

BAB III

METODE PENELITIAN

3.1 Objek Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan membandingkan performa dua algoritma, yaitu Generalized Linear Model dan Linear Regression, dalam memprediksi hujan berdasarkan data cuaca. Penelitian ini dilakukan secara digital dengan memanfaatkan data yang diperoleh dari Kaggle. Kegiatan penelitian dimulai pada bulan November dan berlanjut hingga selesai. Prosesnya mencakup beberapa tahapan penting, yaitu pengumpulan data, preprocessing, penerapan algoritma, perbandingan algoritma GLM dan Linear Regression, evaluasi model, serta analisis hasil penelitian.

3.1.1 Waktu Penelitian

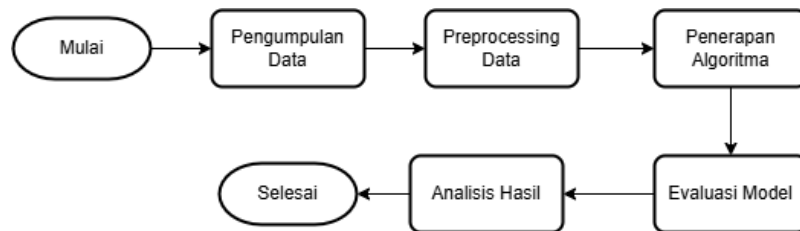
Penelitian dilakukan dengan memproses data yang telah dikumpulkan. Pengumpulan data berlangsung dari tanggal 5 November 2024 sampai Januari 2025.

Tabel 3.1 Waktu Penelitian

Waktu Penelitian	Nov-24				Des-24				Jan-25				Feb-25			
Tahapan Penelitian	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Pengumpulan Data																
Prerocessing																
Penerapan Algoritma																
Evaluasi																
Model																
Analisis Hasil																

3.2 Prosedur Penelitian

Untuk mencapai hasil yang diharapkan dalam penelitian, diperlukan prosedur penelitian yang jelas. Berikut tahapan yang dilakukan digambarkan dalam Gambar 3.1.



Gambar 3.1 Tahapan Penelitian

Alur Penjelasan :

1. Pengumpulan Data

Langkah awal penelitian ini adalah mengumpulkan data historis cuaca dari <https://www.kaggle.com>. Data ini mencakup berbagai parameter penting, seperti suhu, kelembaban, kecepatan angin, tutupan awan, tekanan dan hujan. Selain itu, dilakukan pemeriksaan kelengkapan untuk memastikan data sesuai dengan lokasi, periode waktu, dan relevansi penelitian. Proses validasi awal juga dijalankan guna menjamin kualitas data dalam hal akurasi dan konsistensi.

2. Preprocessing Data

Proses data untuk meningkatkan kualitasnya. Tahapan ini mencakup *encoding*, normalisasi, *class imbalance*.

a. Encoding

Pada tahapan ini akan melakukan encoding yang dimana akan mengubah nilai kategorikal menjadi nilai numerik dengan menggunakan LabelEncoder dari modul sklearn (Istianto, 2023). Variabel kategorikal merupakan jenis data yang terdiri atas nilai berupa kategori atau label teks, seperti "*rain*" dan "*no rain*" pada kolom *Rain* dalam dataset. Untuk dapat digunakan dalam algoritma pembelajaran mesin, yang biasanya hanya mendukung data numerik, variabel ini perlu dikonversi ke dalam format angka, misalnya "*rain*" diberi label 1 dan "*no rain*" diberi label 0.

b. Normalisasi

Normalisasi data adalah suatu teknik penting dalam preprocessing data yang melibatkan penyesuaian ulang nilai-nilai dalam dataset (Suarisman, 2023). Hal ini diperlukan karena perbedaan besar dalam rentang nilai antar atribut dapat mengganggu kinerja optimal atribut dalam analisis data, normalisasi data menjadi suatu langkah krusial untuk memastikan data yang diolah lebih konsisten dan akurat (Suarisman, 2023). Misalnya, jika *Pressure* bernilai ribuan dan *Wind_Speed* bernilai satuan, model cenderung menganggap *Pressure* lebih penting hanya karena skala nilainya lebih besar.

$$Z = \frac{x - \mu}{\sigma} \quad (3)$$

Keterangan:

Z = nilai Z-Score yang dihasilkan.

X = nilai dari data yang akan dinormalisasi.

μ = rata-rata (mean) dari seluruh data.

σ = simpangan baku (standard deviation) dari seluruh data.

c. Class Imbalance

Resampling merupakan teknik dimana mencoba melakukan penyeimbangan data asli melalui proses algoritma sampling dengan menyesuaikan jumlah sampel dalam kelas berbeda (Pratama, 2022). Pada penelitian ini menggunakan metode Random oversampling untuk melakukan duplikasi sampel pada kelas minoritas. Ketidakseimbangan ini dapat menyebabkan model pembelajaran mesin bias terhadap kelas mayoritas, sehingga menghasilkan prediksi yang kurang akurat, terutama untuk kelas minoritas.

3. Penerapan Algoritma

Pada tahap penerapan algoritma, dataset dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian, dengan rasio tertentu seperti

80:20. Data pelatihan digunakan untuk membangun model, sedangkan data pengujian berfungsi untuk menilai kinerja model tersebut. Algoritma *Generalized Linear Model* (GLM) diterapkan untuk menganalisis hubungan linier antara variabel cuaca, dengan *Linear Regression* sebagai metode pembandingan. Untuk memaksimalkan performa model, dilakukan penyesuaian *hyperparameter*, seperti jumlah literasi dan pilihan fungsi aktivasi. Proses ini dilakukan secara terstruktur, mulai dari pembangunan model, pengujian validitas hasil, hingga analisis performa menggunakan data pengujian.

4. Evaluasi Model

Evaluasi dilakukan untuk menilai kinerja model yang telah dibangun. Validasi silang digunakan untuk menguji keandalan model terhadap dataset yang berbeda. Evaluasi ini bertujuan untuk mengidentifikasi keunggulan dan keterbatasan model dalam memprediksi parameter cuaca serta memberikan gambaran menyeluruh tentang performa masing-masing algoritma.

5. Analisis Hasil

Tahap terakhir melibatkan analisis mendalam terhadap hasil evaluasi model. Performa GLM dan *Linear Regression* dibandingkan berdasarkan metrik seperti akurasi, RMSE, dan MAE untuk menentukan algoritma yang lebih efektif. Rekomendasi pengembangan, seperti penambahan data atau penerapan algoritma baru, diusulkan guna meningkatkan akurasi dan stabilitas prediksi. Kesimpulan penelitian dirangkum berdasarkan temuan dan kontribusi yang diberikan terhadap pengembangan ilmu pengetahuan

a. Akurasi

Akurasi merupakan perbandingan antara jumlah prediksi yang tepat dan total data yang diprediksi. Nilai ini menggambarkan seberapa baik kinerja model dalam melakukan prediksi secara keseluruhan, untuk model statistik

yang menggabungkan hubungan linear dengan distribusi probabilitas non-normal dapat dilihat pada Persamaan 3.

$$Akurasi = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

Keterangan:

- y_i : Nilai aktual.
 $\hat{y}_i$: Nilai prediksi.
 $\bar{y}$: Nilai rata-rata dari nilai aktual.
 n : Jumlah data

b. RMSE

Hasil prediksi yang diperoleh dapat dihitung seberapa besar kesalahan yang dihasilkan dengan membandingkannya dengan data aktual. Pada penelitian ini akan menggunakan perhitungan *Root Mean Square Error* (RMSE) untuk mengukur seberapa besar rata-rata kesalahan (*residual*) antara hasil prediksi dan nilai aktualnya, pada Persamaan 4 merupakan rumus perhitungan RMSE (Patriya, 2020). Untuk model statistik yg menggabungkan hubungan linear dengan distribusi probabilitas non- normal dapat dilihat pada Persamaan 4.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

Keterangan:

- y_i : Nilai aktual.
 $\hat{y}_i$: Nilai prediksi.
 n : Jumlah data.

c. MAE

MAE adalah metrik evaluasi yang menghitung rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual. Berbeda dengan RMSE,

MAE memberikan bobot yang sama pada setiap kesalahan tanpa mempengaruhi hasil oleh outlier. MAE memberikan gambaran langsung tentang seberapa besar kesalahan prediksi secara rata-rata, untuk model statistik yg menggabungkan hubungan linear dengan distribusi probabilitas non-normal dapat dilihat pada Persamaan 5.

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

Keterangan:

- n : Jumlah data
 y_i : Nilai aktual pada data ke- i
 $\hat{y}_i$: Nilai prediksi pada data ke- i

