# BAB 2
# LANDASAN TEORI

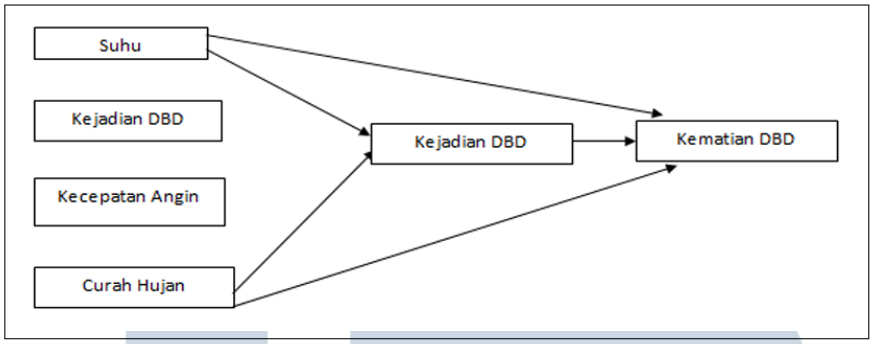
## 2.1 Demam berdarah dengue dan faktor iklim

Demam Berdarah Dengue (DBD) adalah penyakit tular vektor yang disebabkan oleh virus dengue dan ditularkan melalui nyamuk *Aedes aegypti* serta *Aedes albopictus* [2]. Dinamika kasus DBD berkaitan dengan kondisi lingkungan karena suhu, kelembapan, curah hujan, dan penyinaran matahari memengaruhi siklus hidup nyamuk, ketersediaan tempat berkembang biak, serta masa inkubasi virus [3, 11].

DBD menjadi relevan untuk dikaji menggunakan pendekatan prediksi karena jumlah kasusnya dapat berubah dari waktu ke waktu. Perubahan tersebut tidak hanya dipengaruhi oleh faktor medis, tetapi juga oleh kondisi lingkungan yang memengaruhi populasi nyamuk dan peluang penularan. Ketika kondisi lingkungan mendukung perkembangan nyamuk, potensi peningkatan kasus pada periode berikutnya juga dapat meningkat. Oleh karena itu, data iklim sering digunakan sebagai variabel pendukung dalam penelitian prediksi DBD.

Suhu udara berperan dalam memengaruhi perkembangan nyamuk dan virus dengue. Suhu yang sesuai dapat mempercepat siklus hidup nyamuk, mulai dari telur, larva, pupa, hingga nyamuk dewasa. Kelembapan udara berkaitan dengan kemampuan nyamuk bertahan hidup karena kondisi yang terlalu kering dapat menurunkan peluang hidup nyamuk dewasa. Curah hujan dapat meningkatkan jumlah tempat perkembangbiakan nyamuk melalui terbentuknya genangan air, meskipun curah hujan yang terlalu ekstrem juga dapat menghilangkan sebagian tempat berkembang biak. Durasi penyinaran matahari dan kecepatan angin juga dapat memengaruhi aktivitas lingkungan serta kondisi mikrohabitat nyamuk.

Hubungan umum antara faktor meteorologi dan dinamika penularan DBD ditunjukkan pada Gambar 2.1.




Gambar 2.1. Pengaruh faktor meteorologi terhadap dinamika penularan DBD

Sumber: [12]

Pengaruh iklim terhadap kasus DBD tidak terjadi secara langsung pada hari yang sama. Curah hujan dapat menciptakan genangan air, suhu dan kelembapan memengaruhi perkembangan nyamuk, sedangkan gejala klinis baru tercatat setelah masa inkubasi. Selain itu, terdapat jeda antara gigitan nyamuk, munculnya gejala, kunjungan pasien ke fasilitas kesehatan, pencatatan kasus, dan pelaporan data. Oleh karena itu, penelitian prediksi DBD perlu mempertimbangkan efek tunda atau *lag effect*.

Pada penelitian ini, faktor iklim tidak diperlakukan sebagai penyebab tunggal kasus DBD, tetapi sebagai variabel pendukung yang dapat membantu model memahami perubahan kondisi lingkungan. Hal ini penting karena dinamika DBD juga dapat dipengaruhi oleh faktor lain seperti kepadatan penduduk, mobilitas, sanitasi, indeks vektor, dan intervensi kesehatan masyarakat. Dengan demikian, penggunaan variabel iklim dalam penelitian ini dibatasi sebagai bagian dari informasi historis yang digunakan untuk memprediksi jumlah kasus DBD mingguan.

## 2.2 Data deret waktu

Data deret waktu (*time series*) adalah data yang urutan waktunya memiliki makna. Pada data seperti ini, nilai pada waktu ke-$t$ dapat dipengaruhi oleh nilai pada waktu sebelumnya, misalnya $t-1$, $t-2$, dan seterusnya [9, 13]. Jumlah kasus DBD mingguan termasuk data deret waktu karena pola minggu sebelumnya sering menjadi informasi penting untuk memprediksi minggu berikutnya.

Karakter utama data deret waktu adalah adanya ketergantungan antarperiode. Pada data biasa yang tidak mempertimbangkan waktu, setiap baris sering diasumsikan berdiri sendiri. Sebaliknya, pada deret waktu, observasi

pada satu periode dapat berkaitan dengan observasi sebelum dan sesudahnya. Dalam konteks DBD, jumlah kasus pada suatu minggu dapat dipengaruhi oleh jumlah kasus minggu sebelumnya, kondisi iklim beberapa minggu sebelumnya, serta pola musiman dalam satu tahun.

Data deret waktu juga dapat memiliki pola khusus seperti tren, musiman, autokorelasi, dan fluktuasi acak. Tren menunjukkan kecenderungan naik atau turun dalam jangka waktu tertentu. Musiman menunjukkan pola yang berulang pada periode tertentu, misalnya pola tahunan yang berkaitan dengan musim hujan atau perubahan kondisi iklim. Autokorelasi menunjukkan keterkaitan nilai saat ini dengan nilai masa lalu. Fluktuasi acak merupakan perubahan yang tidak sepenuhnya dapat dijelaskan oleh pola historis atau variabel yang tersedia.

Konsekuensi utama dari karakteristik deret waktu adalah proses validasi tidak boleh dilakukan secara acak. Pembagian acak dapat menempatkan data masa depan di data latih dan data masa lalu di data uji, sehingga menimbulkan *data leakage*. Oleh sebab itu, penelitian ini menggunakan pembagian kronologis: bagian awal deret waktu digunakan sebagai data latih dan bagian akhir digunakan sebagai data uji.

Dalam konteks prediksi DBD, bentuk data deret waktu juga memengaruhi cara fitur disusun. Data mentah dapat berbentuk tabular atau matriks, tetapi dapat diperlakukan sebagai deret waktu apabila terdapat atribut waktu dan analisisnya mempertahankan urutan kronologis. Pada penelitian ini, data mentah harian diagregasi menjadi mingguan agar lebih sesuai dengan dinamika pelaporan kesehatan masyarakat dan untuk mengurangi fluktuasi harian yang sangat bising. Setelah agregasi, setiap baris data merepresentasikan satu minggu, bukan lagi observasi harian yang berdiri sendiri.

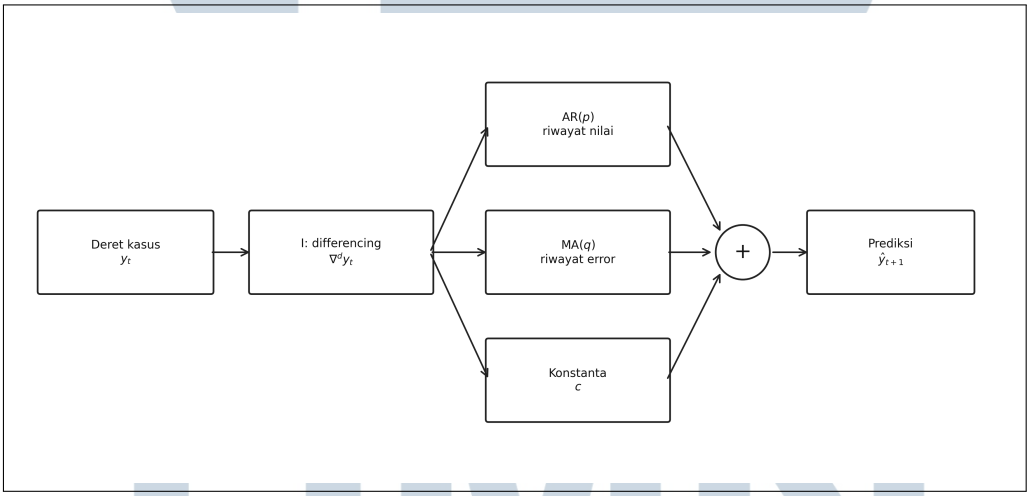
Agregasi mingguan dipilih karena kasus penyakit seperti DBD biasanya tidak selalu stabil jika dibaca secara harian. Data harian dapat dipengaruhi oleh keterlambatan pelaporan, hari libur, atau variasi pencatatan. Dengan mengubah data menjadi mingguan, pola umum kasus dapat terlihat lebih jelas dan lebih sesuai dengan tujuan prediksi jumlah kasus periode berikutnya. Agregasi ini juga membuat target penelitian menjadi jumlah kasus DBD mingguan, sehingga masalah yang diselesaikan adalah regresi deret waktu.

Dalam pemodelan deret waktu, fitur *lag* digunakan untuk membawa informasi masa lalu ke dalam baris data saat ini. Misalnya, CASES_lag1 menunjukkan jumlah kasus satu minggu sebelumnya, sedangkan variabel iklim dengan *lag* menunjukkan kondisi iklim beberapa minggu sebelum minggu yang

diprediksi. Fitur tersebut penting karena model tidak boleh menggunakan informasi masa depan. Dengan menyusun data berdasarkan *lag*, model belajar dari riwayat yang tersedia untuk memprediksi periode berikutnya.

## 2.3 ARIMA

*Autoregressive Integrated Moving Average* (ARIMA) adalah model statistik deret waktu yang menggabungkan komponen autoregresif (AR), diferensiasi (I), dan *moving average* (MA) [8, 9]. Model ini digunakan untuk memprediksi nilai masa depan berdasarkan pola historis pada deret waktu. Dalam konteks penelitian ini, deret waktu yang diprediksi adalah jumlah kasus DBD mingguan, sehingga ARIMA memanfaatkan keterkaitan antara jumlah kasus minggu berjalan dan jumlah kasus pada minggu-minggu sebelumnya.




Gambar 2.2. Komponen model ARIMA
Sumber: Diolah penulis berdasarkan [8, 9]

Gambar 2.2 menunjukkan bahwa ARIMA bekerja melalui tiga komponen utama. Komponen pertama adalah *autoregressive* (AR), yaitu bagian model yang menggunakan nilai masa lalu dari deret itu sendiri. Komponen kedua adalah *integrated* (I), yaitu proses diferensiasi untuk mengurangi tren dan membuat deret lebih stasioner. Komponen ketiga adalah *moving average* (MA), yaitu bagian model yang menggunakan pola galat prediksi masa lalu untuk memperbaiki prediksi saat ini.

Model ARIMA dinyatakan sebagai ARIMA$(p,d,q)$, dengan $p$ sebagai orde autoregresif, $d$ sebagai tingkat diferensiasi, dan $q$ sebagai orde *moving average*. Parameter $p$ menunjukkan berapa banyak nilai historis yang digunakan

model. Parameter $d$ menunjukkan berapa kali data didiferensiasi agar pola deret lebih stabil. Parameter $q$ menunjukkan berapa banyak galat masa lalu yang dipertimbangkan. Sebagai contoh, ARIMA(1,1,0) berarti model melakukan diferensiasi satu kali dan menggunakan satu nilai masa lalu sebagai komponen autoregresif tanpa komponen MA.

Secara umum, setelah diferensiasi, komponen ARMA dapat ditulis sebagai berikut:

$$z_t = c + \sum_{i=1}^{p} \phi_i z_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{2.1}$$

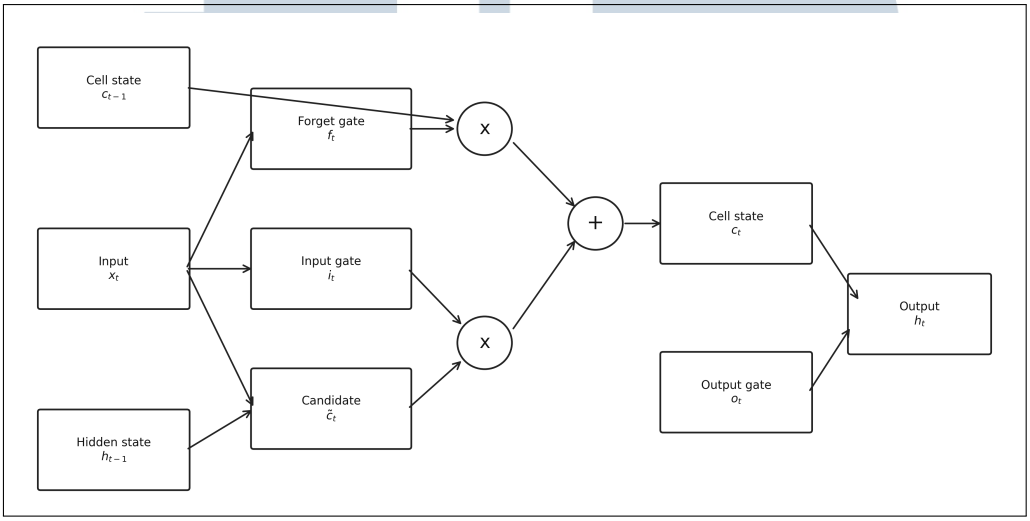
Persamaan 2.1 menunjukkan bentuk umum model ARIMA setelah data didiferensiasi. Pada persamaan tersebut, $z_t$ adalah nilai deret yang telah didiferensiasi pada waktu ke-$t$, $c$ adalah konstanta, $\phi_i$ adalah koefisien autoregresif, $\theta_j$ adalah koefisien *moving average*, dan $\varepsilon_t$ adalah galat pada waktu ke-$t$. Komponen $\sum_{i=1}^{p} \phi_i z_{t-i}$ menunjukkan pengaruh nilai masa lalu terhadap nilai saat ini, sedangkan komponen $\sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$ menunjukkan pengaruh galat prediksi masa lalu terhadap prediksi saat ini. Dengan demikian, ARIMA tidak hanya melihat nilai kasus sebelumnya, tetapi juga memperhitungkan pola kesalahan prediksi sebelumnya setelah data dibuat lebih stasioner melalui proses diferensiasi.

Kelebihan ARIMA adalah strukturnya sederhana, interpretasinya jelas, dan relatif cocok untuk jumlah data yang tidak terlalu besar. Model ini juga kuat sebagai pembanding awal karena dapat menunjukkan apakah pola historis target sudah cukup untuk menghasilkan prediksi yang baik. Namun, ARIMA juga memiliki keterbatasan. Model ini pada dasarnya lebih kuat untuk pola linear dan univariat, sehingga hubungan non-linear antara iklim dan kasus DBD tidak selalu dapat ditangkap secara penuh jika tidak menggunakan pengembangan seperti ARIMAX atau model hibrida.

Pada penelitian ini, ARIMA diposisikan sebagai baseline yang penting. Jika ARIMA sudah menghasilkan galat rendah, maka model yang lebih kompleks seperti LSTM harus menunjukkan peningkatan yang jelas agar kompleksitas tambahannya dapat dibenarkan. Dengan demikian, ARIMA bukan hanya model pembanding, tetapi juga alat untuk menguji seberapa kuat pola autoregresif pada jumlah kasus DBD mingguan.

## 2.4 LSTM

*Long Short-Term Memory* (LSTM) adalah pengembangan dari *Recurrent Neural Network* yang dirancang untuk mempelajari pola sekuensial dan mengatasi masalah *vanishing gradient* [10]. Pada RNN standar, informasi dari waktu yang jauh di masa lalu sering melemah ketika proses pelatihan dilakukan melalui banyak langkah waktu. LSTM mengatasi masalah tersebut dengan menggunakan *cell state* dan mekanisme gerbang yang mengatur aliran informasi.




Gambar 2.3. Arsitektur sel LSTM
Sumber: Diolah penulis berdasarkan [10]

Gambar 2.3 menunjukkan bahwa LSTM menerima tiga informasi utama pada setiap langkah waktu, yaitu input saat ini ($x_t$), *hidden state* sebelumnya ($h_{t-1}$), dan *cell state* sebelumnya ($c_{t-1}$). Input tersebut diproses melalui beberapa gerbang untuk menghasilkan *cell state* baru ($c_t$) dan *hidden state* baru ($h_t$).

Gerbang pertama adalah *forget gate*. Gerbang ini menentukan informasi lama mana yang perlu dipertahankan atau dilupakan dari *cell state* sebelumnya. Jika nilai *forget gate* mendekati 0, informasi lama cenderung dibuang. Jika nilainya mendekati 1, informasi lama cenderung dipertahankan. Dalam konteks prediksi DBD, mekanisme ini berguna karena tidak semua informasi dari minggu-minggu sebelumnya memiliki relevansi yang sama terhadap minggu yang diprediksi.

Gerbang kedua adalah *input gate*. Gerbang ini menentukan informasi baru mana yang akan dimasukkan ke *cell state*. Informasi baru tersebut berasal dari kombinasi input saat ini dan *hidden state* sebelumnya. Pada penelitian ini, input dapat berupa riwayat kasus, fitur iklim tertunda, dan fitur musiman. Dengan

demikian, LSTM memiliki kemampuan untuk mempelajari pola multivariat, bukan hanya pola satu deret target.

Gerbang ketiga adalah *output gate*. Gerbang ini menentukan bagian dari *cell state* yang akan dikeluarkan sebagai *hidden state* pada waktu saat ini. *Hidden state* inilah yang kemudian digunakan untuk menghasilkan prediksi atau diteruskan ke langkah waktu berikutnya. Secara konseptual, LSTM dapat mempertahankan informasi jangka pendek maupun jangka lebih panjang melalui kombinasi *cell state* dan gerbang-gerbang tersebut.

Rangkaian operasi dasar LSTM dapat diringkas sebagai berikut:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.2}$$

Persamaan 2.2 menunjukkan proses pada *forget gate*. Pada persamaan tersebut, $f_t$ adalah keluaran *forget gate*, $\sigma$ adalah fungsi sigmoid, $W_f$ adalah bobot, $h_{t-1}$ adalah *hidden state* sebelumnya, $x_t$ adalah input saat ini, dan $b_f$ adalah bias. Gerbang ini menghasilkan nilai antara 0 dan 1 untuk menentukan seberapa besar informasi lama dari *cell state* sebelumnya perlu dipertahankan.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2.3}$$

Persamaan 2.3 menunjukkan proses pada *input gate*. Pada persamaan tersebut, $i_t$ menentukan seberapa besar informasi baru akan dimasukkan ke memori LSTM. Simbol $W_i$ menunjukkan bobot untuk *input gate*, sedangkan $b_i$ menunjukkan bias. Gerbang ini membantu model memilih informasi baru yang relevan dari input minggu berjalan dan *hidden state* sebelumnya.

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{2.4}$$

Persamaan 2.4 menunjukkan pembentukan kandidat *cell state*. Simbol $\tilde{c}_t$ adalah kandidat informasi baru yang akan dipertimbangkan untuk masuk ke memori, tanh adalah fungsi aktivasi hiperbolik tangent, $W_c$ adalah bobot, dan $b_c$ adalah bias. Kandidat ini dibentuk dari kombinasi input saat ini dan *hidden state* sebelumnya.

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{2.5}$$

Persamaan 2.5 menunjukkan pembaruan *cell state*. Simbol $c_t$ adalah *cell state* saat ini, sedangkan $c_{t-1}$ adalah *cell state* sebelumnya. Bagian $f_t \odot c_{t-1}$ mempertahankan informasi lama yang masih relevan, sedangkan $i_t \odot \tilde{c}_t$ menambahkan informasi baru yang dianggap penting. Simbol $\odot$ menunjukkan perkalian elemen per elemen.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.6}$$

Persamaan 2.6 menunjukkan proses pada *output gate*. Simbol $o_t$ adalah keluaran *output gate*, $W_o$ adalah bobot, dan $b_o$ adalah bias. Gerbang ini menentukan bagian informasi dari *cell state* yang akan dikeluarkan sebagai representasi pada waktu saat ini.

$$h_t = o_t \odot \tanh(c_t) \tag{2.7}$$

Persamaan 2.7 menunjukkan pembentukan *hidden state*. Simbol $h_t$ adalah keluaran LSTM pada waktu ke-*t* yang diperoleh dari hasil *output gate* dan *cell state* saat ini. Dalam penelitian ini, *hidden state* menjadi representasi pola temporal yang kemudian digunakan untuk menghasilkan prediksi jumlah kasus DBD mingguan.

Dalam prediksi DBD, LSTM relevan karena dapat menerima urutan fitur beberapa minggu sebelumnya. Model ini dapat mempelajari hubungan antara riwayat kasus, fitur iklim tertunda, dan pola musiman. Beberapa studi DBD juga mengeksplorasi model cerdas dan pendekatan komputasional untuk menangkap pola temporal penyakit tular vektor [4, 14]. Namun, LSTM membutuhkan jumlah data yang memadai. Pada data deret waktu yang pendek, LSTM berisiko tidak stabil atau kalah dari model statistik yang lebih sederhana.

Alasan tersebut penting dalam interpretasi hasil. Jika LSTM memiliki performa lebih buruk, hasil tersebut tidak otomatis berarti LSTM tidak cocok untuk semua prediksi DBD. Hasil tersebut lebih tepat dibaca sebagai indikasi bahwa ukuran data, panjang deret waktu, dan kelengkapan variabel pada penelitian ini belum cukup untuk memberi keuntungan bagi model *deep learning*.

## 2.5 Efek tunda dan rekayasa fitur

Efek tunda (*lag effect*) menggambarkan jeda antara perubahan kondisi iklim dan perubahan jumlah kasus DBD. Penelitian ini menggunakan *lag* 1–4 minggu untuk tujuh variabel iklim utama, yaitu TN, TX, TAVG, RH_AVG, RR, SS, dan FF_X. Dengan demikian, terbentuk 28 kandidat fitur *lag* iklim. Rentang 1–4 minggu dipilih karena sesuai dengan proses biologis vektor dan inkubasi penyakit yang tidak berlangsung instan [11, 15].

Pada konteks DBD, *lag* diperlukan karena perubahan cuaca tidak langsung berubah menjadi jumlah kasus terlapor pada periode yang sama. Curah hujan dapat membentuk genangan air, kemudian genangan tersebut dapat menjadi tempat berkembang biak nyamuk. Setelah itu masih terdapat proses perkembangan nyamuk, potensi penularan virus, masa inkubasi pada manusia, kemunculan gejala, kunjungan ke fasilitas kesehatan, dan pencatatan kasus. Rangkaian proses tersebut menjelaskan mengapa kondisi iklim beberapa minggu sebelumnya dapat lebih relevan dibandingkan kondisi iklim pada minggu yang sama.

Selain fitur iklim, penelitian ini menambahkan fitur *lag* kasus DBD 1–4 minggu. Fitur tersebut penting karena jumlah kasus penyakit menular biasanya memiliki autokorelasi, yaitu jumlah kasus pada minggu berjalan berkaitan dengan minggu sebelumnya. Jika jumlah kasus pada minggu sebelumnya tinggi, maka minggu berikutnya dapat tetap tinggi karena proses penularan dan pelaporan tidak berhenti secara tiba-tiba. Sebaliknya, jika kasus pada beberapa minggu sebelumnya rendah, model dapat menangkap pola penurunan atau kondisi yang relatif stabil.

Rekayasa fitur *lag* juga membantu mengubah data deret waktu menjadi format yang dapat diproses oleh model pembelajaran mesin. Setiap baris data tidak hanya berisi target minggu berjalan, tetapi juga informasi historis dari minggu-minggu sebelumnya. Dengan cara ini, model LSTM dapat menerima pola sekuensial, sedangkan proses seleksi fitur dapat menilai fitur historis mana yang paling informatif. Namun, pembuatan fitur *lag* menyebabkan beberapa baris awal tidak dapat digunakan karena belum memiliki riwayat lengkap. Hal tersebut merupakan konsekuensi normal pada pemodelan deret waktu.

## 2.6 Seleksi fitur

Seleksi fitur dilakukan untuk mengurangi fitur yang kurang informatif dan menekan risiko *overfitting*. Tahap ini penting karena fitur meteorologi sering

memiliki korelasi antarvariabel, misalnya antara suhu minimum, suhu maksimum, dan suhu rata-rata. Studi prediksi DBD juga menunjukkan bahwa pemilihan fitur dapat memengaruhi akurasi model, khususnya ketika jumlah fitur bertambah akibat pembuatan *lag* [16, 15].

Dalam penelitian ini digunakan metode *SelectKBest* dengan skor *f_regression*. Metode ini menilai hubungan statistik antara setiap fitur kandidat dan target regresi pada data latih. Seleksi fitur hanya dilakukan pada data latih agar informasi dari data uji tidak memengaruhi proses pemilihan fitur.

*SelectKBest* bekerja dengan memberi skor pada setiap fitur secara individual. Pada penelitian ini, skor *f_regression* digunakan karena target yang diprediksi berupa nilai numerik jumlah kasus. Fitur dengan skor lebih tinggi dianggap memiliki hubungan statistik yang lebih kuat dengan target pada data latih. Setelah seluruh fitur diberi skor, hanya sejumlah fitur terbaik yang dipertahankan untuk proses pemodelan LSTM.

Seleksi fitur tidak dimaksudkan untuk menyatakan bahwa fitur yang tidak terpilih sama sekali tidak berpengaruh terhadap DBD. Hasil seleksi hanya menunjukkan bahwa, pada data latih dan konfigurasi penelitian ini, fitur terpilih memiliki hubungan yang lebih kuat terhadap target dibandingkan fitur lain. Hal ini penting karena data yang digunakan relatif terbatas setelah agregasi mingguan dan pembuatan *lag*. Jika terlalu banyak fitur dipaksakan masuk ke model, LSTM berisiko mempelajari pola yang terlalu spesifik pada data latih dan tidak stabil pada data uji.

Keluaran seleksi fitur berupa daftar fitur terpilih yang kemudian digunakan oleh model LSTM. ARIMA tidak menggunakan fitur eksogen karena diposisikan sebagai baseline statistik univariat berbasis riwayat kasus. Dengan pemisahan tersebut, penelitian dapat melihat dua pendekatan yang berbeda: ARIMA menunjukkan kemampuan pola historis target saja, sedangkan LSTM menunjukkan kemampuan model sekuensial ketika diberi fitur historis dan iklim yang telah dipilih.

## 2.7 Normalisasi StandardScaler

*StandardScaler* melakukan transformasi nilai fitur menjadi distribusi dengan rata-rata nol dan standar deviasi satu:

$$x' = \frac{x - \mu}{\sigma} \tag{2.8}$$

Persamaan 2.8 menunjukkan proses normalisasi menggunakan *StandardScaler*. Pada persamaan tersebut, $x'$ adalah nilai hasil normalisasi, $x$ adalah nilai asli, $\mu$ adalah rata-rata data latih, dan $\sigma$ adalah standar deviasi data latih. Nilai asli dikurangi rata-rata agar berada di sekitar nol, kemudian dibagi standar deviasi agar skala antarfitur menjadi lebih sebanding.

Normalisasi digunakan karena fitur iklim memiliki skala berbeda, misalnya curah hujan dalam milimeter dan kelembapan dalam persen. Pada LSTM, normalisasi membantu proses optimasi *gradient descent* agar lebih stabil. Parameter *mean* dan standar deviasi dihitung hanya dari data latih, kemudian diterapkan pada data uji untuk menghindari *data leakage*.

Normalisasi menjadi penting pada model LSTM karena jaringan saraf memperbarui bobot melalui proses optimasi numerik. Jika fitur memiliki skala yang sangat berbeda, fitur dengan nilai numerik besar dapat mendominasi proses pembelajaran, meskipun belum tentu fitur tersebut paling informatif. Sebagai contoh, curah hujan, kelembapan, suhu, dan fitur musiman memiliki rentang nilai yang berbeda. Dengan *StandardScaler*, semua fitur ditempatkan pada skala yang lebih sebanding sehingga proses pelatihan menjadi lebih stabil.

StandardScaler dipilih karena mempertahankan bentuk variasi data dalam satuan standar deviasi. Teknik ini sesuai untuk model LSTM pada penelitian ini karena fitur yang digunakan berasal dari beberapa jenis variabel dengan satuan berbeda, seperti suhu, kelembapan, curah hujan, dan fitur musiman. Dibandingkan transformasi yang hanya memindahkan nilai ke rentang tertentu, StandardScaler membuat nilai fitur berada di sekitar rata-rata nol sehingga proses optimasi jaringan saraf lebih mudah menyesuaikan bobot pada setiap fitur.

Pada data deret waktu, normalisasi harus dilakukan dengan hati-hati. Rata-rata dan standar deviasi tidak boleh dihitung dari seluruh data karena data uji merepresentasikan masa depan yang seharusnya belum diketahui saat model dilatih. Oleh karena itu, parameter normalisasi dihitung hanya dari data latih, kemudian digunakan untuk mentransformasi data latih dan data uji. Prinsip yang sama juga diterapkan pada normalisasi target LSTM sebelum hasil prediksi dikembalikan ke skala kasus asli.

## 2.8 Metrik evaluasi regresi

Penelitian ini menggunakan metrik regresi karena target yang diprediksi adalah jumlah kasus mingguan berbentuk numerik kontinu.

Pemilihan metrik evaluasi disesuaikan dengan tujuan penelitian, yaitu menilai kedekatan hasil prediksi terhadap jumlah kasus DBD aktual. MAE digunakan untuk membaca rata-rata kesalahan dalam satuan kasus, RMSE digunakan untuk memberi perhatian lebih besar pada kesalahan yang tinggi, $R^2$ digunakan untuk melihat kemampuan model menjelaskan variasi data, sedangkan MAPE digunakan sebagai ukuran kesalahan relatif dalam persentase. Keempat metrik tersebut digunakan secara bersamaan karena satu metrik saja belum cukup untuk menggambarkan kualitas model secara menyeluruh.

Kombinasi metrik ini juga penting karena karakter kasus DBD dapat berfluktuasi antarperiode. Model yang memiliki MAE rendah belum tentu aman apabila RMSE-nya tinggi, karena hal tersebut dapat menunjukkan adanya beberapa minggu dengan kesalahan besar. Sebaliknya, nilai $R^2$ yang tinggi perlu tetap dibaca bersama MAE dan RMSE agar performa model tidak hanya dinilai dari kemampuan menjelaskan variasi, tetapi juga dari besar kesalahan prediksi dalam satuan kasus. Oleh karena itu, evaluasi pada penelitian ini tidak menggunakan metrik klasifikasi seperti *accuracy* atau *confusion matrix* sebagai metrik utama.

### 2.8.1 Mean absolute error

*Mean Absolute Error* (MAE) menghitung rata-rata selisih absolut antara nilai aktual dan prediksi.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.9}$$

Persamaan 2.9 menunjukkan cara menghitung *Mean Absolute Error*. Pada persamaan tersebut, $n$ adalah jumlah data pengujian, $y_i$ adalah nilai aktual ke-$i$, dan $\hat{y}_i$ adalah nilai prediksi ke-$i$. Selisih antara nilai aktual dan prediksi dibuat absolut agar kesalahan positif dan negatif tidak saling menghilangkan.

MAE mudah diinterpretasikan karena satuannya sama dengan target, yaitu jumlah kasus DBD. Jika MAE bernilai 28, maka secara rata-rata prediksi model meleset sekitar 28 kasus per minggu dari nilai aktual. Metrik ini sesuai untuk membaca besar kesalahan rata-rata tanpa memperhatikan arah kesalahan, baik

prediksi terlalu tinggi maupun terlalu rendah.

### 2.8.2 Root mean square error

*Root Mean Square Error* (RMSE) memberikan penalti lebih besar pada kesalahan prediksi yang besar.

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

Persamaan 2.10 menunjukkan cara menghitung *Root Mean Square Error*. Pada persamaan tersebut, $n$ adalah jumlah data pengujian, $y_i$ adalah nilai aktual, dan $\hat{y}_i$ adalah nilai prediksi. Selisih antara nilai aktual dan prediksi dikuadratkan terlebih dahulu, kemudian dirata-ratakan, lalu diakarkan kembali agar satuannya kembali sama dengan target.

RMSE lebih sensitif terhadap kesalahan besar karena selisih prediksi dikuadratkan sebelum dirata-ratakan. Pada prediksi DBD, RMSE penting karena model yang sering gagal menangkap lonjakan kasus akan memperoleh nilai RMSE yang tinggi. Oleh karena itu, perbandingan MAE dan RMSE dapat membantu melihat apakah kesalahan model relatif merata atau didominasi oleh beberapa minggu dengan galat besar.

### 2.8.3 Koefisien determinasi

Koefisien determinasi ($R^2$) menunjukkan seberapa baik model menjelaskan variasi data dibandingkan penduga rata-rata.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.11}$$

Persamaan 2.11 menunjukkan cara menghitung koefisien determinasi. Pada persamaan tersebut, $y_i$ adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi, dan $\bar{y}$ adalah rata-rata nilai aktual. Bagian pembilang menunjukkan total kesalahan kuadrat model, sedangkan bagian penyebut menunjukkan total variasi data aktual terhadap rata-ratanya.

Nilai $R^2$ yang mendekati 1 menunjukkan bahwa model mampu menjelaskan variasi data dengan baik. Nilai $R^2$ mendekati 0 menunjukkan bahwa model tidak

jauh lebih baik daripada prediksi menggunakan rata-rata. Nilai $R^2$ juga dapat bernilai negatif apabila model lebih buruk daripada penduga rata-rata. Metrik ini digunakan sebagai pelengkap MAE dan RMSE karena dapat memberikan gambaran tentang seberapa besar variasi kasus yang dapat dijelaskan oleh model.

### 2.8.4 Mean absolute percentage error

*Mean Absolute Percentage Error* (MAPE) menyatakan rata-rata kesalahan dalam bentuk persentase.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{2.12}$$

Persamaan 2.12 menunjukkan cara menghitung *Mean Absolute Percentage Error*. Pada persamaan tersebut, $n$ adalah jumlah data pengujian, $y_i$ adalah nilai aktual, dan $\hat{y}_i$ adalah nilai prediksi. Selisih absolut antara nilai aktual dan prediksi dibagi dengan nilai aktual, kemudian dikalikan 100% agar kesalahan dibaca dalam bentuk persentase.

MAPE membantu membaca kesalahan prediksi dalam bentuk persentase sehingga lebih mudah dibandingkan antarperiode. Namun, MAPE perlu ditafsirkan hati-hati apabila terdapat nilai aktual yang sangat kecil atau nol karena pembagian terhadap nilai aktual dapat membuat persentase menjadi sangat besar atau tidak terdefinisi. Pada penelitian ini, MAPE digunakan sebagai metrik pendukung, sedangkan interpretasi utama tetap diberikan melalui MAE, RMSE, dan visualisasi aktual-prediksi.

## 2.9 Klarifikasi Decision Tree, Gini impurity, entropi, dan confusion matrix

Meskipun model akhir penelitian ini menggunakan ARIMA dan LSTM, konsep *Decision Tree*, Gini impurity, entropi, dan *confusion matrix* perlu diklarifikasi karena sering digunakan pada penelitian *machine learning*.

*Decision Tree* bekerja dengan membagi data secara bertahap berdasarkan fitur yang paling baik memisahkan target. Pada kasus regresi, pemisahan biasanya dipilih berdasarkan penurunan varians atau *mean squared error*. Pada kasus klasifikasi, pemisahan dapat menggunakan Gini impurity atau entropi.

Gini impurity mengukur peluang salah klasifikasi apabila sebuah sampel diberi label secara acak mengikuti distribusi kelas pada simpul. Nilai Gini semakin

kecil berarti simpul semakin murni. Entropi mengukur ketidakpastian distribusi kelas berdasarkan teori informasi. Entropi juga semakin kecil ketika sebuah simpul semakin murni. Perbedaannya, Gini biasanya lebih sederhana secara komputasi, sedangkan entropi menggunakan fungsi logaritmik dan berkaitan dengan konsep *information gain*.

*Confusion matrix* bukan metrik utama untuk masalah regresi karena regresi menghasilkan nilai numerik kontinu, bukan kelas diskrit. Jika *confusion matrix* digunakan, maka penggunaannya hanya sebagai analisis tambahan setelah nilai prediksi regresi dikategorikan menjadi kelas risiko, misalnya rendah, sedang, dan tinggi. Oleh karena itu, penelitian ini tidak menggunakan *confusion matrix* sebagai evaluasi utama; evaluasi utama tetap MAE, RMSE, $R^2$, dan MAPE.

## 2.10 CRISP-DM

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah kerangka kerja pengembangan proyek data mining yang terdiri atas enam tahap utama, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment* [17, 18]. Kerangka kerja ini bersifat iteratif, sehingga peneliti dapat kembali ke tahap sebelumnya apabila ditemukan masalah pada data, model, atau hasil evaluasi.

*Business understanding* adalah tahap untuk memahami tujuan masalah dalam konteks domain. Pada penelitian ini, masalah yang dikaji adalah kebutuhan untuk memperkirakan jumlah kasus DBD mingguan agar dapat menjadi dasar awal sistem peringatan dini. Karena targetnya adalah jumlah kasus, maka masalah penelitian ditetapkan sebagai regresi deret waktu, bukan klasifikasi risiko.

*Data understanding* adalah tahap untuk memahami struktur, kualitas, dan keterbatasan data. Pada penelitian ini, tahap tersebut mencakup identifikasi jumlah baris, jumlah variabel, periode waktu, wilayah administrasi, tipe data, nilai hilang, serta bentuk target CASES. Tahap ini juga memastikan bahwa data harian perlu diagregasi menjadi mingguan agar sesuai dengan tujuan prediksi.

*Data preparation* adalah tahap untuk mengubah data mentah menjadi data siap model. Dalam penelitian ini, tahap tersebut mencakup konversi nilai tidak valid, imputasi median, agregasi harian menjadi mingguan, pembuatan fitur *lag*, pembuatan fitur musiman, pembagian data berbasis waktu, seleksi fitur, dan normalisasi. Tahap ini merupakan bagian paling kritis karena kesalahan seperti pembagian data acak dapat menyebabkan *data leakage*.

*Modeling* adalah tahap pembangunan model prediksi. Penelitian ini menggunakan ARIMA dan LSTM karena keduanya sesuai untuk data deret waktu. ARIMA digunakan sebagai model statistik berbasis autokorelasi kasus, sedangkan LSTM digunakan sebagai model sekuensial yang memanfaatkan fitur historis dan fitur iklim.

*Evaluation* adalah tahap untuk menilai apakah model sudah menjawab tujuan penelitian. Karena target penelitian berupa angka kasus, metrik yang digunakan adalah MAE, RMSE, $R^2$, dan MAPE. Evaluasi juga dilakukan melalui visualisasi aktual-prediksi, residual, dan analisis fitur agar hasil tidak hanya dibaca dari satu angka metrik.

*Deployment* dalam konteks skripsi ini tidak berarti implementasi sistem produksi, melainkan dokumentasi hasil, keterbatasan, dan rekomendasi penggunaan. Model belum diposisikan sebagai sistem operasional kesehatan karena masih membutuhkan validasi lebih lanjut dan penambahan variabel non-iklim seperti kepadatan penduduk, mobilitas, indeks vektor, dan intervensi kesehatan masyarakat.

## 2.11 Tinjauan Penelitian Terdahulu

Tinjauan penelitian terdahulu digunakan untuk memetakan posisi penelitian ini terhadap studi prediksi DBD yang telah dilakukan sebelumnya. Fokus pemetaan diarahkan pada jenis data, pendekatan pemodelan, penggunaan variabel iklim, teknik deret waktu, dan keterbatasan yang masih dapat dikembangkan. Berdasarkan tinjauan tersebut, penelitian ini ditempatkan sebagai prediksi regresi deret waktu kasus DBD mingguan di DKI Jakarta dengan perbandingan ARIMA sebagai model statistik univariat dan LSTM sebagai model sekuensial multivariat berbasis fitur historis dan iklim.

Tabel 2.1. Tinjauan penelitian terdahulu dan celah penelitian

| **Penulis & Tahun** | **Judul** | **Temuan Umum** | **Gap** |
|---|---|---|---|
| Patil dan Pandya (2021) [3] | Forecasting Dengue Hotspots Associated With Variation in Meteorological Parameters Using Regression and Time Series Models | Parameter meteorologi dapat digunakan untuk memprediksi pola DBD dengan pendekatan regresi dan deret waktu. | Studi tersebut belum secara khusus menguji data mingguan DKI Jakarta dengan pembagian data berbasis waktu dan perbandingan ARIMA-LSTM. |

| Penulis & Tahun | Judul | Temuan Umum | Gap |
|---|---|---|---|
| Nor dkk. (2022) [19] | Predicting Dengue Outbreak Based on Meteorological Data Using Artificial Neural Network and Decision Tree Models | ANN dan Decision Tree dapat memanfaatkan data meteorologi untuk memprediksi kejadian DBD. | Fokus penelitian lebih dekat pada prediksi kejadian atau wabah, sedangkan penelitian ini memprediksi jumlah kasus mingguan sebagai masalah regresi. |
| Ong dkk. (2023) [20] | Predicting Dengue Transmission Rates by Comparing Different Machine Learning Models With Vector Indices and Meteorological Data | Kombinasi indeks vektor dan data meteorologi dapat meningkatkan prediksi transmisi DBD. | Data penelitian ini belum memiliki indeks vektor, sehingga perlu menguji sejauh mana riwayat kasus dan iklim saja mampu menjelaskan kasus DBD. |
| Cheng dkk. (2025) [4] | Integrating Meteorological Data and Hybrid Intelligent Models for Dengue Fever Prediction | Integrasi data meteorologi dan model cerdas dapat membantu prediksi DBD, terutama ketika pola hubungan antarvariabel bersifat kompleks. | Penelitian ini belum menggunakan model hibrida, tetapi membandingkan ARIMA dan LSTM terlebih dahulu sebagai dasar evaluasi model statistik dan sekuensial. |
| Jaya dkk. (2025) [14] | Spatiotemporal Dengue Forecasting for Sustainable Public Health in Bandung, Indonesia: A Comparative Study of Classical, Machine Learning, and Bayesian Models | Perbandingan model klasik, *machine learning*, dan Bayesian dapat digunakan untuk memahami perbedaan kemampuan model pada prediksi DBD. | Penelitian ini berfokus pada konteks DKI Jakarta dan membandingkan ARIMA dengan LSTM pada target jumlah kasus mingguan. |
| Tian dkk. (2024) [15] | Precision Prediction for Dengue Fever in Singapore: A Machine Learning Approach Incorporating Meteorological Data | Fitur meteorologi dan fitur historis dapat meningkatkan kemampuan model machine learning dalam memprediksi DBD. | Perlu pengujian pada konteks data Jakarta dengan ukuran data mingguan yang lebih terbatas dan validasi berbasis waktu. |
| Tuan dan Dang (2024) [11] | Leveraging Climate Data for Dengue Forecasting: A Machine Learning Approach in Vietnam | Variabel iklim memiliki hubungan tertunda dengan dinamika DBD dan dapat digunakan dalam rekayasa fitur. | Penelitian ini perlu membatasi lag secara eksplisit, yaitu 1–4 minggu, serta menjelaskan alasan biologis dan teknis pembentukan fitur lag. |
| Mahadee Al Mobin (2024) [16] | Forecasting Dengue in Bangladesh Using Meteorological Variables With a Novel Feature Selection Approach | Seleksi fitur dapat membantu memilih variabel meteorologi yang lebih relevan untuk prediksi DBD. | Penelitian ini menerapkan seleksi fitur hanya pada data latih untuk menghindari data leakage pada skenario deret waktu. |
| da Silva dkk. (2024) [21] | When Climate Variables Improve the Dengue Forecasting: A Machine Learning Approach | Kontribusi variabel iklim terhadap prediksi DBD dapat berbeda antarwilayah dan periode data. | Perlu evaluasi lokal pada DKI Jakarta karena pengaruh iklim terhadap DBD tidak selalu sama dengan wilayah lain. |

| Penulis & Tahun | Judul | Temuan Umum | Gap |
|---|---|---|---|
| Agung Sutriyawan dkk. (2025) [5] | Global Forecasting Models for Dengue Outbreaks in Endemic Regions: A Systematic Review | Model prediksi DBD pada wilayah endemis beragam dan perlu dipilih sesuai karakteristik data, tujuan prediksi, serta kualitas variabel yang tersedia. | Penelitian ini mengambil lingkup lebih terfokus pada ARIMA dan LSTM agar sesuai dengan ketersediaan data dan kebutuhan interpretasi skripsi. |
| Mills dan Donnelly (2024) [22] | Climate-Based Dengue Forecasting Models: Opportunities and Challenges | Model berbasis iklim berpotensi mendukung peringatan dini, tetapi sensitif terhadap kualitas data dan konteks lokal. | Penelitian ini perlu menekankan keterbatasan data serta tidak mengklaim model sebagai sistem peringatan dini operasional. |
| Rahman dan Shiddik (2025) [23] | Explainable Artificial Intelligence for Predicting Dengue Outbreaks in Bangladesh Using Eco-Climatic Triggers | Interpretabilitas model penting agar hasil prediksi DBD dapat dipahami dalam konteks kesehatan masyarakat. | Penelitian ini menggunakan *permutation importance* sebagai analisis interpretasi fitur pada model LSTM. |

Berdasarkan pemetaan pada Tabel 2.1, celah utama penelitian ini terletak pada kebutuhan validasi model prediksi DBD yang disesuaikan dengan karakteristik data lokal DKI Jakarta. Beberapa penelitian terdahulu telah menggunakan variabel iklim, model deret waktu, model *machine learning*, atau *deep learning*, tetapi belum seluruhnya menekankan pembagian data berbasis waktu, pencegahan *data leakage* pada seleksi fitur dan normalisasi, serta perbandingan langsung antara ARIMA dan LSTM pada target jumlah kasus mingguan. Oleh karena itu, penelitian ini berfokus pada prediksi regresi deret waktu dengan alur eksperimen yang disesuaikan dengan sifat data temporal.

Secara eksplisit, *research gap* penelitian ini terdiri atas empat poin. Pertama, masih diperlukan evaluasi prediksi DBD pada konteks lokal DKI Jakarta dengan satuan mingguan. Kedua, masih diperlukan alur validasi yang menjaga urutan waktu agar hasil evaluasi tidak terlalu optimistis. Ketiga, masih diperlukan perbandingan antara model statistik deret waktu yang sederhana dan model sekuensial yang lebih kompleks pada data yang sama. Keempat, masih diperlukan analisis fitur untuk melihat apakah prediksi lebih banyak dipengaruhi oleh riwayat kasus atau oleh faktor iklim tertunda. Empat celah tersebut menjadi dasar penelitian ini dalam membandingkan ARIMA dan LSTM pada prediksi jumlah kasus DBD mingguan.

Posisi penelitian ini tidak hanya terletak pada penggunaan dua model, tetapi juga pada penyesuaian seluruh alur eksperimen terhadap bentuk data deret waktu. Data harian terlebih dahulu diubah menjadi data mingguan, kemudian fitur historis dibentuk menggunakan *lag* 1–4 minggu. Setelah itu, pembagian data dilakukan

secara kronologis, bukan acak, sehingga evaluasi model lebih mendekati kondisi prediksi sebenarnya. Dengan demikian, penelitian ini berusaha menghindari kelemahan umum pada eksperimen deret waktu, yaitu masuknya informasi masa depan ke dalam proses pelatihan.

Dari sisi pemodelan, ARIMA digunakan untuk melihat kemampuan pola historis jumlah kasus dalam menjelaskan kasus pada periode berikutnya. LSTM digunakan untuk menguji apakah model sekuensial yang lebih kompleks dapat memperoleh manfaat tambahan dari fitur iklim dan riwayat kasus. Perbandingan ini penting karena model yang lebih kompleks tidak selalu menghasilkan performa lebih baik, terutama ketika jumlah data terbatas. Oleh karena itu, hasil penelitian tidak hanya dilihat dari nilai metrik terbaik, tetapi juga dari kesesuaian model dengan karakteristik data, stabilitas prediksi, dan kemudahan interpretasi.

Tinjauan penelitian terdahulu juga menunjukkan bahwa prediksi DBD sangat dipengaruhi oleh konteks wilayah. Hubungan antara iklim dan jumlah kasus dapat berbeda karena variasi kepadatan penduduk, perilaku masyarakat, kualitas pelaporan, kondisi lingkungan, serta intervensi kesehatan masyarakat. Oleh sebab itu, hasil penelitian dari negara atau wilayah lain tidak dapat langsung dianggap berlaku untuk DKI Jakarta. Penelitian ini perlu dilakukan sebagai evaluasi lokal agar kesimpulan yang diperoleh sesuai dengan data dan konteks wilayah yang dianalisis.

Berdasarkan uraian tersebut, kontribusi utama penelitian ini adalah menyusun alur prediksi DBD mingguan yang konsisten dengan prinsip deret waktu, membandingkan ARIMA dan LSTM pada periode uji yang sama, serta menjelaskan hasil model menggunakan metrik regresi dan visualisasi aktual-prediksi. Kontribusi tersebut diharapkan dapat menjadi dasar awal untuk pengembangan model prediksi DBD yang lebih lengkap pada penelitian berikutnya, khususnya dengan menambahkan variabel non-iklim seperti kepadatan penduduk, mobilitas, indeks vektor nyamuk, dan data intervensi kesehatan masyarakat.

Selain itu, tinjauan penelitian terdahulu memberikan dasar untuk menentukan batasan penelitian. Penelitian ini tidak diarahkan untuk mengklasifikasikan tingkat risiko DBD, melainkan memprediksi jumlah kasus mingguan. Oleh karena itu, keluaran model tetap berupa angka kasus dan metrik evaluasi yang digunakan adalah metrik regresi.

Tinjauan tersebut juga memperlihatkan bahwa penggunaan variabel iklim perlu disertai rekayasa fitur yang sesuai dengan mekanisme kejadian DBD. Iklim tidak memengaruhi jumlah kasus secara langsung pada hari atau minggu yang

sama, melainkan melalui proses tertunda yang berkaitan dengan genangan air, perkembangan nyamuk, penularan virus, masa inkubasi, dan pelaporan kasus. Oleh sebab itu, penggunaan fitur *lag* 1–4 minggu dalam penelitian ini bukan hanya keputusan teknis, tetapi juga didasarkan pada karakter epidemiologis DBD.

Keterbatasan data menjadi pertimbangan penting dalam membaca hasil penelitian ini. Jumlah sampel setelah agregasi mingguan dan pembuatan fitur *lag* tidak sebesar data harian awal, sehingga model yang kompleks belum tentu selalu unggul. Kondisi tersebut menjadi alasan mengapa ARIMA tetap digunakan sebagai pembanding, sedangkan LSTM diuji untuk melihat manfaat informasi multivariat dari iklim dan riwayat kasus.

Dengan demikian, bagian tinjauan penelitian terdahulu menjadi landasan bagi metode yang dijelaskan pada Bab III. Keputusan mengenai agregasi mingguan, pembagian data berbasis waktu, seleksi fitur, normalisasi, pemilihan ARIMA dan LSTM, serta penggunaan metrik regresi diturunkan dari celah dan keterbatasan penelitian sebelumnya.