision tree yang dapat memprediksi variabel target kontinu, seperti regression tree [14].

Regression tree sendiri adalah algoritma pembelajaran berbasis decision tree yang digunakan dalam analisis regresi. Regression tree dapat digunakan untuk menganalisis hubungan yang kompleks antara fitur input dan variabel target yang kontinu. Model dibangun sedemikian sehingga node internal menyatakan kriteria splitting decision tree dan leaf node atau output yang dihasilkan kontinu, sehingga model dapat digunakan untuk memprediksi target yang kontinu berdasarkan fitur input. Regression tree menghasilkan performa yang baik ketika digunakan pada data di mana variabel fitur dan target memiliki interaksi atau hubungan nonlinear. Selain itu, model regression tree juga cenderung mudah untuk diinterpretasikan dan divisualisasikan [15].

Pada kasus regresi, hasil prediksi random forest secara umum didapatkan dari rata-rata prediksi seluruh regression tree yang menjadi bagian dari random forest. Algoritma random forest dapat menangani data yang sangat besar dan berdimensi tinggi dengan lebih baik dan meningkatkan akurasi prediksi tanpa meningkatkan beban komputasi secara signifikan. Selain itu, algoritma ini tidak sensitif terhadap kolinearitas multivariat, dan hasil perhitungan pada data dengan data yang hilang atau tidak seimbang [16]. Random forest juga sudah digunakan pada berbagai penelitian yang melibatkan forecasting pada data time series [17], [18], [19].

2.4 eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting (XGBoost) adalah algoritma ensemble learning berbasis decision tree dengan optimisasi yang memanfaatkan teknik gradient descent dalam pembangunannya (i.e. ketimbang random) [20]. Pendekatan gradient descent dilakukan untuk meminimalisasi loss dengan tambahan regularization parameters untuk mencegah overfitting. Tujuan utama dari XGBoost adalah meminimalisasi fungsi objektif yang didefinisikan seperti berikut:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

Dengan l adalah suatu loss function konveks yang dapat diturunkan untuk merepresentasikan error antara observasi y_i dengan data prediksi $\hat{y}_i$ dan f_t adalah model tree pada iterasi ke-t pada proses optimasasi. $\Omega(f)$ adalah regularization term yang diatur untuk mencegah overfitting. Perhatikan bahwa nilai $\Omega(f_t)$ yang besar dapat memberikan penalti pada fungsi objektif. Pada algoritma XGBoost, regularization term didefinisikan seperti berikut:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \quad (5)$$

Dengan T adalah total leaf nodes pada pohon, γ dan λ adalah koefisien untuk $L1$ dan $L2$ regularization term, dan w adalah suatu vektor skor pada setiap leaf node [20]. Tujuan penambahan $\Omega(f)$ pada fungsi objektif adalah memberikan penalti pada model yang terlalu kompleks dan memberikan weight akhir yang lebih mulus sehingga model akhir cenderung sederhana dan mencegah overfitting. Sementara f_t ditemukan selama proses pembelajaran, parameter T, γ, dan λ perlu ditentukan sebelum training. Untuk menetukan koefisien yang optimal, dapat dilakukan hyperparameter tuning [20].

Algoritma pembelajaran XGBoost dipilih karena performanya yang sangat baik pada dataset berukuran besar dan dapat memberikan prediksi yang akurat dengan biaya komputasi yang lebih rendah ketimbang neural network [20]. Selain itu, penggunaan XGBoost dalam kasus serupa dengan penelitian (time series forecasting) juga banyak ditemukan pada penelitian-penelitian sebelumnya dan menghasilkan prediksi yang baik [20], [21], [22], [23].

2.5 Long Short Term Memory (LSTM)

LSTM merupakan variant dari unit Recurrent Neural Network (RNN). Pada arsitektur LSTM, terdiri atas neural network dan beberapa blok memori yang berbeda yang disebut dengan *cell*. Informasi dikumpulkan lalu disimpan oleh *cell* dan dilakukan manipulasi memori yang dilakukan oleh komponen. Komponen tersebut disebut dengan *gate.* Pada LSTM, terdapat 3 gate antara lain *input gate, forget gate* dan *output gate.* Pada kasus ini, LSTM *neural network* sangat cocok untuk mengklasifikasi, memproses, dan membuat prediksi berdasarkan data *time series* karena ada kelangkaan durasi yang tidak diketahui di antara peristiwa penting dalam rangkaian waktu [24].

2.6 Evaluation Metrics

Ketika memilih suatu model pembelajaran mesin, muncul kebutuhan untuk mengetahui suatu model yang memiliki performa terbaik di antara beberapa model dalam pengujian. Evaluasi model diperlukan agar peneliti dapat berfokus untuk melakukan optimalisasi pada model dengan kinerja terbaik. Oleh karena itu, muncul kebutuhan untuk melakukan kuantifikasi kinerja model dalam suatu metrik agar dapat dibandingkan. Sebagai contoh, metrik evaluasi dalam model medis dapat dijadikan indikator performa suatu model jika dibandingkan dengan tenaga medis profesional maupun model yang sudah ada sebelumnya [25].

2.6.1 Mean Absolute Error (MAE)

Mean absolute error (MAE) didefinisikan sebagai rerata dari L1 norm (manhattan distance) dari suatu observasi y terhadap prediksi $\hat{y}$ [26], dengan kata lain:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

Salah satu kelebihan mean absolute error adalah unit yang dihasilkan sama dengan unit pada variabel target, sehingga nilai dari suatu error mudah untuk diinterpretasikan [26].

2.6.2 Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) dapat didefinisikan sebagai rerata dari L2 norm (euclidian distance) dari suatu observasi y terhadap prediksi $\hat{y}$ [26], dengan kata lain:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Seperti namanya, RMSE adalah akar kuadrat dari mean squared error (MSE). MSE diperoleh dari rata-rata error yang dikuadratkan. Ini mengakibatkan fungsi memberikan penalti yang sangat besar untuk error yang besar dan penalti yang kecil untuk error yang kecil. Tentu saja, unit yang dihasilkan berbeda dengan unit variabel target, sehingga RMSE melakukan operasi akar kuadrat untuk mengembalikan unit seperti semula. Dengan demikian, nilai yang dihasilkan oleh RMSE juga dapat diinterpretasikan dengan mudah [26].

2.6.3 Coefficient of Determination ($\mathbf{R^2}$)

Coefficient of determination (juga disebut R^2) adalah suatu metrik dalam statistik yang menjelaskan kecocokan suatu prediksi model terhadap observasi. Misalkan terdapat suatu observasi y dan prediksi $\hat{y}$, nilai R^2 didefinisikan sebagai:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{8}$$

Di mana $\bar{y}$ adalah rata-rata dari observasi. Coefficient of determination berada pada rentang $(-\infty, 1]$ dan sering digunakan untuk mengukur kualitas regresi, di mana nilai R^2 yang lebih besar mengacu pada model regresi yang lebih baik. Nilai ini juga dapat diartikan sebagai proporsi varians dalam variabel dependen yang dapat diprediksi dari variabel independen [27].

2.6.4 Symmetric Mean Absolute Percentage Error (SMAPE)

Symmetric mean absolute percentage error (SMAPE) adalah salah satu metrik evaluasi yang menghasilkan persentase ketimbang suatu nilai dalam unit tertentu. Metrik yang bergantung pada persentase seperti SMAPE mengukur error dalam persentase dan memberikan pemikiran yang dapat ditafsirkan mengenai kualitas prediksi. Sebagai contoh, kesalahan pengukuran sebesar 5% dapat diinterpretasikan dengan lebih mudah ketimbang kesalahan pengukuran sebesar 50 cm [28]. Jika diberikan observasi y dan prediksi $\hat{y}$, SMAPE didefinisikan seperti berikut:

$$\mathrm{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{(\hat{y}_i + y_i) \div 2}\right| \tag{9}$$

SMAPE digunakan untuk mengatasi masalah utama pendahulunya, MAPE, yaitu sifatnya yang scale-insensitive. MAPE cenderung memberikan nilai yang ekstrem ketika nilai observasi sangat kecil. Selain itu, MAPE cenderung memberikan penalti atas error positif jauh lebih besar ketimbang negatif. SMAPE memperbaiki hal ini dengan memberikan penalti yang sama pada error positif dan negatif. SMAPE (dan juga MAPE) memberikan nilai pada rentang $[0, +\infty)$, di mana nilai 0 mendeskripsikan kondisi optimal [28].

2.7 Tree-Structured Parzen Estimator (TPE)

Tree-structured Parzen estimator (TPE) adalah varian dari metode optimasi Bayesian yang digunakan dalam melakukan hyperparameter tuning. Disebut tree-structured karena metode optimasi ini dapat digunakan pada search-space model dengan struktur pohon seperti XGBoost. Tujuan dari optimasi Bayesian adalah untuk meminimalisasi fungsi objektif $f(x)$ sedemikian sehingga:

$$x_{opt} \in \arg\min_{x \in X} f(x) \tag{10}$$

Sebagai contoh, hyperparameter tuning bertujuan untuk menemukan suatu parameter (e.g. learning rate, jumlah tree, dan koefisien regularisasi) yang menghasilkan performa terbaik (e.g. error minimal). Optimizer Bayesian secara iteratif mencari x_{opt} menggunakan fungsi akuisisi untuk menimbang trade-off antara eksploitasi dan eksplorasi. Secara sederhana, eksploitasi adalah upaya pencarian observasi terbaik dan eksplorasi adalah pencarian search-space yang belum disentuh [29]. Fungsi akusisi yang umum digunakan adalah menggunakan expected improvement (EI) yang didefisikan sebagai:

$$\mathrm{EI}_{y^*}[x|D] \coloneqq \int_{-\infty}^{y^*} (y^* - y) p(y|x, D)\, \mathrm{d}y \tag{11}$$

Pilihan lain dalam fungsi akusisi adalah probability of improvement (PI) yang didefinisikan seperti berikut:

$$\mathrm{P}(y \leq y^*|x, D) \coloneqq \int_{-\infty}^{y^*} p(y|x, D)\, \mathrm{d}y \tag{12}$$

Di mana y^* adalah suatu parameter kontrol yang diberikan oleh pengguna. PI cenderung mengeksploitasi pengetahuan sedangkan EI cenderung mengeksplorasi search region baru. TPE menggunakan fungsi akusisi PI sehingga TPE cenderung untuk melakukan pencarian secara lokal. Untuk menghitung fungsi akusisi, diperlukan suatu model yang menyatakan nilai $p(y|x, D)$ [29]. Pada TPE, digunakan kernel density estimators (KDEs) dengan pemodelan $p(y|x, D)$ mengikuti asumsi berikut:

$$p(x, y|D) \coloneqq \begin{cases} p\left(x \middle| D^{(l)}\right) & (y \leq y^{\gamma}) \\ p\left(x \middle| D^{(g)}\right) & (y > y^{\gamma}) \end{cases} \tag{13}$$

Di mana top-quantile γ dihitung pada tiap iterasi berdasarkan jumlah observasi $N = |D|$ [29].

3. METODOLOGI

3.1 Dataset

Pada penelitian ini, digunakan dataset *Regional Datasets for Air Quality Monitoring in European Cities* yang dipublikasikan dalam konferensi 2024 IEEE IGARS-24 atau IEEE International Geoscience and Remote Sensing pada tanggal 7-12 Juli yang bertempat di Athena, Yunani [30]. Dataset juga tersedia melalui platform Kaggle melalui Vladimir Demidov. Dataset membahas kualitas udara dari beberapa tempat di kawasan Eropa diantaranya Ancona Italia, Zaragoza Spanyol, dan Athena Yunani yang mana dataset pada kawasan ini akan menjadi fokus utama dalam penelitian ini. Dataset diambil dari AQMS atau *Air Quality Monitoring System* di Kota Athena, Yunani. Berikut adalah informasi kolom yang tersedia pada dataset menurut data card:

- **Date:** Tanggal dan waktu pengukuran yang dicatat.
- **Latitude:** Koordinat lintang dari lokasi pengukuran.
- **Longitude:** Koordinat bujur dari lokasi pengukuran.
- **station_name:** Nama stasiun tempat pengukuran dilakukan.
- **Wind-Speed (U):** Komponen U dari kecepatan angin, mewakili arah timur-barat.
- **Wind-Speed (V):** Komponen V dari kecepatan angin, mewakili arah utara-selatan.
- **Dewpoint Temp:** Suhu titik embun, menunjukkan suhu di mana udara menjadi jenuh dengan uap air.
- **Soil Temp:** Suhu tanah di lokasi pengukuran.
- **Total Percipitation:** Total jumlah presipitasi yang tercatat.
- **Vegitation (High):** Jumlah vegetasi tinggi yang ada di lokasi pengukuran.
- **Vegitation (Low):** Jumlah vegetasi rendah yang ada di lokasi pengukuran.
- **Temp:** Suhu lingkungan saat pengukuran dilakukan.
- **Relative Humidity:** Persentase kelembaban relatif pada saat pengukuran.
- **PM10:** Konsentrasi partikel materi dengan diameter 10 mikrometer atau kurang.
- **PM2.5:** Konsentrasi partikel materi dengan diameter 2,5 mikrometer atau kurang.
- **NO2:** Konsentrasi nitrogen dioksida.
- **O3:** Konsentrasi ozon.
- **code:** Kode yang mewakili detail spesifik terkait data (tergantung konteks).
- **id:** Pengidentifikasi unik untuk setiap catatan dalam dataset.

Pada pengelolaannya, data time series menjadi alasan mengapa perlu dijadikan kolom waktu terlebih dahulu sebagai indeks. Sehingga untuk mengimplementasikannya, kolom date perlu di-parse dan dijadikan indeks. Dataset pada awalnya berisi jutaan baris entri dalam rentang 3 tahun yaitu sekitar 2020-2023 dengan resolusi temporal per jamnya dan resolusi spasial dengan derajatnya yaitu 0.005. adalah hingga kemudian suatu sumber daya komputasi yang terbatas menjadikannya dataset ini tereduksi menjadi 26 ribu baris dengan melakukan data sampling dengan mengambil data sensor tiap jamnya.

Dalam dataset ini, terdapat beberapa fitur penting yang disediakan antaranya yaitu tanggal, lokasi pengukuran, tempat pengukuran dilakukan, kecepatan angin, suhu, jumlah vegetasi, kelembaban, konsentrasi partikel, konsentrasi nitrogen, dan konsentrasi ozon. Fitur-fitur tersebut kebanyakan memiliki korelasi yang memiliki peranan penting dalam penelitian ini khususnya terhadap target fitur yaitu PM2.5. Ukuran PM2.5 yang kecil dapat menyebabkan masalah kesehatan yang serius. Selain itu, PM2.5 dapat bertahan di udara hingga beberapa jam tergantung kondisi lingkungan, di mana PM2.5 dengan ukuran yang lebih kecil dapat bertahan di udara hingga satu minggu [31]. Properti ini memiliki potensi untuk memberikan relevansi yang tinggi dalam *forecasting* pada data *time series* sehingga PM2.5 dipilih menjadi variabel target.

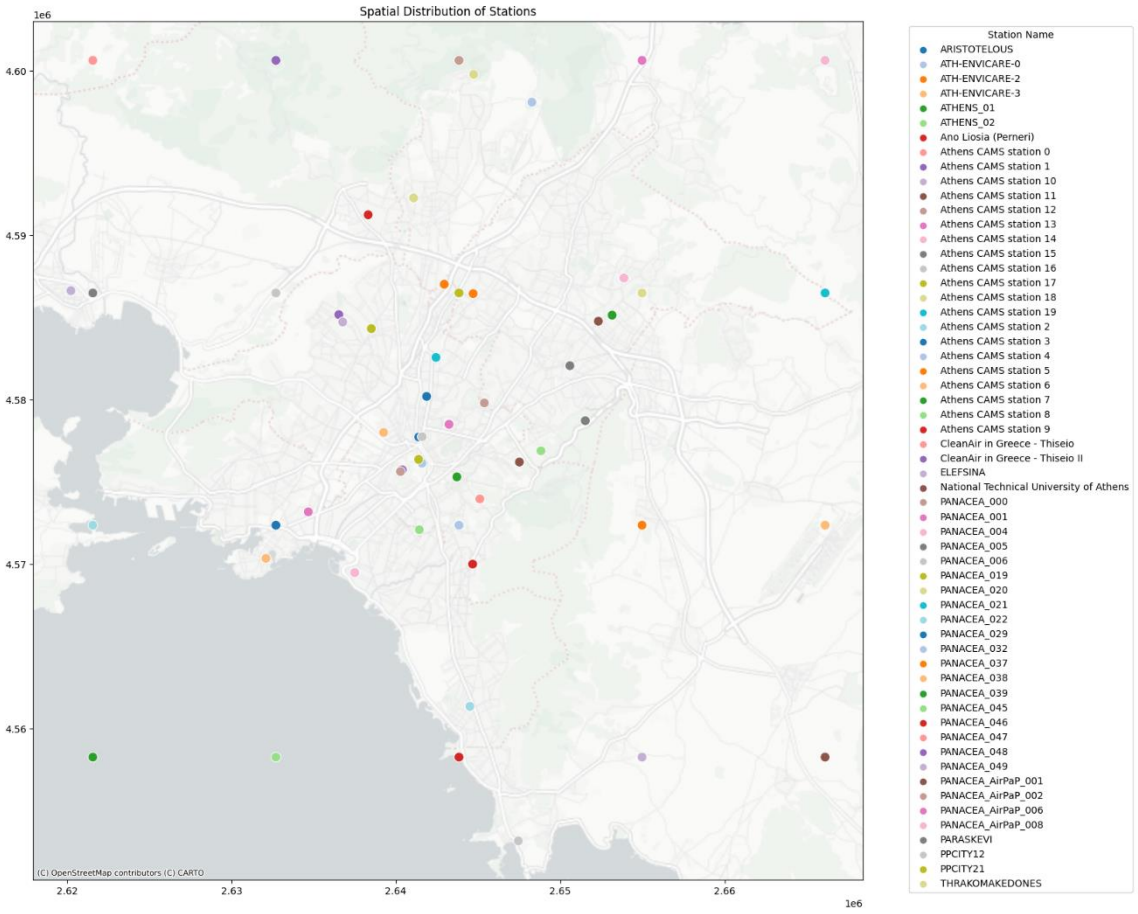
Hal pertama yang paling mudah dilakukan dalam melakukan analisis data ini adalah menganalisis statistik dataset secara keseluruhan untuk mempelajari karakteristik data secara sederhana. Berikut adalah statistik dari sampel data (polutan) yang digunakan dalam penelitian:

	PM10	**PM2.5**	**NO2**	**O3**
count	26976	26976	26976	26976
mean	20,23471	13,82152	17,41365	71,06858
std	16,6143	12,36514	17,03888	27,83211
min	0	0	0,313335	0,760259
25%	10,87977	7,491281	5,312168	53,91167
50%	15,99838	10,82617	11,48135	72,04289
75%	24	16,21273	23,99086	90
max	396,4567	343,5833	312	246,048

Tabel 3.1 Statistik Polutan pada Dataset

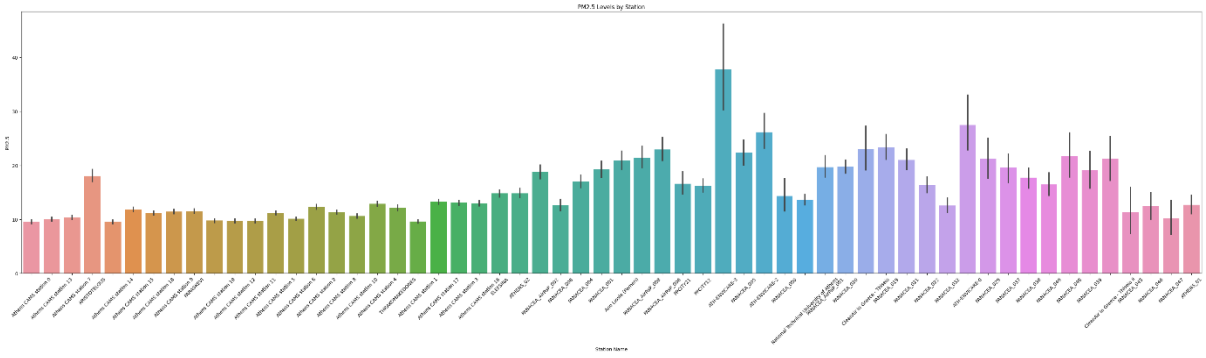
Konsentrasi rata-rata ground level ozone 71.068 μg/m³ dengan standar deviasi 27.832. Guideline WHO untuk rerata semitahunan konsentrasi O3 adalah 60 μg/m³ [31]. Konsentrasi rata-rata NO2 17.413 μg/m³ dengan standar deviasi 17.038, menunjukkan variasi yang cukup besar dari data pengamatan. Guideline WHO untuk rerata tahunan konsentrasi NO2 adalah 10 μg/m³ [31]. Konsentrasi rata-rata PM2.5 13.821 μg/m³ dengan standar deviasi 12.365, variasi yang juga cukup besar. Gudeiline WHO untuk rerata tahunan konsentrasi PM2.5 adalah 5μg/m³ [31]. Konsentrasi rata-rata PM10 20.234 μg/m³ dengan standar deviasi 16.614, variasi yang juga cukup besar. Guideline WHO untuk rerata tahunan konsentrasi PM10 adalah 15 μg/m³ [31]. Rerata konsentrasi polutan yang terekam pada dataset cukup fluktuatif (sebagai contoh, rerata PM2.5 yang terpaut hingga 330 μg/m³ dari nilai maksimalnya) dan berada di atas batas yang ditetapkan oleh panduan WHO.

Data konsentrasi polutan dan pembacaan sensor lainnya didapat dari bukan satu melainkan 58 stasiun monitoring berbeda yang tersebar di seluruh Kota Athena. Berikut adalah distribusi spasial dari AQMS yang menjadi sumber dari data yang digunakan:



Gambar 3.1 Distribusi Spasial AQMS

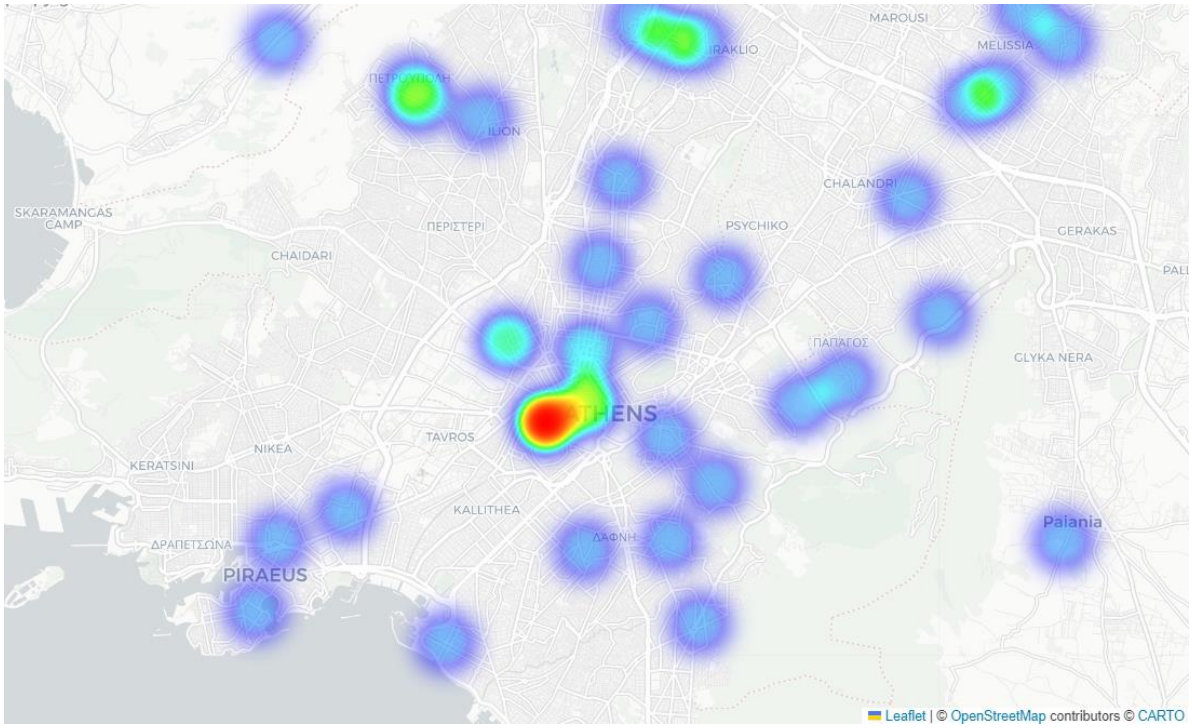
AQMS cenderung terpusat pada pusat kota Athena yang padat penduduk, namun AQMS juga terdapat pada daerah-daerah lain sehingga hasil pembacaan sensor AQMS dapat dikatakan representatif. Pusat kota cenderung memiliki aktivitas yang lebih pesat ketimbang daerah pinggir, sehingga terjadi kemungkinan perbedaan rerata konsentrasi polutan yang dibaca oleh stasiun di pusat kota dan di pinggir kota. Perbedaan ini dapat divisualisasikan menggunakan plot berikut:



Gambar 3.2 Rata-Rata Konsentrasi PM2.5 pada Tiap Stasiun

Seperti yang terlihat dalam barplot, terdapat stasiun yang cenderung mendapat pembacaan konsentrasi PM2.5 lebih besar ketimbang stasiun lainnya. Stasiun ATH-ENVICARE-3 yang terletak di Marathonomachon dan berada dekat Jalan Nasional memiliki

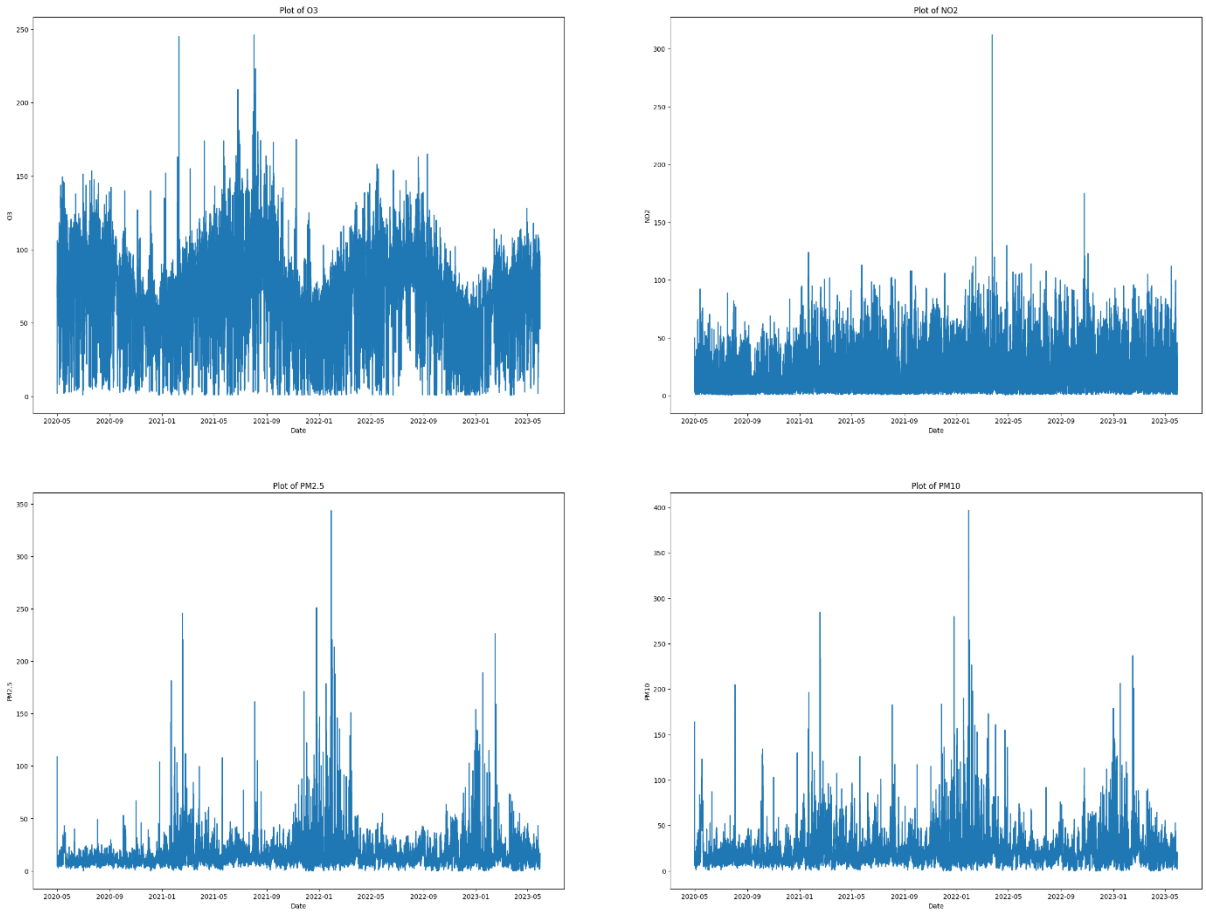
pembacaan konsentrasi PM2.5 tertinggi. Berikut adalah heatmap konsentrasi PM2.5 terhadap lokasi:



Gambar 3.3 *Heatmap* Konsentrasi PM2.5 pada Tiap Stasiun

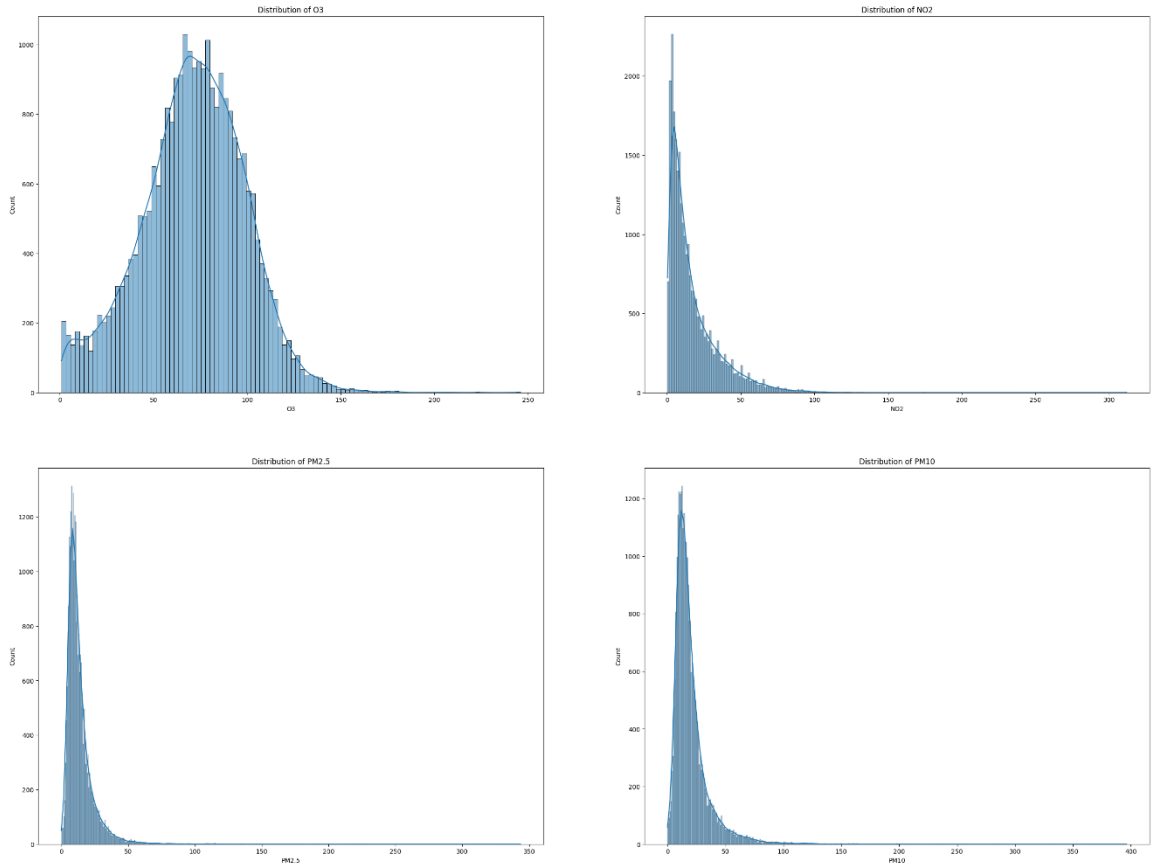
Stasiun dengan reratan konsentrasi PM2.5 tertinggi memang berada di pusat Kota Athena, namun terdapat juga beberapa stasiun di pinggiran kota yang memiliki rerata pembacaan konsentrasi PM2.5 yang cukup tinggi. Hal ini bisa saja dikarenakan tingginya emisi dari kota yang berdekatan. Harap diperhatikan bahwa radius pada heatmap tidak merepresentasikan jangkauan pembacaan AQMS.

Polutan menjadi kuantitas utama yang dimonitor oleh AQMS. Bagaimana konsentrasi polutan-polutan tersebut berubah setiap waktu? Mengetahui bagaimana pergerakan polutan dapat memberikan informasi yang berharga mengenai karakteristik data.



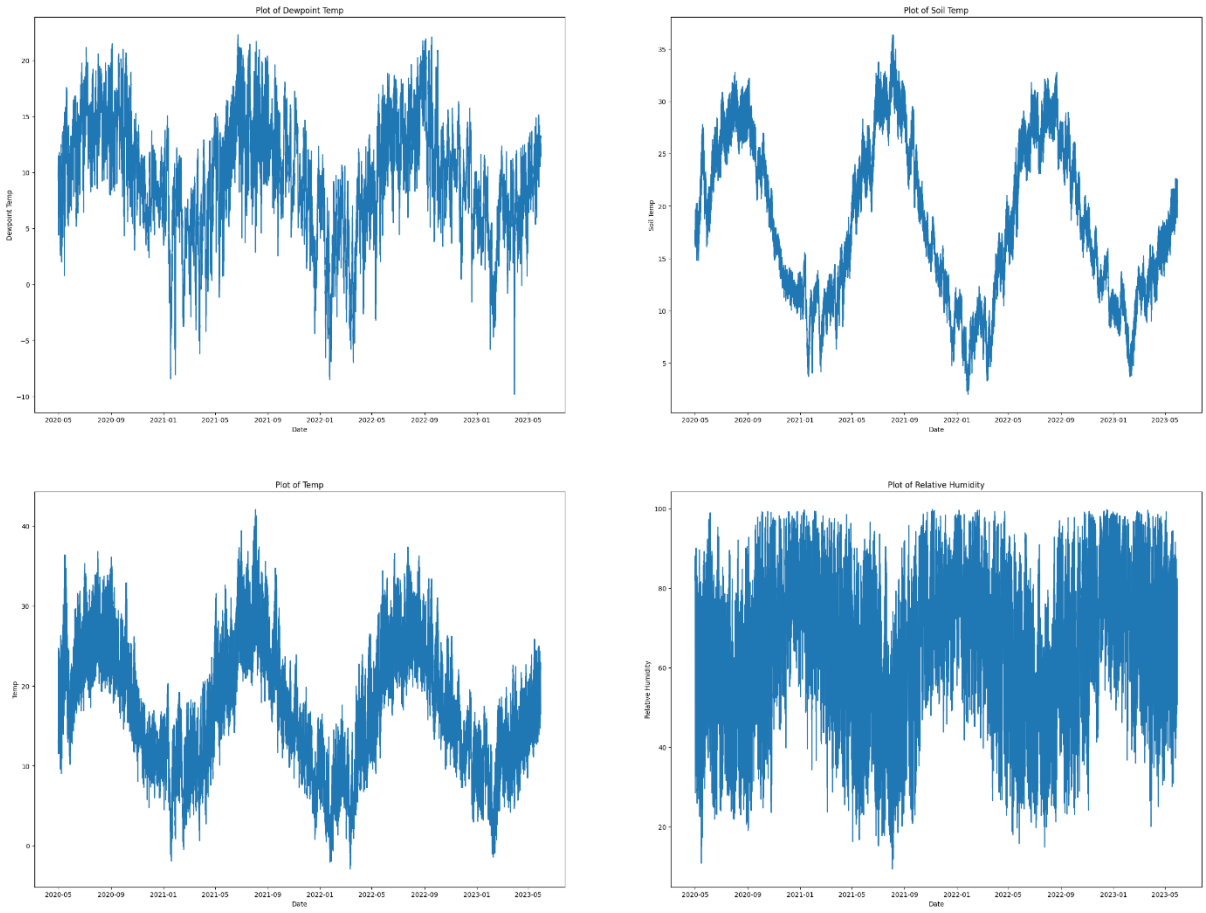
Gambar 3.4 Grafik Konsentrasi Polutan terhadap Waktu

Secara sekilas, konsentrasi PM2.5 dan PM10 setiap waktu memiliki pola yang sangat mirip secara visual. Hal ini tidak mengejutkan mengingat korelasinya yang sangat tinggi. Selain itu, seluruh PM2.5 juga secara teknis merupakan PM10. Terdapat lonjakan konsentrasi PM2.5 dan PM10 sekitar awal tahun. Penelitian ini tidak menggunakan PM10 sebagai salah satu target, namun observasi ini cukup menarik. Konsentrasi NO2 cenderung berada pada interval 50-120 μg/m³ setiap saat dengan lonjakan tinggi yang terjadi beberapa saat setelah terjadi lonjakan PM2.5 dan PM10. Ground level ozone (O3) menunjukkan pola yang naik turun sepanjang waktu. Bagaimana dengan distribusi data polutan?



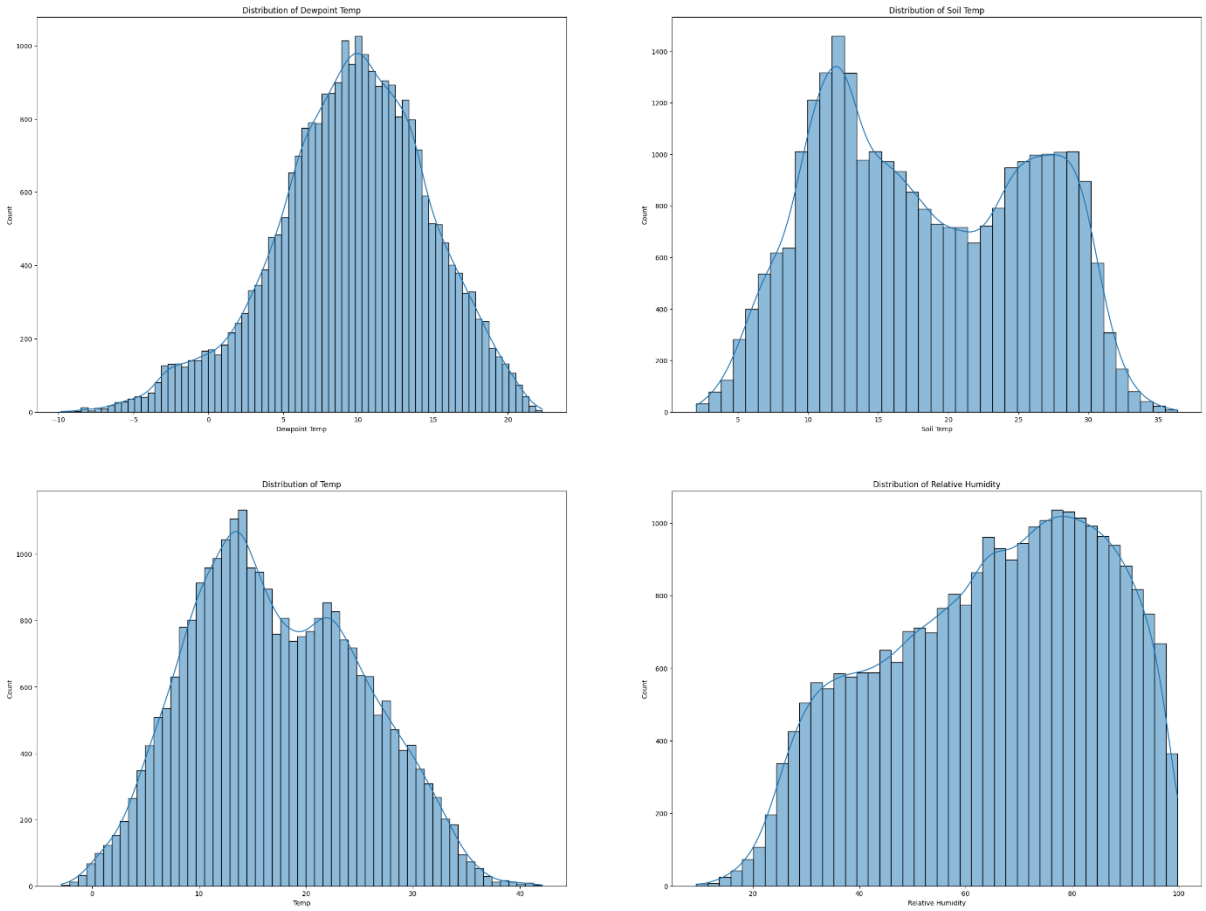
Gambar 3.5 Grafik Distribusi Polutan

Data cenderung terdistribusi secara log-normal dengan negative skew pada ground-level ozone dan postive skew pada polutan lainnya. Selain polutan, AQMS juga mengumpulkan data temperatur. Bagaimana perubahan temperatur dari waktu ke waktu yang terekam oleh stasiun monitoring?



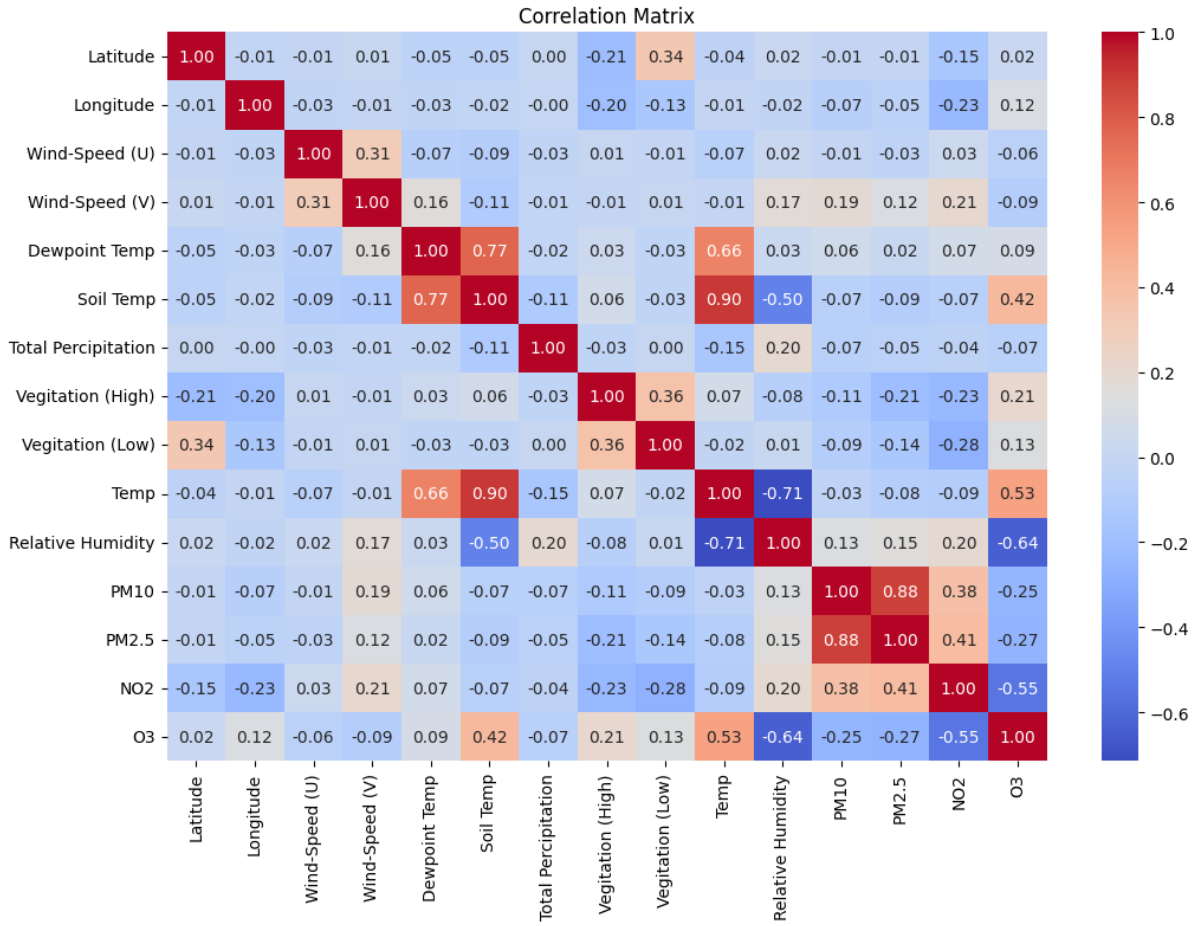
Gambar 3.6 Grafik Temperatur terhadap Waktu

Secara umum, temperatur lingkungan mengikuti pola perubahan yang sama dengan temperatur dewpoint dan temperatur tanah. Kelembaban cenderung naik ketika temperatur turun dan sebaliknya. Perubahan suhu yang terekam oleh AQMS mengikuti pola pergantian musim pada Kota Athena. Selanjutnya, bagaimana distribusi dari data-data temperatur ini?



Gambar 3.7 Grafik Distribusi Temperatur

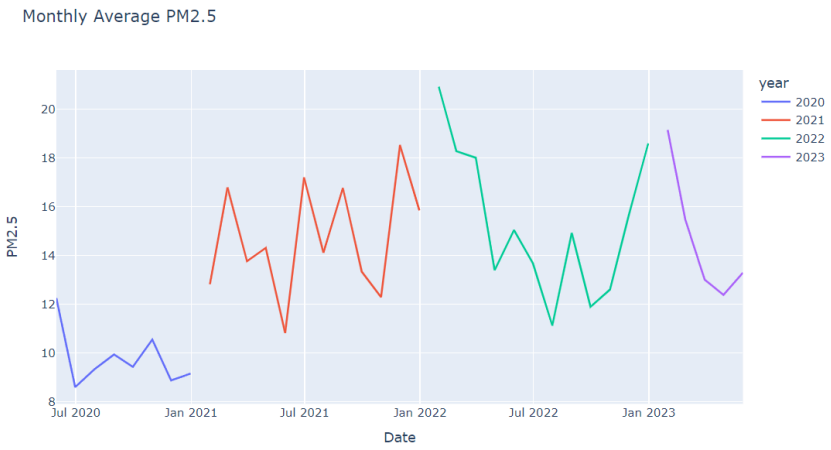
Setelah mengetahui karakteristik dari fitur-fitur yang dikumpulkan melalui sensor, analisis korelasi menggunakan correlation matrix dapat membantu lebih lanjut dalam mendapatkan insight mengenai data dan bagaimana hubungan setiap fitur terhadap target.



Gambar 3.8 *Correlation Matrix*

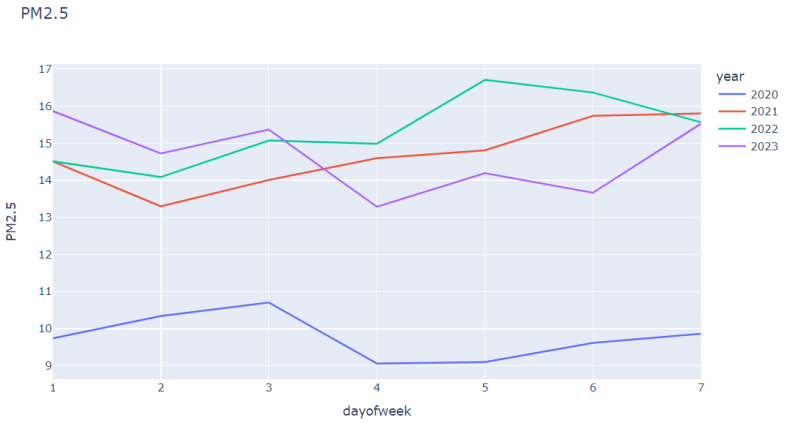
Terdapat beberapa insight yang diperoleh dari matriks korelasi. Temperatur berbanding terbalik dengan kelembaban. Hal ini terjadi karena suhu tinggi cenderung membuat lingkungan menjadi kering dan sebaliknya. Kemudian, konsentrasi ground level ozone juga dipengaruhi oleh temperatur dan kelembaban. Semakin panas suhu, semakin kering lingkungan, semakin tinggi konsentrasi ground level ozone. Ground level Ozone sendiri berpengaruh, kendati tidak secara signifikan, terhadap konsentrasi polutan PM10 dan PM2.5

Konsentrasi O3 serta NO2 memiliki korelasi linear terhadap PM10 dan PM2.5. Konsentrasi PM2.5 sangat dipengaruhi PM10. Hal ini masuk akal, karena PM2.5 secara teknis juga merupakan bagian dari PM10. Setelah mengetahui korelasi antarfitur pada data yang dimiliki, dilakukan analisis secara khusus pada properti time series untuk data kelas target.



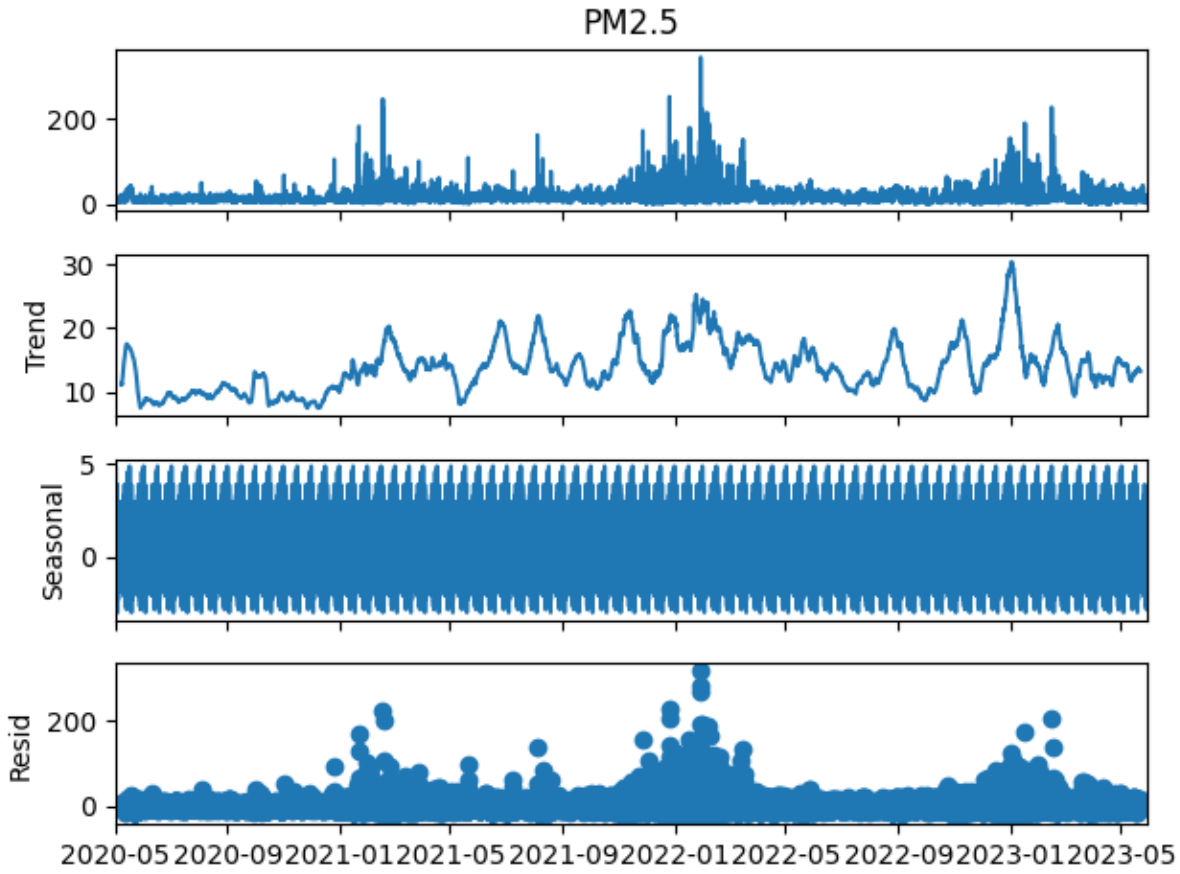
Gambar 3.9 Rata-Rata Bulanan Konsentrasi PM2.5

Terlihat lonjakan monthly average dari konsentrasi PM2.5 yang terjadi pada pergantian tahun. Terjadi kenaikan dan penurunan sepanjang tahun dan monthly average dari awal pengukuran pada bulan Mei 2020 tidak terpaut jauh dengan akhir pengukuran pada bulan Mei 2023, meskipun terdapat lonjakan dan penurunan yang signifikan sepanjang waktu pengamatan. Selanjutnya, dilakukan analisis rata-rata konsentrasi PM2.5 setiap harinya.



Gambar 3.10 Rata-Rata Harian Konsentrasi PM2.5

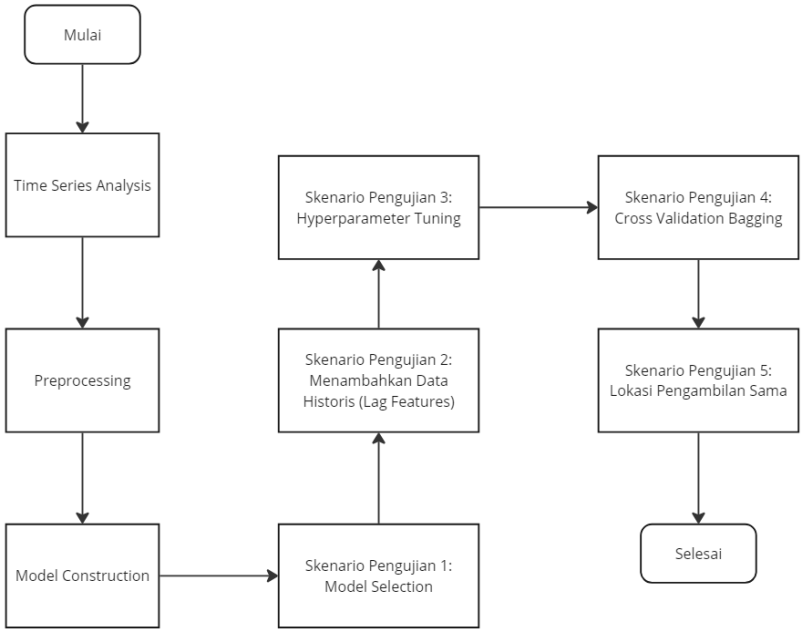
Terdapat beberapa insight yang didapat melalui visualisasi ini. Hal yang paling terlihat adalah bagaimana rata-rata konsentrasi PM2.5 cenderung jauh lebih rendah sepanjang hari pada tahun 2020 ketimbang tahun-tahun sesudahnya. Hal ini mungkin atau mungkin tidak dipengaruhi oleh faktor luar seperti pandemi dan aktivitas masyarakat yang berkurang. Hari rabu, kamis, minggu, dan rabu masing-masing menjadi hari dengan rata-rata konsentrasi PM2.5 tertinggi pada tahun 2020, 2021, 2022, dan 2023. Kemudian, dilakukan dekomposisi untuk mengetahui pola-pola time series pada data target.



Gambar 3.11 Plot Dekomposisi Konsentrasi PM2.5 terhadap Waktu

3.2 Desain Sistem

Pada tahapan pengujian, akan dilakukan beberapa langkah kerja yang telah disusun dalam *flowchart* berikut:



Gambar 3.12 *Flowchart* Langkah Kerja

Berikut penjelasan langkah kerja di atas:

1. Data Analysis
   Memahami data dengan melakukan visualisasi data, analisis data berdasarkan waktu (tren, pola musim), analisis distribusi serta korelasi.
2. Data Preprocessing
   Sampling data berdasarkan jam, menghilangkan baris yang memiliki nilai *NaN*, menambahkan fitur temporal, membagi dataset menjadi *train* dan *test set*, *drop* fitur redundan atau tidak berguna.
3. Model Construction
   Mendefinisikan loss function, model statistika, dan arsitektur model LSTM. Dalam implementasinya, model *Linear Regression* dan *Random Forest* akan menggunakan *library* scikit learn, XGBoost menggunakan *library* xgboost, dan LSTM menggunakan *library* keras.

```
model = LinearRegression()
```

Gambar 3.13 Mendefinisikan Model Linear Regression

```
model = RandomForestRegressor(random_state=random_state)
```

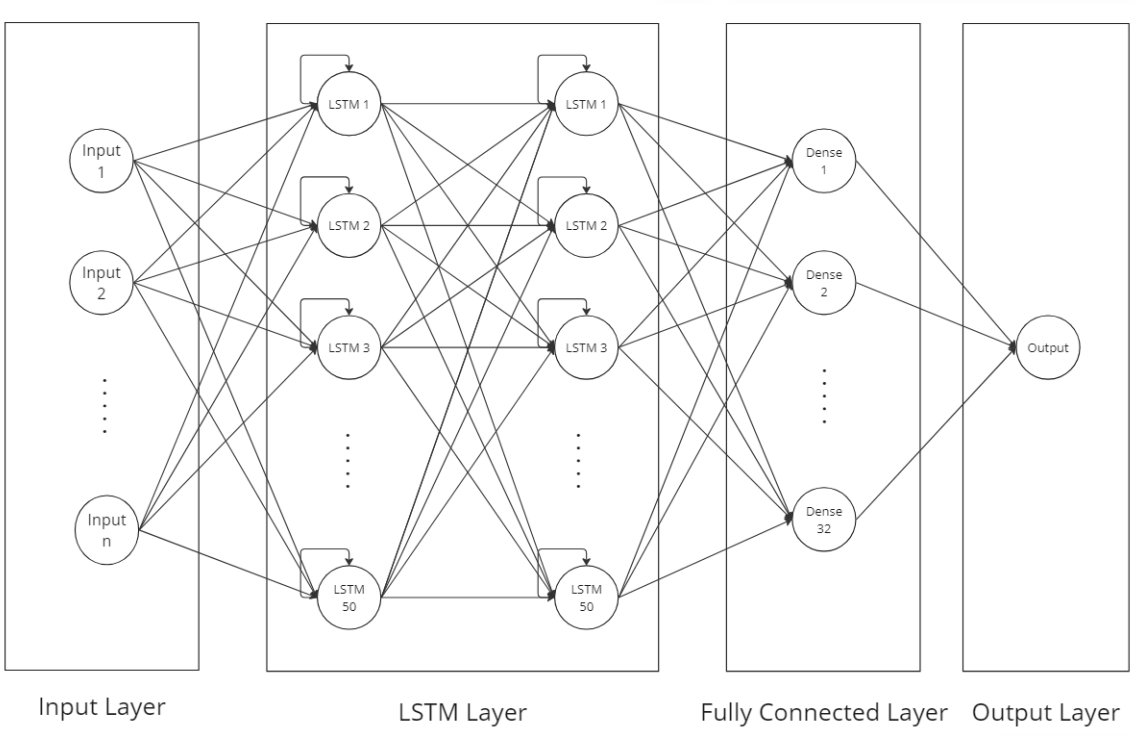
Gambar 3.14 Mendefinisikan Model Random Forest

```
model = XGBRegressor(random_state=random_state, verbose=-1)
```

Gambar 3.15 Mendefinisikan Model XGBoost Objective Squared Error

```
model = XGBRegressor(random_state=random_state, verbose=-1, objective='reg:absoluteerror')
```

Gambar 3.16 Mendefinisikan Model XGBoost Objective Absolute Error



Gambar 3.17 Arsitektur LSTM Model

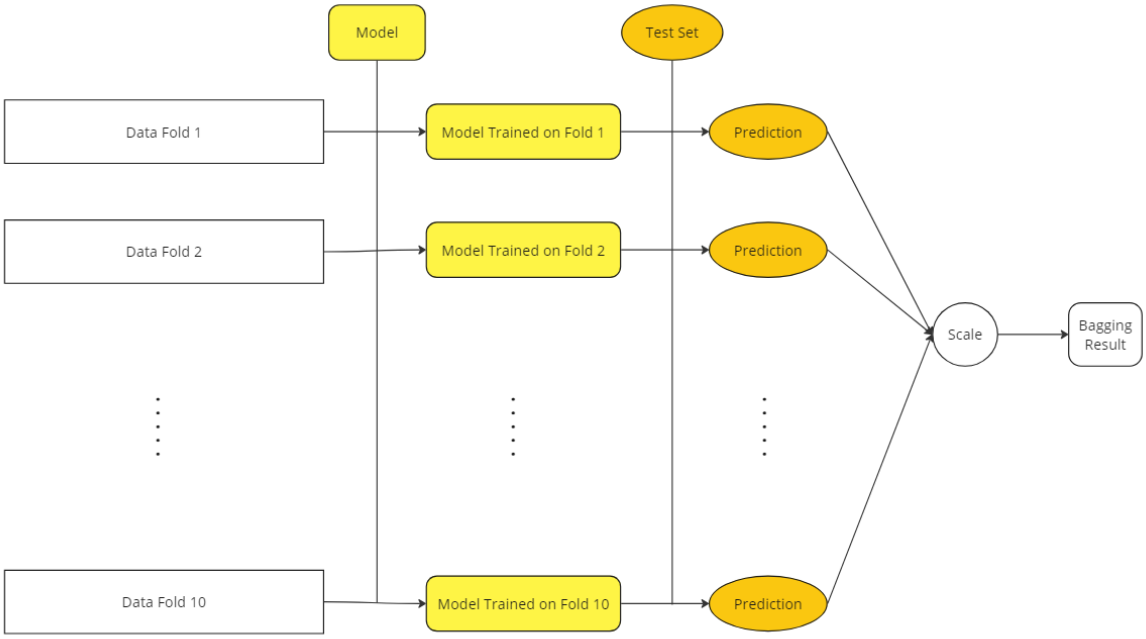
```
model = Sequential()

model.add(LSTM(50, activation='relu', return_sequences=True, input_shape=(1, X_train_seq.shape[2])))
model.add(LSTM(50, activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(1))

opt = keras.optimizers.Adam(1e-5)
model.compile(loss=smape_loss, optimizer=opt)
```

Gambar 3.18 Mendefinisikan Model LSTM (*lag feature*=1, smape loss)

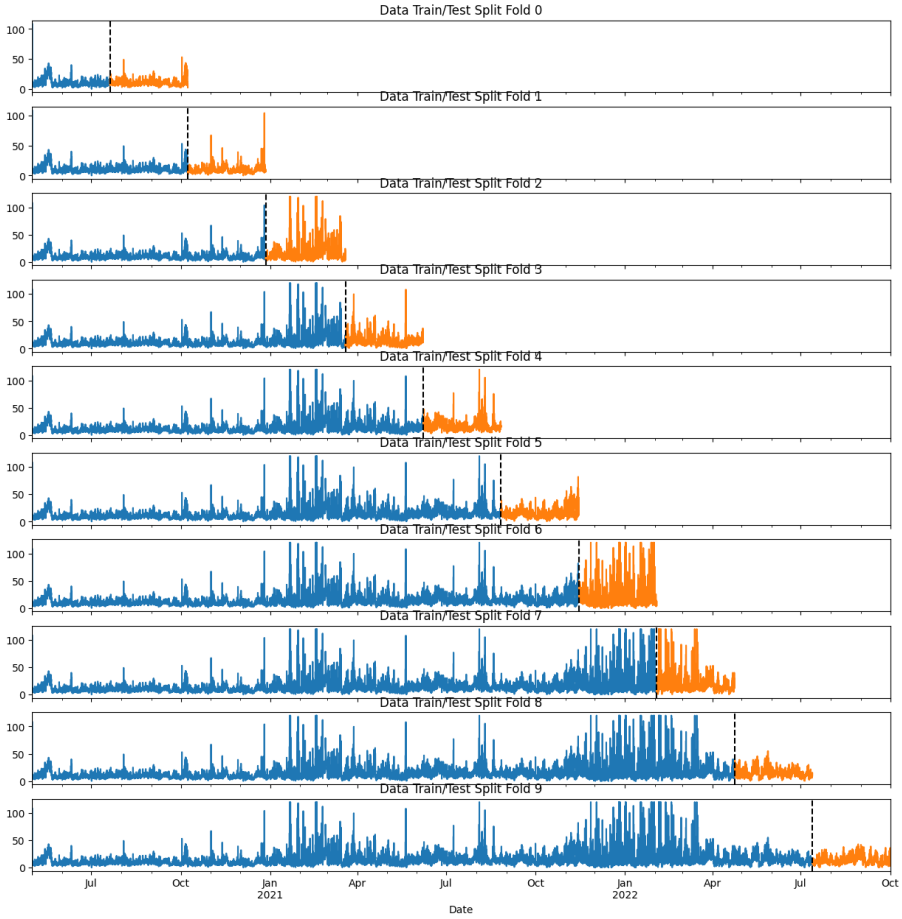
4. Skenario Pengujian 1
   Melakukan perbandingan performa (RMSE, MAE, R^2, dan SMAPE) antar model dimana model yang berperforma terbaik akan digunakan untuk skenario kedepannya. LSTM akan menggunakan 1 jam data historis agar adil.
5. Skenario Pengujian 2
   Melakukan perbandingan performa model terbaik dan LSTM jika ditambahkan data historis pada beberapa *time step* yang berbeda yakni 3 jam, 6 jam, 12 jam, 1 hari, 3 hari,

dan 1 minggu. Selang waktu dipilih berdasarkan fakta yang dipaparkan sebelumnya bahwa PM2.5 dapat bertahan di udara selama beberapa jam tergantung kondisi lingkungan, di mana PM2.5 dengan ukuran yang lebih kecil memiliki potensi untuk bertahan di udara selama satu hari hingga satu minggu [31]. LSTM akan kita *benchmark* lebih lanjut karena ada kemungkinan 1 jam kurang untuk menangkap hubungan fitur sebenarnya serta sebagai .

6. Skenario Pengujian 3
   Melakukan *hyperparameter tuning* menggunakan TPE pada model terbaik dan membandingkan performanya dengan model dengan *default parameter*. Dalam implementasinya akan digunakan *library* optuna.
7. Skenario Pengujian 4
   Model terbaik akan dievaluasi menggunakan *Expanding Window Cross Validation* 10 lipatan, sekaligus pada setiap lipatan model juga akan dilatih pada data lipatan tersebut dan memprediksi *test set*. Kemudian hasil-hasil prediksi *test set* tersebut akan di-*scale* (*bagging*) dan dibandingkan dengan prediksi model tanpa *bagging*.



Gambar 3.19 Skema Prediksi dengan *Bagging*



Gambar 3.20 Data Setiap Lipatan

8. Skenario Pengujian 5
   Membandingkan performa model menggunakan skema pengujian yang sama pada data yang dikumpulkan di stasiun yang sama.

4. HASIL DAN PEMBAHASAN

4.1 Model Selection

Berikut hasil perhitungan metrik evaluasi setiap model:

1. LSTM

Metric	Score
RMSE (*lower better*)	13.022997
MAE (*lower better*)	7.326285
R^2 (*higher better*)	-0.030483
SMAPE (*lower better*)	48.027059

Tabel 4.1 Metrik Evaluasi LSTM

LSTM menunjukkan RMSE dan MAE tertinggi, menandakan kinerja prediksi yang buruk. R^2 negatif menunjukkan bahwa model berkinerja lebih buruk daripada garis horizontal (mean data). Selain itu, nilai SMAPE-nya relatif tinggi.

2. *Linear Regression*

Metric	Score
RMSE (*lower better*)	11.298094
MAE (*lower better*)	7.584804
R^2 (*higher better*)	0.224323
SMAPE (*lower better*)	49.666334

Tabel 4.2 Metrik Evaluasi *Linear Regression*

Linear Regression berkinerja lebih baik daripada LSTM, dengan RMSE yang lebih rendah dan R^2 positif. Namun, MAE dan SMAPE-nya menunjukkan bahwa masih terdapat banyak kesalahan dalam prediksi.

3. *Random Forest*

Metric	Score
RMSE (*lower better*)	9.445501
MAE (*lower better*)	5.972396
R^2 (*higher better*)	0.457849
SMAPE (*lower better*)	40.197155

Tabel 4.3 Metrik Evaluasi *Random Forest*

Random Forest menunjukkan peningkatan yang signifikan dibandingkan LSTM dan *Linear Regression*. Model memiliki RMSE dan MAE yang lebih rendah, dan R^2 lebih tinggi, yang menunjukkan akurasi dan kecocokan prediksi yang lebih baik. Nilai SMAPE juga menunjukkan model yang lebih akurat dibandingkan model sebelumnya.

4. XGBoost (AE)

Metric	Score
RMSE (*lower better*)	9.622778
MAE (*lower better*)	5.676192
R^2 (*higher better*)	0.437307
SMAPE (*lower better*)	38.492745

Tabel 4.4 Metrik Evaluasi XGBoost (AE)

XGBoost dengan objective absolute error memberikan kinerja yang sedikit lebih baik dibandingkan Random Forest dalam hal MAE dan SMAPE, namun memiliki RMSE yang sedikit lebih tinggi. Nilai R^2 sebanding dengan Random Forest, menunjukkan kesesuaian yang baik.

5. XGBoost (SE)

Metric	Score
RMSE (*lower better*)	9.391797
MAE (*lower better*)	6.034848
R^2 (*higher better*)	0.463996
SMAPE (*lower better*)	41.617070

Tabel 4.5 Metrik Evaluasi XGBoost (SE)

XGBoost dengan objective squared error menunjukkan performa terbaik untuk RMSE dan R^2. Namun, MAE dan SMAPE-nya sedikit lebih tinggi dibandingkan model XGBoost (AE).

Berdasarkan hasil metrik evaluasi seluruh model, model XGBoost (AE) adalah yang berkinerja terbaik dengan MAE dan SMAPE, yang menunjukkan bahwa model tersebut memberikan prediksi paling akurat. RMSE dan R^2 untuk model tersebut sedikit lebih buruk dibandingkan model XGBoost (SE) namun dapat diabaikan karena perbedaan tidak terlalu signifikan.

4.3 Menambahkan Data Historis (Lag Features)

Berikut adalah tabel perbandingan performa model XGBoost dan LSTM dengan jumlah *lag feature* berbeda:

	Model-Time Step	RMSE	MAE	R^2	SMAPE
0	XGB-0 Jam	9.622778e+00	5.676192e+00	4.373072e-01	38.492745
1	XGB-3 Jam	1.227863e+01	7.049465e+00	8.399430e-02	45.776023
2	XGB-6 Jam	1.209477e+01	6.865667e+00	1.114391e-01	44.765481
3	XGB-12 Jam	1.211711e+01	6.913394e+00	1.089735e-01	45.237686
4	XGB-24 Jam	1.202294e+01	6.818276e+00	1.240288e-01	44.671468
5	XGB-72 Jam	1.212663e+01	6.871640e+00	1.115718e-01	44.466417
6	XGB-168 Jam	1.232596e+01	7.019906e+00	8.913654e-02	44.919695
7	LSTM-1 Jam	1.326782e+01	7.421870e+00	-6.959271e-02	48.606126
8	LSTM-3 Jam	1.296277e+01	7.246912e+00	-2.092511e-02	47.292854
9	LSTM-6 Jam	1.711408e+01	9.494013e+00	-7.790931e-01	57.728166
10	LSTM-12 Jam	4.024608e+01	1.776824e+01	-8.829681e+00	71.595893
11	LSTM-24 Jam	1.122071e+03	9.421464e+02	-7.628735e+03	193.224243
12	LSTM-72 Jam	6.145721e+03	3.154833e+03	-2.281844e+05	197.306292
13	LSTM-168 Jam	8.700310e+12	4.291645e+11	-4.538174e+23	199.860066

Gambar 4.1 Perbandingan Metrik Evaluasi Time-Step Berbeda

Model XGBoost tanpa *lag feature* (XGB-0) menunjukkan performa terbaik di semua metrik. Meningkatkan jumlah *lag feature* hanya menurunkan performa model seiring bertambahnya lag feature. Sama untuk model LSTM, meskipun secara teoritis cocok untuk data deret waktu, model tidak berfungsi dengan baik dengan data ini, performanya menurun tajam seiring bertambahnya *lag feature*.

Model kemungkinan besar berperforma lebih buruk dengan data historis karena pola dasar data sulit ditangkap secara efektif, sehingga menimbulkan *noise* atau *overfitting* pada model yang bergantung pada data historis seperti LSTM.

4.4 Hyperparameter Terbaik

Setelah melakukan hyperparameter tuning, performa model XGBoost meningkat sedikit. Parameter optimal yang diidentifikasi adalah sebagai berikut:

Parameter	Optimal Value
n_estimators	285
learning_rate	0.07000036538325966
max_depth	8
min_child_weight	9
subsample	0.566833353799025
colsample_bytree	0.9724450047679658
alpha	0.3386725243333994

Parameter	Optimal Value
lambda	0.5692566677681892

Tabel 4.6 Setelan Parameter Terbaik XGBoost

Dengan hyperparameter yang dioptimalkan ini, model mencapai metrik berikut pada trial terbaik:

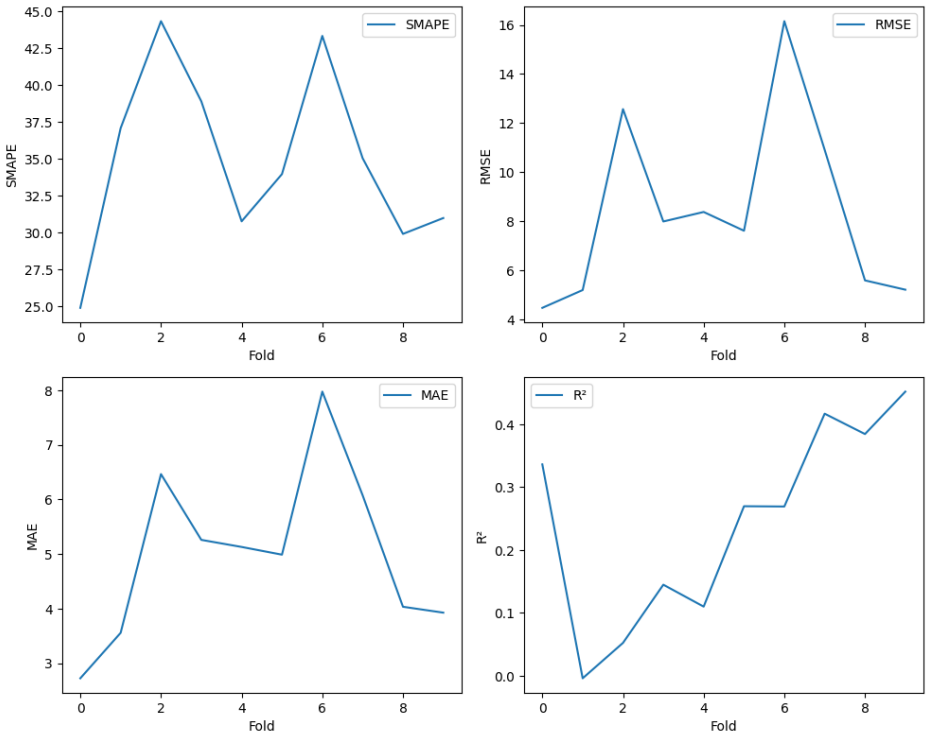
Metric	Tanpa Parameter Tuning	Dengan Parameter Tuning
RMSE (*lower better*)	9.622778	9.393333
MAE (*lower better*)	5.676192	5.486492
R^2 (*higher better*)	0.437307	0.463820
SMAPE (*lower better*)	38.492745	36.943337

Tabel 4.7 Perbandingan Metrik Evaluasi XGBoost tanpa/dengan *Hyperparameter Tuning*

Dari peningkatan dalam metrik evaluasi, khususnya pengurangan SMAPE, RMSE, dan SMAPE, serta R^2 yang relatif stabil, dapat disimpulkan bahwa proses hyperparameter tuning berhasil meningkatkan akurasi dan keandalan prediksi model.

4.5 Time Series Cross Validation Bagging

Setelah dilakukan *time series cross validation* 10 lipat, berikut adalah hasil metrik evaluasi setiap lipatannya:

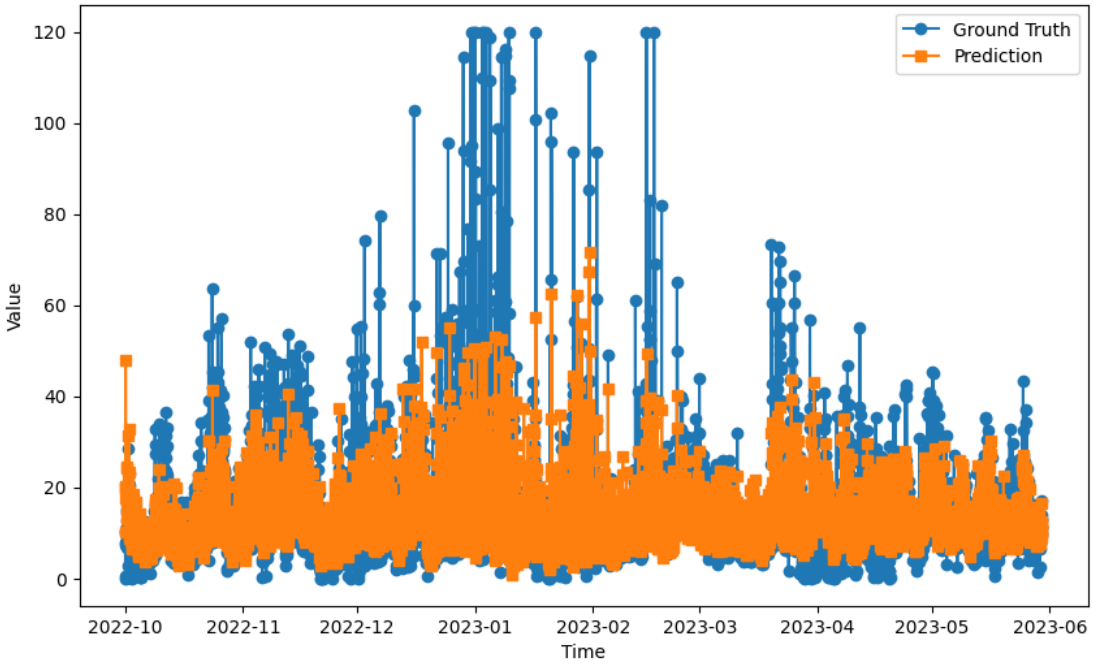


Gambar 4.2 Metrik Evaluasi setiap Lipatan

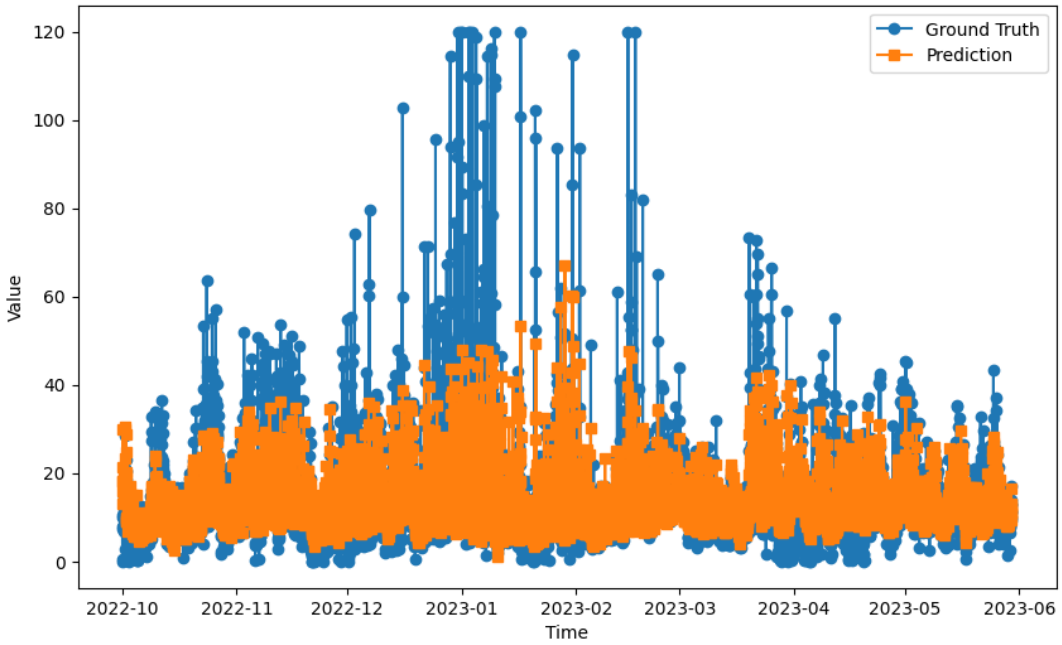
Metric	Mean
RMSE (*lower better*)	8.295249
MAE (*lower better*)	4.971455
R^2 (*higher better*)	0.254872
SMAPE (*lower better*)	34.826916

Table 4.8 Rata-Rata Metrik Evaluasi *Cross Validation*

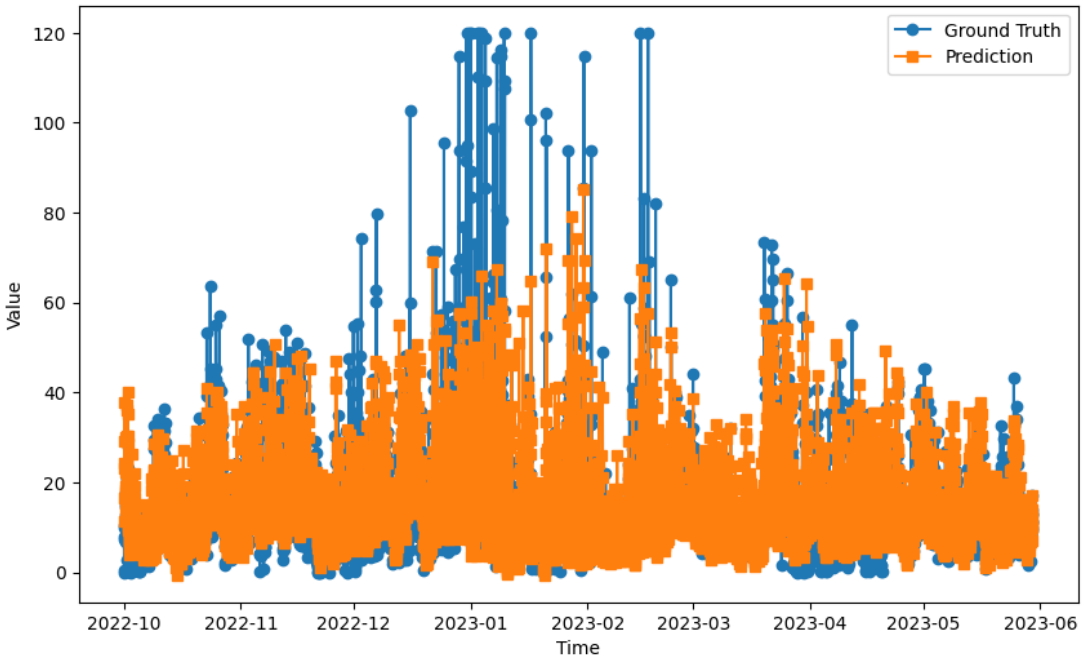
Ada variabilitas dalam metrik evaluasi di berbagai bidang. Misalnya, Lipatan ke-1 memiliki RMSE terendah (4,47) dan SMAPE (24,88), kondisi dimana pola pada data relatif sederhana. Sebaliknya, Lipatan ke-7 menunjukkan RMSE tertinggi (16,15) dan SMAPE (43,34), yang menunjukkan model kesulitan memprediksi fluktuasi secara efektif.



Gambar 4.3 Prediksi *Base* XGBoost



Gambar 4.4 Prediksi XGBoost Parameter Teroptimasi



Gambar 4.5 Prediksi XGBoost Parameter Teroptimasi dan *Bagging*

Plot yang disediakan membandingkan *ground truth* dan prediksi membuktikan bahwa model mengalami kesulitan dalam memprediksi data dengan nilai ekstrem secara akurat. *Ground truth* menunjukkan beberapa lonjakan tajam yang gagal ditangkap oleh prediksi secara akurat. Memprediksi dengan *bagging* membantu sedikit dalam memprediksi nilai ekstem, namun masalah yang sama tetap ada. Model ini secara konsisten meng-*underestimate* nilai dalam data.

Meskipun terdapat kesulitan dalam menentukan puncak, model ini mampu menangkap tren umum dan distribusi data secara keseluruhan dengan cukup baik. Prediksi tersebut

mengikuti bentuk *ground truth* yang luas, meskipun dengan magnitudo yang lebih rendah untuk puncaknya.

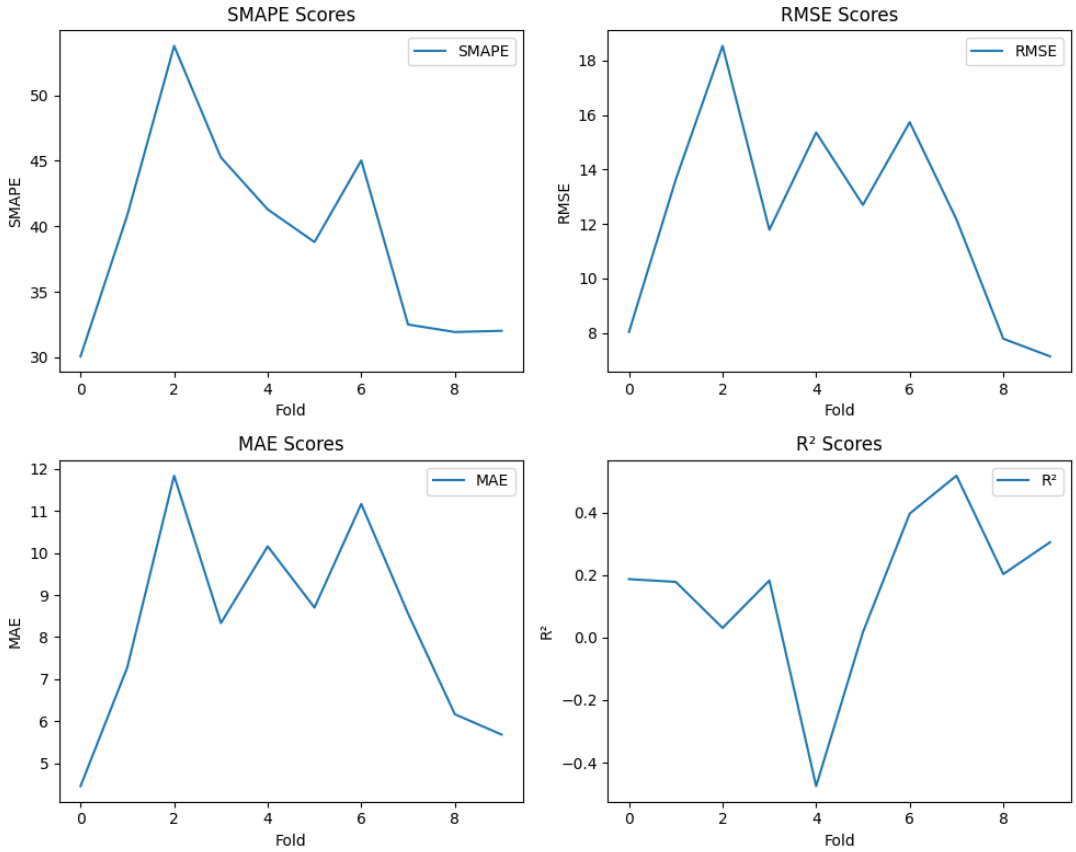
4.6 Segmentasi Lokasi Data

Berdasarkan hasil pada sub-bab sebelumnya, variabilitas yang disebabkan oleh penggunaan data dari stasiun yang berbeda kemungkinan besar berkontribusi terhadap performa model yang kurang optimal. Hal tersebut menimbulkan sebuah pertanyaan, bagaimana performa model yang dilatih pada data dari satu stasiun?

Terdapat beberapa alasan yang mendasari pertanyaan ini. PM2.5 dipilih karena kemampuannya untuk bertahan di udara selama beberapa jam hingga dalam kasus ukuran partikel yang sangat kecil satu minggu. Dengan kata lain, *lag features* PM2.5 memiliki potensi untuk memberikan kontribusi besar terhadap hasil *forecasting* dalam kasus ini, karena konsentrasi PM2.5 saat prediksi dipengaruhi oleh konsentrasi PM2.5 beberapa jam hingga satu minggu sebelumnya.

Berangkat dari premis ini, pendekatan yang dilakukan sebelumnya dalam menyiapkan data memiliki suatu kecacatan. Data pembacaan sensor diperoleh dari lebih dari 50 stasiun yang tersebar di seluruh penjuru Kota Athena. Keputusan untuk mengambil sampel data dari seluruh stasiun yang tersedia diperoleh atas dasar diperlukan banyak titik pengambilan data agar hasil pembacaan sensor merepresentasikan Kota Athena, ketimbang suatu desa di Kota Athena, misalnya. Akibat yang ditimbulkan adalah, karena data yang diperoleh dari suatu waktu t dan $t + 1$ bisa saja berasal dari stasiun yang berbeda, terjadi ketidaksinambungan data. Hal ini diperkuat dengan hasil analisis data yang memberikan informasi bahwa rata-rata konsentrasi PM2.5 yang diperoleh AQMS di pusat kota cenderung lebih besar ketimbang AQMS yang jauh dari pusat kota. Dampak yang lebih buruk bisa saja terjadi pada kasus di mana data AQMS pusat kota diikuti dengan data AQMS yang jauh dari pusat kota pada urutan deret waktu setelahnya.

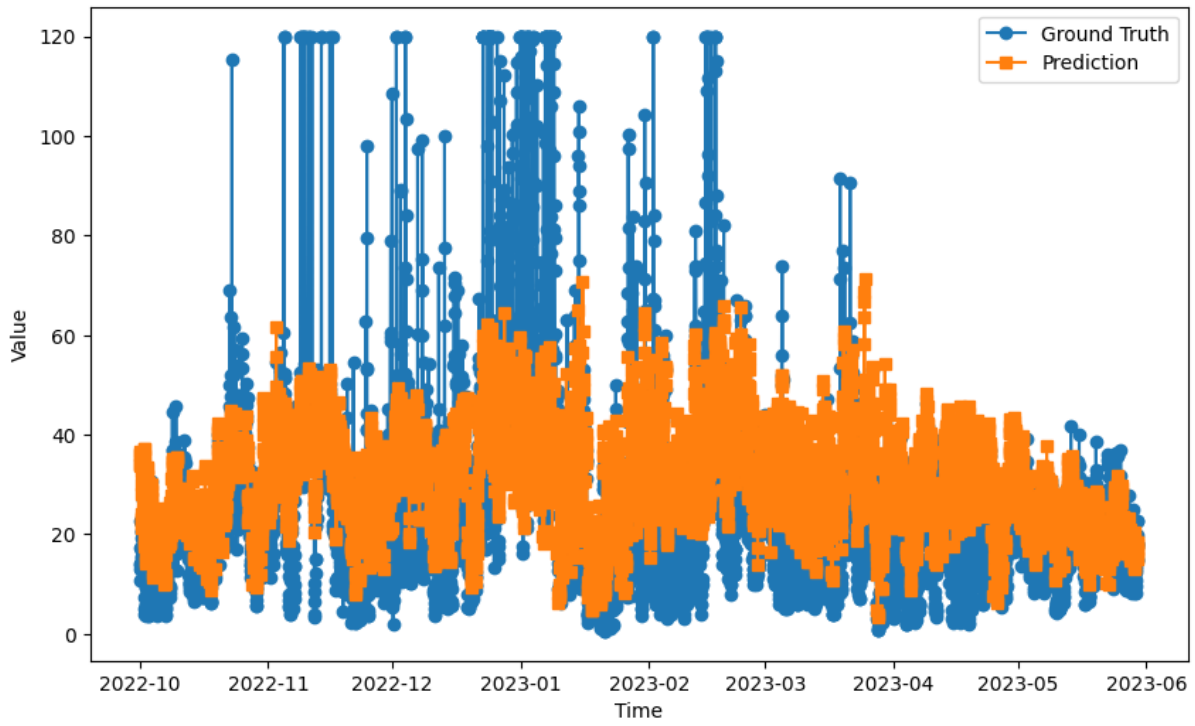
Setelah dilakukan tahap pengujian yang sama pada data di stasiun ‘ATH-ENVICARE-0', didapatkan hasil sebagai berikut:



Gambar 4.6 Metrik Evaluasi setiap Lipatan Data Satu Stasiun

Metric	Mean
RMSE (*lower better*)	12.287573
MAE (*lower better*)	8.234418
R^2 (*higher better*)	0.154376
SMAPE (*lower better*)	39.153712

Tabel 4.7 Rata-Rata Metrik Evaluasi *Cross Validation* Data Satu Stasiun



Gambar 4.8 Prediksi Data Satu Stasiun, XGBoost Parameter Teroptimasi dan *Bagging*

Prediksi model pada data stasiun yang sama, seperti yang ditunjukkan pada gambar 4.18, menunjukkan penangkapan tren umum yang lebih baik dibandingkan dengan hasil sebelumnya. Namun, model ini masih kesulitan dalam memprediksi puncak secara akurat. Demikian pula, model ini mengalami kesulitan dalam memprediksi nilai yang lebih rendah, dan sering kali *overestimate*.

Metrik evaluasi secara keseluruhan model menunjukkan penurunan performa dibandingkan sebelumnya. Heterogenitas dalam pola polusi dan variabilitas waktu di berbagai stasiun kemungkinan besar berkontribusi terhadap penurunan kinerja ini.

5. KESIMPULAN

5.1 Kesimpulan

Konsentrasi PM2.5 pada Kota Athena cenderung menunjukkan pola naik dan turun sepanjang tahun dan pola seasonal cenderung konsisten. Terdapat beberapa lonjakan dari waktu ke waktu yang mungkin disebabkan karena anomali seperti cuaca ekstrem dan lainnya. Konsentrasi PM2.5 dan polutan-polutan lainnya memiliki variabilitas yang sangat tinggi namun tersebar mengikuti distribusi log normal dengan kecondongan pada satu arah. Ditemukan juga korelasi linear antara polutan, dan hingga batas tertentu, temperatur lingkungan, terhadap konsentrasi PM2.5. Pembacaan PM2.5 dilakukan melalui sensor AQMS yang tersebar di seluruh penjuru kota Athena, dengan pola AQMS yang berada di pusat kota cenderung memperoleh rata-rata konsentrasi PM2.5 yang lebih tinggi ketimbang AQMS yang jauh dari pusat kota.

Setelah dilakukan berbagai skenario pengujian, evaluasi model menunjukkan bahwa XGBoost mengungguli model lain seperti LSTM, *Linear Regression,* dan *Random Forest* dalam memprediksi kualitas udara PM 2,5 di Athena, Yunani. H*yperparameter tuning* memberikan sedikit peningkatan performa, dimana inklusi data historis memperburuk performa dikarenakan pola fluktuatif data yang sulit ditangkap secara efektif, sehingga menimbulkan *noise* atau *overfitting* pada model. Hasil *Time Series* C*ross Validation* menunjukkan variabilitas kinerja di berbagai lipatan terutama pada lipatan yang memiliki nilai ekstrem. Model memiliki kesulitan dalam memprediksi nilai ekstrem secara akurat yang dibuktikan dalam prediksi model yang cenderung halus dibandingkan *ground truth*. Metode *bagging* dapat membantu kecenderungan tersebut meskipun hanya sedikit serta dengan membantu model agar mampu menangkap tren umum dan mendistribusikan data secara keseluruhan dengan cukup baik meskipun dengan nilai yang sedikit lebih rendah. Terdapat kemungkinan bahwa data historis kurang representatif dikarenakan penggunaan data di lokasi yang berbeda, namun hasil mengatakan sebaliknya karena tidak terdapat peningkatan prediksi yang signifikan jika model dilatih pada data pada satu lokasi.

5.2 Saran

Pada penelitian selanjutnya, masih terdapat beberapa algoritma regresi yang dapat dieksplorasi untuk kasus permasalahan yang sama, terutama model *tree-based* karena performa yang relatif konsisten. *Feature engineering* yang lebih komprehensif juga dapat membantu model dalam menangkap pola yang lebih kompleks. Metode *ensemble learning* lainnya juga dapat dieksplorasi lebih lanjut karena seperti yang sudah disertakan dalam bab sebelumnya, metode *bagging* membantu dalam kecenderungan pola prediksi model yang meng-*underestimate* nilai dalam data. Selain itu, metode *ensemble learning* tersebut dapat digunakan pada data dengan metode *cross validation* yang berbeda, yang mungkin lebih representatif untuk data yang digunakan.

Selain itu, penggunaan interval waktu sampling yang lebih kecil dapat memberikan insight yang lebih baik dengan trade-off terhadap kebutuhan komputasi. Peneliti selanjutnya

juga dapat bereksperimen menggunakan dataset lain dengan polutan berbeda untuk mendapat insight lebih lanjut mengenai bagaimana tiap polutan mempengaruhi konsentrasi PM2.5.

DAFTAR PUSTAKA

[1] C. J. L. Murray *et al.*, "Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1223–1249, Oct. 2020, doi: 10.1016/S0140-6736(20)30752-2.

[2] M. Theil, "Ella Kissi-Debrah death: Mum continues clean air fight 10 years on," *Big Issue*, London, Feb. 15, 2023. Accessed: Jun. 22, 2024. [Online]. Available: https://www.bigissue.com/news/environment/air-pollution-nothing-fills-that-void-rosamund-adoo-kissi-debrah-on-ten-years-since-her-daughter/

[3] S. Laville, "Air pollution a cause in girl's death, coroner rules in landmark case," *The Guardian*, London, Dec. 16, 2020.

[4] S. Holgate, S. Wilson, D. E. Davies, and M. Loxham, "The health effects of air pollution: Driving policy and legislative change to improve air quality and protect public health," *University of Southampton*, Southampton, 2020. Accessed: Jun. 22, 2024. [Online]. Available: https://www.southampton.ac.uk/medicine/research/impact/health-effects-of-air-pollution.page

[5] A. Gentleman, "Mother of girl whose death was linked to air pollution sues UK government," *The Guardian*, London, Jan. 25, 2024. Accessed: Jun. 22, 2024. [Online]. Available: https://www.theguardian.com/environment/2024/jan/25/mother-of-girl-who-died-from-air-pollution-sues-uk-government

[6] J. Lelieveld *et al.*, "Air pollution deaths attributable to fossil fuels: observational and modelling study," *BMJ*, p. e077784, Nov. 2023, doi: 10.1136/bmj-2023-077784.

[7] Health Effects Institute, "State of Global Air 2019," Boston, 2019.

[8] D. P. Turner, "Sampling Methods in Research Design," *Headache: The Journal of Head and Face Pain*, vol. 60, no. 1, pp. 8–12, Jan. 2020, doi: 10.1111/head.13707.

[9] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[10] A. K. Kuchibhotla, L. D. Brown, A. Buja, and J. Cai, "All of Linear Regression," Oct. 2019.

[11] J. M. Sangeetha and K. J. Alfia, “Financial stock market forecast using evaluated linear regression based machine learning technique,” *Measurement: Sensors*, vol. 31, p. 100950, Feb. 2024, doi: 10.1016/j.measen.2023.100950.

[12] F. X. Diebold, M. Göbel, and P. Goulet Coulombe, “Assessing and comparing fixed-target forecasts of Arctic sea ice: Glide charts for feature-engineered linear regression and machine learning models,” *Energy Econ*, vol. 124, p. 106833, Aug. 2023, doi: 10.1016/j.eneco.2023.106833.

[13] I. Ilic, B. Görgülü, M. Cevik, and M. G. Baydoğan, “Explainable boosted linear regression for time series forecasting,” *Pattern Recognit*, vol. 120, p. 108144, Dec. 2021, doi: 10.1016/J.PATCOG.2021.108144.

[14] S. L. KUMAR, V. R. SAROBIN M, and J. ANBARASI L, “Predictive Analytics of COVID-19 Pandemic: Statistical Modelling Perspective,” *Walailak Journal of Science and Technology (WJST)*, vol. 18, no. 16, Aug. 2021, doi: 10.48048/wjst.2021.15583.

[15] J. Jayasinghe, P. Ekanayake, O. Panahatipola, C. I. Madhushani, and U. Rathnayake, “Forecasting the power generation at renewable power plants in Sri Lanka using regression trees,” *Results in Engineering*, vol. 22, p. 102111, Jun. 2024, doi: 10.1016/J.RINENG.2024.102111.

[16] Q. Lei, H. Yu, and Z. Lin, “Understanding China’s CO2 emission drivers: Insights from random forest analysis and remote sensing data,” *Heliyon*, vol. 10, no. 7, p. e29086, Apr. 2024, doi: 10.1016/J.HELIYON.2024.E29086.

[17] H. Zhang, J. Peng, R. Wang, M. Zhang, C. Gao, and Y. Yu, “Use of random forest based on the effects of urban governance elements to forecast CO2 emissions in Chinese cities,” *Heliyon*, vol. 9, no. 6, p. e16693, Jun. 2023, doi: 10.1016/J.HELIYON.2023.E16693.

[18] B. K. Meher, M. Singh, R. Birau, and A. Anand, “Forecasting stock prices of fintech companies of India using random forest with high-frequency data,” *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 1, p. 100180, Mar. 2024, doi: 10.1016/J.JOITMC.2023.100180.

[19] B. Gaertner, “Geospatial patterns in runoff projections using random forest based forecasting of time-series data for the mid-Atlantic region of the United States,” *Science of The Total Environment*, vol. 912, p. 169211, Feb. 2024, doi: 10.1016/J.SCITOTENV.2023.169211.

[20] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, “An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach,” *MethodsX*, vol. 10, p. 102119, 2023, doi: 10.1016/j.mex.2023.102119.

[21] Z. Jiang, J. Che, M. He, and F. Yuan, “A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition,” *Renew Energy*, vol. 203, pp. 802–827, Feb. 2023, doi: 10.1016/j.renene.2022.12.124.

[22] L. Zhang and D. Jánošík, “Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches,” *Expert Syst Appl*, vol. 241, p. 122686, May 2024, doi: 10.1016/j.eswa.2023.122686.

[23] Y. Xu, S. Zheng, Q. Zhu, K. Wong, X. Wang, and Q. Lin, “A complementary fused method using GRU and XGBoost models for long-term solar energy hourly forecasting,” *Expert Syst Appl*, vol. 254, p. 124286, Nov. 2024, doi: 10.1016/j.eswa.2024.124286.

[24] A. Yadav, C. K. Jha, and A. Sharan, “Optimizing LSTM for time series prediction in Indian stock market,” *Procedia Comput Sci*, vol. 167, pp. 2091–2100, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.257.

[25] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci Rep*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[26] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geosci Model Dev*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[27] 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, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[28] A. Jierula, S. Wang, T.-M. OH, and P. Wang, “Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data,” *Applied Sciences*, vol. 11, no. 5, p. 2314, Mar. 2021, doi: 10.3390/app11052314.

[29] S. Watanabe, “Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance,” Apr. 2023.

[30] G.-F. Angelis, “Regional Datasets for Air Quality Monitoring in European Cities,” in *2024 IEEE International Geoscience and Remote Sensing Symposium (IEEE IGARSS-24)*, Athens: Zenodo, May 2024. doi: 10.5281/zenodo.11220965.

[31] Organización Mundial de la Salud (OMS), “WHO global air quality guidelines,” *Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*, pp. 1–360, 2021.