# BAB II
# KAJIAN LITERATUR

## 2.1 Kualitas Udara dan Indeks Kualitas Udara (AQI)

Pemahaman mendalam tentang kualitas udara dan indeks yang digunakan untuk mengukurnya merupakan fondasi penting dalam studi prediksi pencemaran udara. Sub-bab ini akan membahas secara sistematis mengenai definisi pencemaran udara, dampak berbagai polutan terhadap kesehatan masyarakat, konsep *Air Quality Index* (AQI) beserta metode perhitungannya berdasarkan standar US EPA, serta sistem pemantauan kualitas udara yang berlaku di Indonesia.

Pencemaran udara didefinisikan sebagai masuknya atau tercampurnya zat, energi, dan/atau komponen lain ke dalam udara ambien oleh kegiatan manusia atau proses alam, sehingga kualitas udara turun sampai ke tingkat tertentu yang dapat membahayakan kesehatan manusia dan lingkungan [17]. Polutan utama yang menjadi indikator kualitas udara secara global diklasifikasikan menjadi *criteria pollutants* oleh badan lingkungan Amerika Serikat (EPA). Polutan tersebut meliputi *Particulate Matter (PM2.5* dan *PM10)*, *Ozon Troposferik ($O_3$)*, *Nitrogen Dioksida ($NO_2$)*, *Sulfur Dioksida ($SO_2$)*, dan *Karbon Monoksida (CO)* [1].

PM2.5 dan PM10 merujuk pada partikel dengan diameter aerodinamik kurang dari 2.5 μm dan 10 μm, yang mampu menembus saluran pernapasan [2]. PM2.5 dapat menembus hingga ke *alveolus* dan masuk ke aliran darah, memicu penyakit *kardiovaskular*, *respiratori kronis*, dan penurunan fungsi paru [3]. $O_3$ terbentuk melalui reaksi fotokimia antara *Nitrogen Oksida (NOx)* dan Senyawa Organik Volatil (VOC) di bawah sinar matahari [18]. $O_3$ bersifat oksidan kuat yang dapat menyebabkan peradangan saluran napas, memperburuk asma, dan mengurangi fungsi paru [4]. Sementara itu, $NO_2$ dan $SO_2$ umumnya dihasilkan dari proses pembakaran bahan bakar fosil [19]. Kedua gas ini berkontribusi terhadap terjadinya hujan asam serta iritasi pada mukosa saluran pernapasan, meningkatkan risiko infeksi pernapasan [4]. Kelompok rentan seperti anak-anak, lansia, dan penderita penyakit paru obstruktif kronis (PPOK) memiliki sensitivitas lebih tinggi terhadap penurunan kualitas udara [8].

*Air Quality Index* (AQI) adalah angka tidak berdimensi yang digunakan untuk melaporkan kualitas udara harian secara sederhana dan mudah dipahami masyarakat [20]. Semakin tinggi nilai AQI, semakin besar tingkat pencemaran dan potensi risiko kesehatan. *United States Environmental Protection Agency (US EPA)* menetapkan metode perhitungan

AQI dengan pendekatan *piecewise linear function* untuk setiap polutan [21]. Konsentrasi polutan di udara diubah menjadi nilai indeks (sub-indeks) menggunakan rumus *breakpoint* sebagai berikut [18]:

$$AQI_p = \frac{I_{High} - I_{Low}}{C_{High} - C_{Low}} * (C_p - C_{Low}) + I_{Low} \quad \text{.....(2.1)}$$

Dalam persamaan (2.1), $C_p$ merepresentasikan konsentrasi polutan yang terukur di lapangan. Variabel $C_{Low}$\$ dan $C_{High}$ adalah titik potong konsentrasi terendah dan tertinggi pada rentang breakpoint yang relevan, sementara $I_{Low}$ serta $I_{High}$ merupakan nilai indeks yang bersesuaian dengan titik-titik potong tersebut.

Setelah nilai indeks setiap polutan diperoleh, indeks kualitas udara akhir yang dilaporkan kepada publik ditentukan berdasarkan nilai tertinggi dari seluruh parameter yang diukur [21]:

$$AQI = max(AQI_{PM2.5}, AQI_{PM10}, AQI_{O3}, AQI_{SO2}, AQI_{CO}) \quad \text{.....(2.2)}$$

Penerapan prinsip nilai maksimum ini mencerminkan pendekatan manajemen risiko yang konservatif dan preventif. Nilai AQI final yang dilaporkan adalah nilai tertinggi dari seluruh sub-indeks yang dihitung. Kategori kualitas udara berdasarkan nilai AQI menurut US EPA ditunjukkan pada Tabel 2.1.

Tabel 2.1 Kategori Indeks Kualitas Udara (AQI) menurut US EPA [20], [21]

| Rentang AQI | Kategori | Warna | Tingkat Kekhawatiran Kesehatan |
|---|---|---|---|
| 0 - 50 | Baik | Hijau | Memuaskan |
| 51 - 100 | Sedang | Kuning | Dapat diterima |
| 101 - 150 | Tidak Sehat bagi Kelompok Sensitif | Oranye | Kelompok sensitif terdampak |
| 151 - 200 | Tidak Sehat | Merah | Semua orang mulai terdampak |
| 201 - 300 | Sangat Tidak Sehat | Ungu | Peringatan kesehatan |
| 301 - 500 | Berbahaya | Marun | Gawat darurat |

Di Indonesia, pemantauan kualitas udara ambien dilakukan oleh Kementerian Lingkungan Hidup dan Kehutanan (KLHK) melalui Stasiun Pemantau Kualitas Udara (SPKU) yang tersebar di berbagai kota [19]. Data dari SPKU diintegrasikan dan dipublikasikan melalui aplikasi dan situs *web* ISPU (Indeks Standar Pencemar Udara), yang metode perhitungannya mengacu pada Peraturan Menteri LHK No. 14 Tahun 2020 [6]. Selain data resmi pemerintah, platform global seperti *OpenAQ* berperan dalam menyediakan

informasi kualitas udara *real-time*. *OpenAQ* mengintegrasikan data dari stasiun pemantauan milik pemerintah, sensor swasta, dan mitra komunitas [6].

## 2.2 Data Mining dan Prediksi Time Series

*Data mining* merupakan inti dari *proses Knowledge Discovery in Databases* (KDD), yaitu serangkaian proses ekstraksi informasi potensial, implisit, tidak diketahui sebelumnya, dan bermanfaat dari suatu kumpulan data besar [22]. Proses KDD bersifat iteratif dan interaktif yang terdiri dari beberapa tahapan, dimulai dari seleksi data, pembersihan (*preprocessing*), transformasi, data mining (penambangan pola), hingga interpretasi dan evaluasi hasil. Data mining sendiri merujuk pada tahapan spesifik di mana algoritma cerdas diterapkan untuk menemukan pola-pola tersembunyi, asosiasi, atau struktur data.

Prediksi dalam *data mining* adalah proses estimasi nilai dari suatu variabel atau kejadian di masa depan berdasarkan pola historis dari data yang tersedia [23]. Tujuan utamanya adalah meminimalkan kesalahan antara nilai prediksi dengan nilai aktual. Berbeda dengan klasifikasi yang memprediksi label kategori, prediksi untuk variabel kontinu termasuk dalam regression task.

*Data time series* adalah serangkaian observasi terurut berdasarkan waktu, biasanya dengan interval yang sama. Untuk keperluan prediksi, *data time series* memiliki karakteristik utama yang harus dipahami, yaitu [24]:

1. Tren (*Trend*): pergerakan jangka panjang yang menunjukkan arah data, baik meningkat, menurun, atau stagnan.
2. Musiman (*Seasonality*): pola yang berulang secara periodik dalam interval waktu tetap, misalnya pola harian, mingguan, atau tahunan.
3. Siklus (*Cyclicity*): fluktuasi naik turun data yang tidak terikat pada kalender tetap dan umumnya berdurasi lebih panjang dari musiman.
4. Ketidakstasioneran: sifat data di mana mean, varians, atau kovarians berubah seiring waktu, yang seringkali perlu diatasi sebelum pemodelan.
5. Autokorelasi: korelasi antara nilai observasi pada waktu tertentu (*lag*) dengan nilai observasi pada waktu sebelumnya.

*Preprocessing* merupakan tahapan krusial dalam proyek prediksi *time series* untuk menjamin kualitas *input* data sebelum pemodelan. Tiga teknik utama yang umum diterapkan adalah [23]:

1. Penanganan Data Hilang *(Handling Missing Values)*

   *Data time series* seringkali memiliki nilai yang hilang karena gangguan sensor atau kesalahan pencatatan. Teknik penanganan yang umum adalah interpolasi linier, di mana nilai yang hilang diperkirakan berdasarkan nilai sebelum dan sesudahnya. Formula interpolasi linier untuk dua titik ($t_0$, $y_0$) dan ($t_1$, $y_1$) untuk mencari nilai pada t adalah:

$$y = y_0 + \frac{(y1 - y_0)}{(t1 - t_0)} \times (t - t_0) \quad \text{.....(2.3)}$$

2. Normalisasi Data *(Normalization)*

   Normalisasi dilakukan untuk mengubah skala data ke rentang tertentu, umumnya [0, 1], guna menghindari dominasi fitur dengan skala besar terhadap algoritma pembelajaran (terutama *neural network*). Metode *Min-max scaling* didefinisikan sebagai:

$$Xnorm = \frac{Xmax - Xmin}{X - Xmin} \quad \text{.....(2.4)}$$

3. Pembuatan Fitur Kelambanan *(Lag Features)*

   *Data time series* dimodelkan dengan asumsi bahwa nilai masa depan dipengaruhi oleh nilai masa lalunya. Oleh karena itu, dibuat fitur baru berupa nilai kelambanan (lag) dari variabel target. Lag ke-$t-1$ adalah nilai observasi satu langkah waktu sebelumnya. Sebagai contoh, untuk memprediksi nilai pada waktu $t$, fitur yang digunakan dapat mencakup $X_{t-1}, X_{t-2}, \ldots, X_{t-n}$. Hal ini mengubah data deret waktu menjadi format *supervised learning* seperti pada Tabel 2.2.

Tabel 2.2 Transformasi Data Time Series ke Format Supervised [12], [23]

| **Waktu (t)** | **Nilai Aktual ($X_t$)** | **Lag-1 ($X_{t-1}$)** | **Lag-2 ($X_{t-2}$)** | **Target** |
|---|---|---|---|---|
| 2024-01-01 01:00 | 45.2 | - | - | - |
| 2024-01-01 02:00 | 48.5 | 45.2 | - | - |
| 2024-01-01 03:00 | 52.1 | 48.5 | 45.2 | 52.1 |
| 2024-01-01 04:00 | 50.8 | 52.1 | 48.5 | 50.8 |
| 2024-01-01 01:00 | 45.2 | - | - | - |

## 2.3 Multiple linear regression (MLR)

*Multiple linear regression (MLR)* merupakan metode statistik parametrik yang digunakan untuk memodelkan hubungan linear antara satu variabel dependen (target) dengan dua atau lebih variabel independen (prediktor) [9]. Apabila hanya melibatkan satu variabel independen, metode tersebut disebut regresi linear sederhana. Sementara itu, *Multiple linear regression* (MLR) atau regresi linear berganda digunakan ketika terdapat dua

atau lebih variabel independen yang memengaruhi variabel dependen. Bentuk umum persamaan MLR dinyatakan sebagai berikut:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 + \cdots + +\beta_k X_k + \epsilon \quad \text{.....(2.5)}$$

Di mana $Y$ adalah variabel target $AQI_{t+1}$, $\beta_0$ merupakan konstanta (intercept), $\beta_k$ adalah koefisien regresi untuk setiap variabel independen $X_k$, dan $\epsilon$ merepresentasikan kesalahan residu. Tujuan utama MLR adalah mencari nilai koefisien regresi yang meminimalkan jumlah kuadrat error (*Sum of squared errors*/SSE) sehingga model dapat menghasilkan prediksi yang akurat [9].

Agar model MLR menghasilkan estimasi yang *Best linear unbiased estimator* (BLUE) dan valid secara statistik, beberapa asumsi klasik harus dipenuhi. Pelanggaran terhadap asumsi-asumsi ini dapat menyebabkan bias, ketidakkonsistenan, atau ketidakefisienan parameter estimasi [11].

1. Asumsi Linearitas

   Asumsi yang menyatakan bahwa hubungan antara variabel dependen dan seluruh variabel independen harus bersifat linear dalam parameter. Untuk mendeteksi pemenuhan asumsi ini, peneliti dapat melakukan analisis diagnostik melalui *plot residual versus predicted values*; jika plot tersebut menunjukkan distribusi titik yang acak dan tidak membentuk pola kurva tertentu, maka asumsi linearitas dapat dianggap terpenuhi. Selain itu, *partial regression plot* juga sering digunakan untuk meninjau korelasi linear parsial antara variabel independen spesifik dengan target.

2. Asumsi Independensi Residual *(Non-autokorelasi)*

   Asumsi yang menuntut agar nilai residu (*error*) pada suatu observasi tidak berkorelasi dengan nilai residu pada observasi lainnya. Dalam data *time series*, pemenuhan asumsi ini sangat menantang karena struktur data yang berurutan cenderung menyimpan korelasi serial.

   Untuk mendeteksi adanya pelanggaran asumsi ini, pengujian statistik seperti *Durbin-Watson Test* menjadi standar utama. Nilai statistik d yang dihasilkan berkisar antara 0 hingga 4, di mana nilai yang mendekati 2 mengindikasikan ketiadaan autokorelasi, sementara nilai $d < 1$ atau $d > 3$ menjadi sinyal kuat adanya autokorelasi yang perlu diwaspadai. Selain itu, penggunaan *Autocorrelation function* (ACF) *plot* pada residu dapat memberikan konfirmasi visual; jika tidak ditemukan *lag* yang signifikan melampaui batas kepercayaan, maka asumsi independensi dapat dianggap terpenuhi.

3. Asumsi Homoskedastisitas

Asumsi yang mensyaratkan bahwa varians dari residu harus bersifat konstan untuk seluruh rentang nilai prediksi, di mana ketidakpatuhan terhadap asumsi ini akan membentuk pola sistematis tertentu pada sebaran residu, seperti pola corong *(fan-shape)* yang mengindikasikan terjadinya *heteroskedastisitas*. Deteksi terhadap pelanggaran ini dilakukan melalui analisis visual pada plot residu versus nilai prediksi, atau melalui pengujian formal seperti *Breusch-Pagan Test* dan *Goldfeld-Quandt Test* yang dilakukan dengan membagi data menjadi dua kelompok guna membandingkan varians antar kelompok tersebut.

4. Asumsi Normalitas Residual

   Asumsi normalitas residual mensyaratkan bahwa residu model harus terdistribusi secara normal dengan nilai rata-rata (mean) sebesar nol. Deteksi terhadap pemenuhan asumsi ini dapat dilakukan secara visual melalui histogram residu yang harus membentuk pola kurva lonceng, atau melalui Q-Q Plot dengan memastikan titik-titik observasi mengikuti garis diagonal. Secara statistik, pengujian formal juga dapat dilakukan menggunakan *Shapiro-Wilk Test* atau *Kolmogorov-Smirnov Test*.

5. Asumsi Non-multikolinearitas

   Asumsi yang mensyaratkan tidak adanya korelasi linear yang sempurna atau tinggi antar variabel independen dalam model. Deteksi multikolinearitas dilakukan dengan menghitung *Variance inflation factor* (VIF) menggunakan rumus:

$$VIF_j = \frac{1}{1-R_j^2} \quad \text{.....(2.6)}$$

   di mana $R_j^2$ adalah koefisien determinasi ketika variabel $X_j$ diregresikan terhadap semua variabel independen lainnya. Nilai VIF yang melebihi 10 mengindikasikan adanya multikolinearitas serius, sementara nilai di atas 5 memerlukan perhatian lebih lanjut; selain itu, penggunaan matriks korelasi juga dapat mendeteksi masalah ini apabila ditemukan korelasi antar variabel yang melebihi 0,8.

Model *Multiple linear regression* (MLR) memiliki beberapa keunggulan utama yang relevan dengan penelitian ini. Pertama, metode ini bersifat sederhana dan mudah diinterpretasikan, di mana koefisien regresi memberikan informasi eksplisit mengenai arah dan besaran pengaruh setiap variabel, sehingga memudahkan pemahaman kontribusi masing-masing polutan terhadap prediksi AQI. Kedua, MLR sangat efisien secara komputasi karena estimasi parameter melalui *Ordinary least squares* (OLS) dapat diselesaikan dengan cepat bahkan untuk dataset berukuran besar, yang sangat mendukung pengembangan prototipe serta implementasi sistem *real-time*. Ketiga, metode ini memiliki dasar statistik

yang kuat dengan tersedianya berbagai uji validasi seperti uji t, F, dan nilai $R^2$, sehingga memungkinkan evaluasi ketat sebelum model diintegrasikan ke dalam *meta-ensemble*. Keempat, MLR memberikan *baseline* yang solid karena kerap digunakan sebagai model *benchmark* untuk membandingkan peningkatan performa dari *meta-ensemble*. Terakhir, model ini bersifat transparan karena hubungan antara *input* dan *output* dapat dilacak secara jelas, yang menjadi poin penting bagi sistem yang memerlukan akuntabilitas seperti model peringatan dini (*early warning model*).

Sebagaimana dibuktikan oleh penelitian Sharma dkk., MLR terbukti efektif sebagai model *baseline* untuk prediksi PM2.5 di wilayah dengan keterbatasan data di India, karena mampu memberikan interpretasi yang transparan mengenai faktor-faktor dominan yang memengaruhi kualitas udara setempat [25].

Model *Multiple linear regression* (MLR) memiliki beberapa keterbatasan struktural yang membatasi efektivitasnya dalam memprediksi kualitas udara yang bersifat dinamis. Keterbatasan utama terletak pada asumsi linearitas yang kaku, di mana MLR gagal menangkap hubungan non-linear—seperti efek sinergis antar polutan atau fenomena ambang batas (*threshold effects*)—yang sering ditemukan dalam data kualitas udara. Selain itu, sensitivitas metode *Ordinary least squares* (OLS) terhadap data pencilan (*outlier*) dapat mendistorsi model, terutama pada hari-hari dengan konsentrasi polusi ekstrem.

Model ini juga mengasumsikan independensi residual, padahal data *time series* umumnya memiliki autokorelasi serial yang melanggar asumsi tersebut dan menyebabkan bias pada *standard error*. Keterbatasan lain mencakup ketidakmampuan model untuk menangani interaksi kompleks antara variabel secara otomatis, sehingga peneliti harus menentukan interaksi tersebut secara manual, yang sering kali sulit dilakukan mengingat kompleksitas hubungan antara polutan dan faktor meteoro*log*is. Selain itu, potensi *overfitting* dapat terjadi jika banyak variabel ditambahkan tanpa seleksi yang cermat, sementara banyaknya asumsi statistik yang ketat sering kali tidak terpenuhi oleh data riil di lapangan.

Sebagaimana temuan Lei dkk. dalam penelitian prediksi kualitas udara di Macao, MLR memiliki keterbatasan dalam memprediksi fluktuasi polutan yang kompleks, terutama pada kondisi ekstrem. Hal inilah yang mendasari perlunya penggunaan pendekatan *machine learning* yang lebih fleksibel untuk meningkatkan akurasi model [12].

## 2.4 Random forest regressor (RFR)

*Random forest regressor* (RFR) adalah algoritma *machine learning* berbasis *ensemble learning* yang menggabungkan banyak pohon keputusan *(decision trees)* untuk menghasilkan prediksi regresi yang lebih stabil dan akurat dibandingkan pohon keputusan tunggal. RFR termasuk dalam metode *bagging (bootstrap aggregating)*, di mana setiap pohon dilatih menggunakan sampel data yang diambil secara acak dengan pengembalian *(bootstrap sample)* dan pada setiap pemecahan simpul *(splitting)* hanya mempertimbangkan subset fitur secara acak. Dalam konteks penelitian ini, RFR digunakan untuk memprediksi nilai AQI hari berikutnya serta berperan sebagai *meta-model* pada arsitektur *stacking*. RFR dipilih karena kemampuannya menangani hubungan non-linear antara variabel prediktor dan target, ketahanannya terhadap data pencilan (*outlier*), serta tidak memerlukan normalisasi data. Sub-bab ini akan menguraikan konsep dasar *decision tree* dan *ensemble learning*, cara kerja *Random forest* untuk regresi, kelebihan RFR dalam menangani data non-linear dan high dimensional, serta kelemahan RFR yang perlu diperhatikan.

*Decision tree* (pohon keputusan) adalah algoritma pembelajaran terawasi yang bekerja dengan cara membagi data secara rekursif (*recursive partitioning)* berdasarkan fitur tertentu hingga mencapai simpul daun *(leaf node)* yang berisi nilai prediksi. Untuk kasus regresi, nilai prediksi pada simpul daun umumnya adalah rata-rata dari seluruh sampel yang mencapai simpul tersebut. Meskipun *interpretable*, pohon keputusan tunggal cenderung memiliki varians tinggi *(high variance)*, di mana perubahan kecil pada data dapat mengubah struktur pohon secara drastis. Untuk mengatasi kelemahan ini, diperkenalkan konsep *ensemble learning*, yaitu teknik yang menggabungkan banyak model untuk menghasilkan satu model dengan performa prediksi yang lebih stabil dan akurat [13]. *Random forest* termasuk dalam *metode ensemble* berbasis *bagging (bootstrap aggregating)*.

*Random forest* untuk regresi bekerja dengan membangun kumpulan *(forest)* dari banyak pohon keputusan *(decision trees)* dan menggabungkan prediksi dari seluruh pohon tersebut. Terdapat dua mekanisme utama yang membuat *Random forest* unggul [14]:

1. *Bagging (Bootstrap aggregating)*: Setiap pohon dalam *Random forest* dilatih menggunakan sampel data yang diambil secara acak dengan pengembalian *(bootstrap sample)* dari dataset asli. Akibatnya, setiap pohon melihat data yang sedikit berbeda, sehingga mengurangi korelasi antar pohon. Sekitar sepertiga data tidak terpilih dalam sampel *bootstrap*, yang disebut *Out-of-Bag* (OOB) data, yang dapat digunakan untuk validasi internal.

2. *Random feature selection*: Pada setiap proses pemecahan simpul *(splitting)* dalam pohon, tidak semua fitur dipertimbangkan sebagai kandidat pemisah. Algoritma hanya memilih subset fitur secara acak (sebanyak mtry) dari total fitur yang ada. Hal ini memastikan bahwa pohon-pohon yang dihasilkan tidak terlalu mirip satu sama lain dan memberikan kesempatan bagi fitur-fitur yang mungkin lemah untuk berkontribusi.

Untuk prediksi regresi, nilai akhir dari *Random forest* adalah rata-rata *(averaging)* dari seluruh prediksi pohon penyusunnya, yang secara matematis dapat dituliskan sebagai:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B}\widehat{y_b} \quad \text{.....(2.7)}$$

Dimana:

$\hat{y}$ = Prediksi akhir *Random forest*

$y_b$ = Prediksi dari pohon (tree) ke-$b$

$B$ = Jumlah total pohon dalam forest

*Random forest regressor* memiliki sejumlah kelebihan yang membuatnya sangat efektif untuk pemodelan kualitas udara [10]:

1. Pertama, RFR mampu menangani hubungan non-linear antara fitur prediktor dan target tanpa memerlukan transformasi data yang kompleks, berbeda dengan regresi linear yang terbatas pada asumsi linearitas.
2. Kedua, RFR robust terhadap data high dimensional (jumlah fitur banyak) karena proses *random feature selection* secara implisit melakukan seleksi fitur, sehingga hanya fitur penting yang dominan berpengaruh.
3. Ketiga, algoritma ini tidak terlalu sensitif terhadap skala data (tidak memerlukan normalisasi ketat) dan relatif robust terhadap *outliers* karena menggunakan rata-rata dari banyak pohon.
4. Keempat, RFR menyediakan fitur *feature importance* yang dapat digunakan untuk mengukur kontribusi setiap variabel prediktor terhadap akurasi model.

Meskipun unggul dalam banyak aspek, *Random forest regressor* tetap memiliki beberapa kelemahan [13]:

1. Pertama, model ini cenderung memiliki variansi tinggi pada rentang data di luar nilai pelatihan (*extrapolation*).

   Karena setiap pohon hanya membuat prediksi berdasarkan rata-rata nilai pada simpul daun, RFR tidak dapat memprediksi nilai di luar rentang data latih.

2. Kedua, meskipun dirancang untuk mengurangi *overfitting*, pada data dengan noise tinggi, RFR dapat mengalami *overfitting* jika jumlah pohon terlalu banyak atau kedalaman pohon *(depth)* tidak dibatasi.
3. Ketiga, model yang dihasilkan memiliki interpretabilitas rendah dibandingkan regresi linear tunggal, karena sulit untuk memahami hubungan sebab-akibat dari ratusan pohon yang terbentuk.
4. Keempat, RFR membutuhkan sumber daya komputasi yang lebih besar (memori dan waktu proses) terutama ketika jumlah pohon ($n\_estimators$) dan kedalaman pohon ditetapkan tinggi.

Tabel 2.3 merangkum perbandingan karakteristik antara MLR dan RFR berdasarkan berbagai literatur.

Tabel 2.3 Perbandingan Karakteristik MLR dan RFR [9], [10], [11], [13]

| **Karakteristik** | ***Multiple linear regression* (MLR)** | ***Random forest regressor* (RFR)** |
|---|---|---|
| Hubungan Data | Linear | Non-linear |
| Interpretabilitas | Tinggi (koefisien jelas) | Rendah (*black box*) |
| Skala Data | Perlu normalisasi | Tidak perlu normalisasi |
| *Overfitting* | Rentan jika multikolinearitas | Cukup robust, namun bisa terjadi |
| Kemampuan Extrapolasi | Baik (dalam batas linear) | Buruk (tidak mampu) |

Berdasarkan Tabel 2.3, MLR unggul dalam hal interpretabilitas dan kemampuan extrapolasi, namun terbatas pada hubungan linear dan rentan terhadap multikolinearitas. Sebaliknya, RFR mampu menangani hubungan non-linear tanpa perlu normalisasi data, tetapi bersifat *black box* dan tidak mampu melakukan extrapolasi di luar rentang data.

## 2.5 Meta-Ensemble

*Meta-ensemble,* yang juga dikenal sebagai *stacking (stacked generalization)*, merupakan teknik *ensemble learning* lanjutan yang menggabungkan prediksi dari beberapa model dasar (*base models*) menggunakan model pembelajaran tingkat tinggi yang disebut *meta-model*. Berbeda dengan metode ensemble konvensional seperti *bagging (Random forest)* yang menggabungkan prediksi dengan mekanisme statis seperti rata-rata, *stacking* mempelajari cara terbaik untuk mengkombinasikan prediksi dari berbagai *base model* secara adaptif berdasarkan data. Dalam konteks penelitian ini, *meta-ensemble*

digunakan untuk menggabungkan kelebihan *Multiple linear regression* (MLR) yang unggul dalam interpretabilitas dan menangkap hubungan linear, dengan *Random forest regressor* (RFR) yang unggul dalam menangkap pola non-linear. Sub-bab ini akan menguraikan arsitektur *meta-ensemble* dua tingkat, konsep koreksi kesalahan dalam *meta-ensemble*, perbedaan *meta-ensemble* dengan ensemble konvensional, serta keunggulan *meta-ensemble* untuk prediksi data kompleks seperti kualitas udara.

*Stacking* (singkatan dari *stacked generalization*) merupakan teknik meta-ensemble yang menggabungkan prediksi dari beberapa model dasar (*base models*) menggunakan model pembelajaran tingkat tinggi yang disebut *meta-model* [14]. Arsitektur *stacking* secara umum terdiri dari dua lapisan. Lapisan pertama (Level-0) berisi kumpulan *base models* yang dilatih secara independen menggunakan data pelatihan asli. Lapisan kedua (Level-1) berisi *meta-model* yang dilatih menggunakan output (prediksi) dari *base models* sebagai fitur input, sementara target asli tetap menjadi variabel dependen [15], [16]. Dengan kata lain, *meta-model* mempelajari bagaimana cara terbaik mengkombinasikan prediksi dari berbagai *base models* untuk menghasilkan prediksi akhir yang lebih akurat. Secara formal, jika terdapat M*M base model*, maka untuk setiap *instance* data, vektor fitur baru P*P* dibentuk sebagai prediksi dari seluruh *base models*, kemudian *meta-model* mempelajari fungsi $f: P \rightarrow y$ untuk meminimalkan *error* prediksi [26].

Secara formal, jika terdapat *M base model*, maka untuk setiap instance data, vektor fitur baru *P* dibentuk sebagai:

$$P_i = [\hat{y}_{i1,}\ \hat{y}_{i2,}\ \dots, \hat{y}_{iM,}] \quad \text{.....(2.8)}$$

di mana$\hat{y}_{ij,}$ adalah prediksi dari *base model* ke-*j* untuk instance ke-*i*. *Meta-model* kemudian mempelajari fungsi $f: P \rightarrow y$ untuk meminimalkan error prediksi.

Keunggulan utama *stacking* terletak pada kemampuannya melakukan koreksi kesalahan (*error correction*) secara sistematis [26]. Setiap *base model* memiliki kelemahan dan bias yang berbeda-beda dalam mempelajari pola data. Ketika *base model* membuat kesalahan prediksi, kesalahan tersebut cenderung tidak independen satu sama lain. *Meta-model* berperan sebagai *referee* yang mempelajari pola kesalahan dari *base models*. Dengan menganalisis residu atau deviasi prediksi *base models* terhadap nilai aktual, *meta-model* dapat mengidentifikasi kondisi di mana *base model* tertentu cenderung overestimate atau underestimate. *Meta-model* kemudian belajar memberikan bobot yang sesuai atau mengkoreksi output *base models* untuk menghasilkan prediksi yang lebih mendekati nilai

sebenarnya. Proses ini pada dasarnya adalah pembelajaran tingkat tinggi (*meta-learning*) tentang kelemahan dan kelebihan masing-masing model dasar.

*Stacking* memiliki perbedaan fundamental dengan metode ensemble konvensional seperti *bagging* (*Random forest*) dan *boosting* (AdaBoost, *Gradient boosting*). Perbedaan utama tersebut dirangkum pada Tabel 2.4.

Tabel 2.4 Perbandingan Ensemble Konvensional dengan *Stacking* [10], [13], [15], [16]

| **Karakteristik** | ***Bagging* / *Boosting*** | ***Stacking* (Meta-Ensemble)** |
|---|---|---|
| Arsitektur | Satu level (model homogen atau terintegrasi) | Dua level (*base models* + *meta-model*) |
| Model Dasar | Umumnya menggunakan algoritma yang sama (homogen) | Menggunakan algoritma yang berbeda (heterogen) |
| Fungsi Penggabung | Rata-rata (*averaging*) atau voting terbobot sederhana | Model pembelajaran (*meta-model*) yang mempelajari cara menggabung |
| Kompleksitas | Relatif lebih sederhana | Lebih kompleks karena memerlukan pelatihan dua tahap |
| Tujuan | Mengurangi varians (*bagging*) atau bias (*boosting*) | Mengkombinasikan kelebihan berbagai model secara adaptif |

Pada *bagging*, penggabungan dilakukan dengan mekanisme statis (rata-rata). Pada *stacking*, penggabungan bersifat dinamis dan dipelajari dari data, sehingga mampu menangkap interaksi yang lebih kompleks antar prediksi *base models* [14].

Penerapan meta-ensemble seperti *stacking* menawarkan sejumlah keunggulan signifikan untuk prediksi data kompleks seperti kualitas udara [25]:

1. Pertama, *stacking* mampu mengkombinasikan kekuatan berbagai algoritma.
   MLR unggul dalam interpretabilitas dan menangkap tren linear, sementara RFR unggul dalam menangkap pola non-linear. *Stacking* memungkinkan kedua kelebihan ini bekerja secara sinergis.
2. Kedua, *stacking* secara adaptif mempelajari bobot optimal untuk setiap *base model* berdasarkan konteks data, tidak hanya berdasarkan rata-rata sederhana.
3. Ketiga, pendekatan ini secara signifikan dapat meningkatkan akurasi prediksi karena meminimalkan bias dan varians secara simultan melalui proses koreksi kesalahan.
4. Keempat, *stacking* relatif robust terhadap kelemahan individual *base model*; jika satu model memberikan prediksi buruk pada kondisi tertentu, *meta-model* dapat belajar untuk mengurangi kontribusinya.

Hal ini menjadikan *stacking* sebagai pendekatan yang sangat menjanjikan untuk memodelkan dinamika polusi udara yang bersifat kompleks, non-linear, dan dipengaruhi oleh berbagai faktor meteoro*log*is.

## 2.6 Evaluasi Kinerja Model Regresi

Evaluasi kinerja model regresi bertujuan untuk mengukur seberapa akurat model dalam memprediksi nilai kontinu, khususnya dalam konteks prediksi Indeks Kualitas Udara (AQI). Metrik evaluasi yang umum digunakan meliputi MAE, RMSE, dan $R^2$, yang masing-masing memberikan perspektif berbeda tentang kualitas prediksi [27].

*Mean absolute error* (MAE) mengukur rata-rata selisih absolut antara nilai prediksi dan nilai aktual. Metrik ini menghitung besaran error tanpa mempertimbangkan arahnya (overestimate atau underestimate). MAE didefinisikan dengan formula [28]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{.....(2.9)}$$

Dimana:

$n$ = jumlah sampel data

$y_i$ = nilai aktual ke-i*i*

$\hat{y}_i$ = nilai prediksi ke-i*i*

MAE memiliki satuan yang sama dengan variabel target (dalam konteks ini, satuan AQI). Keunggulan MAE adalah interpretasinya yang intuitif karena merepresentasikan rata-rata kesalahan absolut dalam skala asli. MAE bersifat robust terhadap *outlier* karena tidak mengkuadratkan error [28].

*Root mean square error* (RMSE) mengukur akar kuadrat dari rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual. RMSE memberikan bobot lebih besar pada error yang besar karena proses kuadratisasi sebelum dirata-rata. Formula RMSE adalah sebagai berikut [29]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{.....(2.10)}$$

Seperti MAE, RMSE juga memiliki satuan yang sama dengan variabel target. RMSE lebih sensitif terhadap *outlier* dibandingkan MAE, sehingga jika terdapat perbedaan signifikan antara MAE dan RMSE, hal ini mengindikasikan adanya varians error yang besar atau keberadaan *outlier* dalam data. RMSE sangat berguna ketika error besar tidak diinginkan dalam Model Prediksi [29].

Koefisien determinasi ($R^2$) mengukur proporsi variansi dalam variabel dependen yang dapat dijelaskan oleh variabel independen dalam model. Nilai $R^2$ berkisar antara 0

hingga 1, meskipun dalam praktiknya dapat bernilai negatif jika model lebih buruk daripada rata-rata sederhana. $R^2$ didefinisikan sebagai [30]:

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Tot}} \quad \text{.....(2.11)}$$

Dimana:

$SS_{Res}$=$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ = jumlah kuadrat residual (error)

$SS_{Tot}$ $\sum_{i=1}^{n}(y_i - \bar{y})^2$ = jumlah kuadrat total (selisih nilai aktual dengan rata-rata $\bar{y}$)

Interpretasi $R^2$ adalah sebagai berikut [30]:

1. $R^2 = 1$: Model sempurna, semua variansi data dapat dijelaskan.
2. $R^2$ mendekati 1: Model sangat baik, variansi error sangat kecil.
3. $R^2 = 0$: Model tidak lebih baik dari rata-rata data.
4. $R^2 < 0$: Model lebih buruk dari prediksi menggunakan rata-rata.

Interpretasi metrik evaluasi harus dilakukan secara kontekstual dengan mempertimbangkan skala dan kategori AQI. Tabel 2.5 menyajikan panduan interpretasi hasil evaluasi untuk prediksi AQI berdasarkan rentang nilai MAE, RMSE, dan $R^2$.

Tabel 2. 5 Interpretasi Metrik Evaluasi dalam Konteks Prediksi AQI [29], [30], [31]

| Metrik | Rentang Nilai | Interpretasi dalam Prediksi AQI |
|---|---|---|
| MAE / RMSE | < 10 | Sangat baik; kesalahan prediksi kurang dari satu tingkat kategori AQI (misal: "Sedang" ke "Tidak Sehat") |
| | 10 - 25 | Cukup baik; kesalahan masih dalam batas toleransi untuk peringatan dini |
| | > 25 | Kurang baik; berisiko salah klasifikasi kategori kesehatan masyarakat |
| $R^2$ | > 0.85 | Sangat baik; model mampu menjelaskan hampir seluruh variabilitas data AQI |
| | 0.70 - 0.85 | Baik; model cukup akurat untuk keperluan prediksi operasional |
| | 0.50 - 0.70 | Cukup; dapat digunakan dengan hati-hati untuk estimasi kasar |
| | < 0.50 | Kurang; model tidak mampu menangkap pola AQI dengan memadai |

Perbedaan signifikan antara MAE dan RMSE (RMSE jauh lebih besar) mengindikasikan adanya prediksi dengan error sangat besar yang perlu diinvestigasi lebih lanjut.

## 2.7 Sistem Informasi dan Visualisasi Prediksi

Sistem informasi dan visualisasi prediksi merupakan komponen pendukung dalam menjembatani hasil analisis model *machine learning* dengan kebutuhan pengguna akhir, khususnya dalam konteks model peringatan dini kualitas udara. Model prediksi AQI yang akurat akan lebih bermanfaat jika hasilnya disajikan kepada masyarakat dan pemangku kebijakan dalam bentuk yang mudah dipahami dan cepat diakses. Oleh karena itu, diperlukan sebuah sistem informasi berbasis *web* yang dapat menampilkan hasil prediksi AQI secara visual, informatif, dan intuitif. Sub-bab ini akan menguraikan konsep sistem informasi berbasis *web*, arsitektur *website* untuk visualisasi data prediksi, peran antarmuka pengguna dalam model peringatan dini, serta teknologi pendukung yang digunakan dalam pengembangan *prototipe website (front-end, back-end*, dan *API)*.

Sistem informasi berbasis *web* merupakan suatu sistem yang dirancang untuk mengelola, menyimpan, memproses, dan menyajikan informasi melalui platform *web* (intranet atau internet) dengan menggunakan arsitektur client-server [31]. Pengguna dapat mengakses sistem ini melalui peramban *web* (*web browser*) tanpa memerlukan instalasi perangkat lunak khusus di sisi pengguna. Karakteristik utama sistem berbasis *web* meliputi aksesibilitas lintas platform, kemudahan distribusi informasi secara *real-time*, dan kemampuan untuk diintegrasikan dengan berbagai sumber data eksternal. Dalam konteks prediksi kualitas udara, sistem ini berfungsi sebagai media implementasi yang menjembatani model prediksi kompleks dengan pengguna akhir, menyajikan hasil analisis dalam format yang informatif dan mudah dipahami.

Arsitektur prototipe *website* untuk visualisasi data prediksi umumnya mengadopsi pola arsitektur tiga lapis *(three-tier architecture)* yang terdiri dari lapisan presentasi, lapisan aplikasi, dan lapisan data [7]. Lapisan data bertanggung jawab menyimpan data historis, hasil prediksi model, dan *metadata* lainnya, biasanya menggunakan sistem manajemen basis data seperti MySQL atau PostgreSQL. Lapisan aplikasi (*backend*) berisi logika bisnis, termasuk skrip untuk menjalankan model prediksi, melakukan pra-pemrosesan data, dan menyediakan layanan API. Lapisan presentasi (*frontend*) menampilkan data kepada pengguna dalam bentuk visualisasi interaktif seperti grafik deret waktu dan indikator AQI.

Antarmuka pengguna *(user interface/UI)* memegang peranan penting dalam efektivitas penyampaian informasi hasil prediksi [7]. UI yang baik tidak hanya menampilkan data, tetapi juga mengomunikasikan tingkat risiko secara intuitif sehingga pengguna dapat mengambil tindakan cepat dan tepat. Desain yang berpusat pada pengguna *(user-centered*

*design)* memastikan bahwa informasi penting tidak tenggelam dalam kompleksitas data, sehingga fungsi peringatan dini dapat berjalan optimal.

Pada sisi *front-end*, teknologi seperti HTML, CSS, dan *JavaScript* digunakan untuk membangun struktur, tampilan, dan interaktivitas halaman *web*. *Library* visualisasi seperti Chart.js, D3.js, atau ECharts dapat diintegrasikan untuk membuat grafik yang dinamis dan responsif. Pada sisi *back-end*, bahasa pemrograman seperti *Python* dengan *framework Django* atau Node.js umum digunakan untuk menangani logika *server*, menjalankan model prediksi, dan mengelola sesi pengguna. *Application Programming Interface* (API) berperan sebagai penghubung antara *front-end* dan *back-end*, memungkinkan pertukaran data secara terstruktur, biasanya dalam format JSON. API juga dapat digunakan untuk mengakuisisi data *real-time* dari sumber eksternal seperti OpenWeatherMap atau *IQAir*. Integrasi teknologi ini memungkinkan terciptanya sistem yang modular, skalabel, dan mudah dipelihara sebagai media implementasi model prediksi.

## 2.8 Penelitian Terkait (*State of the art*)

Berbagai penelitian telah dilakukan dalam bidang prediksi kualitas udara menggunakan pendekatan *machine learning*, mulai dari regresi linear sederhana hingga meta-ensemble. Tabel 2.6 merangkum penelitian-penelitian terkait yang menjadi acuan dalam studi ini, meliputi metode yang digunakan, hasil utama, serta keterbatasan masing-masing penelitian.

Tabel 2. 6 Rangkuman Penelitian Terkait Prediksi Kualitas Udara

| **Penulis (Tahun)** | **Metode** | **Hasil Utama** | **Keterbatasan / Gap** |
|---|---|---|---|
| Qonitan et al. (2025) | MLR | $R^2 = 0{,}522$ | Akurasi masih perlu ditingkatkan [33] |
| Dutta & Jinsart (2021) | MLR + *lag* | MAE = 21,97–27,35 | Terbatas pada data lokal [32] |
| Mohapatra et al. (2020) | MLR vs ANN | ANN lebih akurat, MLR lebih interpretabel | Berhenti pada evaluasi statistik [34] |
| Alsayadi et al. (2022) | RFR | $R^2 = 0{,}7882$ | Tidak ada implementasi sistem [35] |
| Ovaliadis et al. (2019) | RFR vs *deep learning* | RFR lebih cepat, akurasi sebanding | Data terbatas (2 bulan) [36] |
| Kesten & Panhwar (2026) | RFR vs OLS | RFR unggul (RMSE=7,03 vs 25,68) | Tidak ada implementasi sistem [37] |
| Tian et al. (2024) | *Stacking* (LSTM+*Xgboost*→*Lightgbm*) | MAE = 3,9142 (lebih baik dari RFR) | *Base models* kompleks, tidak diimplementasikan ke sistem [15] |
| Morshed-Bozorgdel et al. (2022) | *Stacking* (9 algoritma) | $R^2 = 0{,}89$, akurasi +43% | *Base models* terlalu banyak, tidak praktis untuk sistem *real-time* [16] |
| Janani et al. (2026) | *Dynamic ensemble* (SG-SKRDX) | Akurasi rekomendasi 93,41% | Domain pertanian, bukan kualitas udara [38] |

Berdasarkan kajian terhadap penelitian-penelitian terdahulu, dapat diidentifikasi beberapa celah penelitian (*research gap*) yang menjadi fondasi bagi penelitian ini:

1. Keterbatasan eksplorasi *meta-ensemble* dengan *base models heterogen* dalam prediksi kualitas udara

   Penelitian oleh Tian et al. (2024) dan Morshed-Bozorgdel et al. (2022) telah menunjukkan efektivitas pendekatan *meta-ensemble* (*stacking*) dalam prediksi kualitas udara, namun kedua studi tersebut mengkombinasikan *base models* yang kompleks seperti LSTM, *Xgboost*, ANN, dan SVM [15], [16]. Sementara itu, penelitian yang secara khusus mengeksplorasi kombinasi *base models heterogen* yang menggabungkan pendekatan linear (MLR) dan non-linear (RFR) dalam arsitektur *stacking* dua tingkat masih terbatas, terutama dalam konteks prediksi AQI di perkotaan Indonesia. Padahal, kombinasi MLR dan RFR menawarkan sinergi antara interpretabilitas linear dan fleksibilitas non-linear yang potensial untuk menangkap kompleksitas data kualitas udara.

2. Minimnya studi komparatif sistematis antara MLR, RFR, dan *stacking* pada domain spasial yang sama.

   Penelitian oleh Alsayadi et al. (2022) dan Ovaliadis et al. (2019) telah mendemonstrasikan keunggulan RFR dibandingkan model linear [32], [33], sementara Kesten dan Panhwar (2026) membandingkan RFR dengan OLS [34]. Namun, studi komparatif yang secara sistematis membandingkan performa MLR (mewakili pendekatan linear dan interpretabel), RFR (mewakili ensemble konvensional non-linear), dan *stacking* (mewakili meta-ensemble) pada dataset yang sama masih sangat terbatas, khususnya dalam konteks prediksi AQI di Kota Medan. Padahal, komparasi komprehensif ketiga pendekatan ini penting untuk mengidentifikasi peningkatan akurasi secara kuantitatif yang dihasilkan oleh arsitektur meta-ensemble.

3. Keterbatasan implementasi model prediksi ke dalam prototipe sistem informasi berbasis *web*.

   Penelitian-penelitian terdahulu, seperti yang dilakukan oleh Qonitan et al. (2025), Dutta dan Jinsart (2021), serta Mohapatra et al. (2020), umumnya berhenti pada tahap evaluasi model secara statistik tanpa mengintegrasikan hasil prediksi ke dalam sistem informasi yang dapat diakses publik [35], [36], [37]. Sementara itu, Janani et al. (2026) telah mengintegrasikan model prediksi dengan sistem rekomendasi, namun dalam domain pertanian [38]. Integrasi hasil prediksi kualitas udara ke dalam prototipe sistem informasi berbasis *web* yang berfungsi sebagai model peringatan dini masih jarang diimplementasikan, padahal hal ini krusial untuk menjembatani hasil prediksi dengan kebutuhan masyarakat akan informasi yang akurat dan tepat waktu.

Dengan demikian, penelitian ini mengisi celah tersebut dengan mengembangkan arsitektur *stacking* dua tingkat dengan *base models heterogen* (MLR dan RFR) yang secara spesifik menggabungkan keunggulan interpretabilitas linear dan fleksibilitas non-linear untuk prediksi AQI, melakukan evaluasi komprehensif yang membandingkan performa tiga pendekatan prediksi (MLR, RFR, dan *Stacking*) pada dataset AQI Kota Medan periode 2023-2025, serta mengimplementasikan model terbaik ke dalam prototipe *website* visualisasi prediksi sebagai media implementasi model peringatan dini kualitas udara.

Kebaruan (*novelty*) penelitian ini terletak pada penerapan arsitektur *stacking* dengan *base models heterogen* (MLR dan RFR) yang menggabungkan keunggulan interpretabilitas linear dan fleksibilitas non-linear, sebuah pendekatan yang belum banyak dieksplorasi dalam studi prediksi kualitas udara sebelumnya; evaluasi

komprehensif ketiga pendekatan (MLR, RFR, dan *Stacking*) pada konteks spasial Kota Medan dengan menggunakan data periode 2023-2025 yang memberikan kontribusi signifikan pada literatur prediksi kualitas udara di perkotaan Indonesia; serta pengembangan prototipe sistem informasi berbasis *web* yang menjembatani hasil prediksi dengan kebutuhan masyarakat sebagai media implementasi model.