

BAB 1

PENDAHULUAN

1.1 Latar belakang masalah

Demam Berdarah Dengue (DBD) merupakan penyakit menular yang disebabkan oleh virus dengue dan ditularkan melalui nyamuk *Aedes aegypti* atau *Aedes albopictus* [1, 2]. Penyakit ini menjadi perhatian kesehatan masyarakat karena jumlah kasusnya dapat berubah dari waktu ke waktu dan dapat meningkat pada periode tertentu. Dalam konteks wilayah perkotaan seperti DKI Jakarta, dinamika DBD menjadi semakin kompleks karena penularan tidak hanya dipengaruhi oleh faktor biologis, tetapi juga oleh kondisi lingkungan, kepadatan aktivitas penduduk, dan proses pelaporan kasus kesehatan.

Salah satu faktor yang sering dikaji dalam prediksi DBD adalah faktor iklim. Suhu udara, kelembapan, curah hujan, durasi penyinaran matahari, dan kecepatan angin dapat memengaruhi tempat perkembangbiakan nyamuk, masa hidup nyamuk dewasa, aktivitas menggigit, serta perkembangan virus di dalam tubuh nyamuk [3, 4]. Namun, pengaruh tersebut tidak selalu muncul pada hari yang sama karena terdapat jeda antara perubahan kondisi lingkungan, perkembangan populasi vektor, infeksi manusia, munculnya gejala, pemeriksaan, dan pencatatan kasus. Oleh karena itu, hubungan antara iklim dan kasus DBD lebih tepat dipahami sebagai hubungan temporal yang memiliki efek tunda.

Karakter temporal tersebut membuat data DBD perlu diperlakukan sebagai data deret waktu (*time series*). Jumlah kasus pada suatu minggu tidak berdiri sendiri, tetapi dapat berkaitan dengan jumlah kasus pada minggu sebelumnya dan kondisi iklim pada beberapa minggu sebelumnya. Jika urutan waktu diabaikan, model dapat belajar dari pola yang tidak realistis karena data masa depan dapat tercampur ke dalam data pelatihan. Pada penelitian prediksi, kondisi yang lebih sesuai dengan situasi nyata adalah model dilatih menggunakan data masa lalu, kemudian diuji untuk memperkirakan periode yang lebih baru. Oleh karena itu, pembagian data secara acak tidak tepat untuk penelitian ini karena berisiko menyebabkan *data leakage* dan membuat nilai evaluasi terlihat lebih baik daripada kemampuan sebenarnya.

Berbagai penelitian telah menggunakan model statistik, *machine learning*, dan *deep learning* untuk memprediksi kasus DBD. Kajian sistematis menunjukkan

bahwa model prediksi DBD di wilayah endemis semakin banyak memanfaatkan data iklim, data epidemiologi, dan pendekatan komputasional untuk mendukung peringatan dini [5, 6]. Studi di Indonesia juga menunjukkan bahwa pendekatan *machine learning* mulai digunakan untuk mendukung prediksi dini DBD, meskipun pendekatan, target, satuan data, dan strategi validasi yang digunakan dapat berbeda [7]. Perbedaan tersebut penting karena model yang bekerja baik pada satu wilayah atau satu bentuk data belum tentu memberikan hasil yang sama pada wilayah dan struktur data lain.

Model statistik deret waktu seperti ARIMA memiliki keunggulan dalam menangkap autokorelasi, tren historis, dan pola jangka pendek pada data kasus [8, 9]. ARIMA relatif mudah diinterpretasikan dan dapat menjadi pembanding yang kuat ketika jumlah data tidak terlalu besar. Di sisi lain, model *deep learning* seperti LSTM memiliki kemampuan mempelajari pola sekuensial dan hubungan non-linear dari fitur multivariat karena menggunakan mekanisme memori jangka panjang [10, 2, 11]. LSTM berpotensi memanfaatkan riwayat kasus, fitur iklim tertunda, dan fitur musiman secara bersamaan. Namun, model yang lebih kompleks tidak selalu memberikan hasil lebih baik, terutama jika jumlah data terbatas atau variabel yang tersedia belum menggambarkan seluruh faktor penularan.

Penelitian ini berfokus pada *dataset* "Jakarta Dengue Weather 2021–2024" yang berisi 2.922 data harian dari dua wilayah administrasi Jakarta. Secara bentuk awal, *dataset* tersebut merupakan data tabular atau matriks karena tersusun atas baris observasi dan kolom variabel. Akan tetapi, karena memiliki atribut tanggal dan tujuan penelitian adalah memprediksi jumlah kasus pada periode berikutnya, data tersebut tidak diperlakukan sebagai observasi acak. Data harian diagregasi menjadi deret waktu mingguan Jakarta sehingga target penelitian menjadi jumlah kasus DBD mingguan. Agregasi mingguan dipilih karena data kasus harian dapat dipengaruhi oleh variasi pencatatan, sedangkan pola mingguan lebih sesuai untuk membaca dinamika kasus dan mengevaluasi model prediksi periode berikutnya.

Dalam penelitian ini, masalah yang diselesaikan adalah regresi deret waktu, bukan klasifikasi. Target model tetap berupa angka jumlah kasus DBD mingguan, bukan label kelas seperti rendah, sedang, atau tinggi. Konsekuensinya, evaluasi utama menggunakan metrik regresi seperti MAE, RMSE, R^2 , dan MAPE, bukan *confusion matrix*. Keputusan ini penting karena jenis target menentukan bentuk label, metode validasi, model yang digunakan, serta cara membaca hasil. Jika target diubah menjadi kelas, maka makna penelitian akan berubah menjadi klasifikasi risiko. Karena tujuan penelitian adalah memperkirakan jumlah kasus, target

dipertahankan sebagai nilai numerik.

Pemilihan ARIMA dan LSTM didasarkan pada kesesuaian keduanya terhadap data deret waktu. ARIMA digunakan sebagai model statistik univariat yang memanfaatkan riwayat kasus untuk menguji kekuatan pola autoregresif. LSTM digunakan sebagai model sekuensial multivariat yang dapat menerima fitur hasil seleksi, termasuk fitur *lag* kasus, fitur *lag* iklim, dan fitur musiman. Dengan membandingkan keduanya, penelitian ini tidak hanya mencari model dengan galat terkecil, tetapi juga menilai apakah kompleksitas model *deep learning* benar-benar memberikan keuntungan pada ukuran data yang tersedia. Hasil pembandingan sederhana seperti *naive lag-1* juga digunakan agar interpretasi performa model tidak hanya bergantung pada perbandingan ARIMA dan LSTM.

Tinjauan terhadap penelitian terdahulu dan celah penelitian disajikan secara khusus pada Bab 2 bagian *Tinjauan Penelitian Terdahulu*. Bagian tersebut digunakan sebagai dasar untuk menegaskan posisi penelitian ini, yaitu prediksi regresi deret waktu DBD mingguan dengan perbandingan ARIMA dan LSTM, pembagian data berbasis waktu, pencegahan *data leakage*, serta seleksi fitur yang hanya dilakukan pada data latih.

1.2 Rumusan masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah penelitian ini adalah sebagai berikut:

1. Bagaimana mengubah *dataset* harian DBD dan iklim menjadi deret waktu mingguan yang sesuai untuk tugas regresi?
2. Bagaimana menerapkan *preprocessing*, pembuatan fitur *lag*, normalisasi, seleksi fitur, dan pembagian data berbasis waktu tanpa menyebabkan *data leakage*?
3. Bagaimana kinerja ARIMA dan LSTM dalam memprediksi jumlah kasus DBD mingguan berdasarkan metrik MAE, RMSE, R^2 , dan MAPE?
4. Model manakah yang lebih sesuai untuk karakteristik *dataset* DBD Jakarta 2021–2024?

1.3 Batasan permasalahan

Agar penelitian terarah, batasan masalah yang digunakan adalah sebagai berikut:

1. Data yang digunakan adalah *dataset* sekunder "Jakarta Dengue Weather 2021–2024" yang memuat data harian iklim dan kasus DBD dari Jakarta Pusat dan Jakarta Utara.
2. Target prediksi adalah jumlah kasus DBD mingguan Jakarta setelah agregasi lintas wilayah; variabel iklim mingguan dihitung sebagai rata-rata.
3. Penelitian berfokus pada tugas regresi, bukan klasifikasi risiko. Oleh karena itu, metrik utama yang digunakan adalah MAE, RMSE, R^2 , dan MAPE.
4. Algoritma yang dibandingkan adalah ARIMA dan LSTM karena keduanya sesuai untuk data deret waktu.
5. Fitur iklim yang digunakan terbatas pada tujuh variabel utama, yaitu TN, TX, TAVG, RH_AVG, RR, SS, dan FF_X, beserta fitur *lag* 1–4 minggu.
6. Penelitian tidak menggunakan variabel non-iklim seperti kepadatan penduduk, mobilitas, indeks jentik, fogging, sanitasi, atau kebijakan kesehatan.

1.4 Tujuan penelitian

Tujuan penelitian ini adalah:

1. Menyusun data DBD dan iklim harian menjadi deret waktu mingguan yang sesuai untuk prediksi jumlah kasus.
2. Menerapkan alur *preprocessing* yang tepat untuk data deret waktu, termasuk imputasi median, fitur *lag*, normalisasi *StandardScaler*, seleksi fitur, dan split kronologis.
3. Mengembangkan dan mengevaluasi model ARIMA dan LSTM untuk memprediksi kasus DBD mingguan.
4. Menentukan model yang lebih sesuai dengan karakteristik data Jakarta berdasarkan metrik regresi dan analisis hasil prediksi.

1.5 Manfaat penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Manfaat akademis: memberikan contoh penerapan pemodelan deret waktu untuk prediksi DBD dengan alur evaluasi yang menghindari *data leakage*.
2. Manfaat metodologis: menunjukkan pentingnya pemilihan model berdasarkan struktur data, bukan hanya berdasarkan popularitas algoritma.
3. Manfaat praktis: menyediakan dasar awal untuk pengembangan sistem pendukung keputusan prediksi DBD mingguan, dengan catatan bahwa implementasi operasional masih memerlukan variabel non-iklim dan validasi tambahan.

1.6 Sistematika penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
Menjelaskan latar belakang, rumusan masalah, batasan, tujuan, manfaat, dan sistematika penelitian.
- Bab 2 LANDASAN TEORI
Menjelaskan teori DBD dan faktor iklim, konsep deret waktu, ARIMA, LSTM, seleksi fitur, normalisasi, metrik regresi, klarifikasi mengenai *Decision Tree*, Gini impurity, entropi, penggunaan *confusion matrix*, serta tinjauan penelitian terdahulu.
- Bab 3 METODOLOGI PENELITIAN
Menjelaskan metodologi penelitian, alur CRISP-DM, persiapan data, agregasi mingguan, *time-based split*, seleksi fitur, pengembangan model ARIMA dan LSTM, serta prosedur evaluasi.
- Bab 4 HASIL DAN DISKUSI
Menyajikan hasil implementasi, hasil seleksi fitur, hasil evaluasi ARIMA dan LSTM, visualisasi aktual dan prediksi, serta pembahasan keterbatasan.
- Bab 5 KESIMPULAN DAN SARAN
Menyajikan simpulan dan saran berdasarkan hasil evaluasi model dan karakteristik data.