BAB III
METODOLOGI PENELITIAN

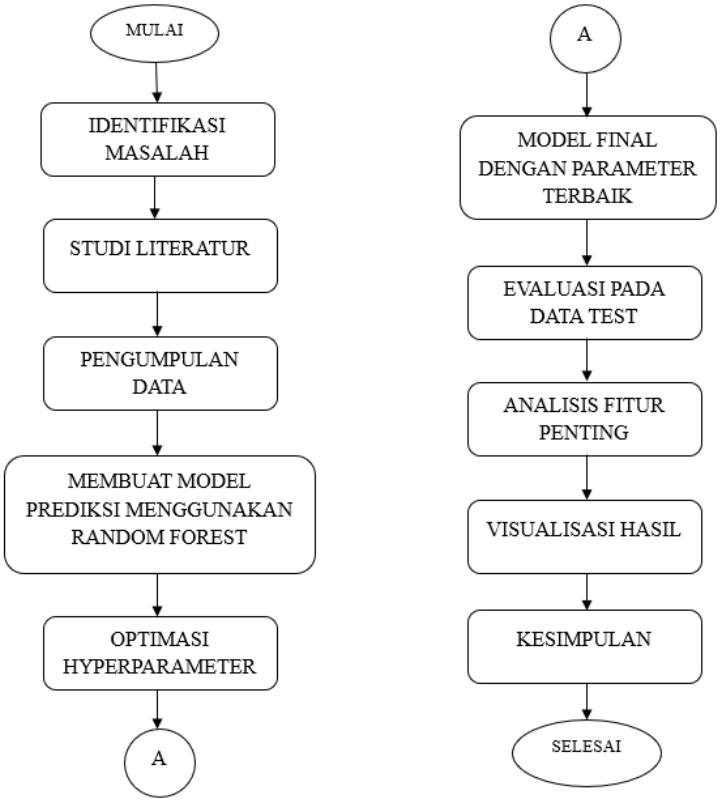
3.1 Desain Penelitian

Penelitian ini dirancang dengan pendekatan kuantitatif berbasis data historis untuk memprediksi daya keluaran sistem fotovoltaik menggunakan algoritma *machine learning*, khususnya metode *Random Forest*. Data yang digunakan merupakan data historis atau dataset yang sudah ada sebelumnya, meliputi suhu udara, kelembaban, intensitas cahaya, dan kecepatan angin. Seluruh proses penelitian dilakukan secara komputasional melalui perangkat lunak MATLAB, mulai dari pemrosesan data, pelatihan model, hingga evaluasi hasil. Desain penelitian ini mencakup beberapa tahapan utama, yaitu pengumpulan data, pra-pemrosesan, pembuatan model prediksi dengan *Random Forest*, optimasi parameter model, validasi model, serta evaluasi performa. Selain itu, untuk memberikan pemahaman lebih lanjut terhadap hasil model, dilakukan juga visualisasi prediksi dalam bentuk grafik.

3.2 Prosedur Metode Penelitian

Penelitian ini mengikuti alur sistematis sebagaimana digambarkan dalam flowchart pada Gambar 3.1 Proses diawali dari identifikasi masalah yang menekankan pada perlunya model prediksi daya fotovoltaik berbasis variabel lingkungan tanpa mengandalkan fitur internal panel. Langkah berikutnya adalah studi literatur untuk menentukan pendekatan yang sesuai, dilanjutkan dengan pengumpulan dan pra-pemrosesan data lingkungan (suhu, kelembaban, intensitas cahaya, dan kecepatan angin). Data kemudian dinormalisasi dan dibagi menjadi data latih dan data uji. Model prediksi dikembangkan menggunakan algoritma *Random Forest*, yang disertai proses optimasi hyperparameter seperti jumlah pohon (ntree) dan jumlah fitur terpilih per split (mtry). Evaluasi model dilakukan menggunakan metrik MAE, MSE, RMSE, R^2, dan MAPE. Untuk memperkuat analisis, hasil prediksi divisualisasikan melalui berbagai grafik seperti scatter plot,

residual plot, boxplot, serta grafik feature importance. Seluruh tahapan ini dirancang untuk memastikan bahwa model yang dikembangkan dapat digunakan secara efektif dalam konteks sistem fotovoltaik di wilayah tropis.



Gambar 3. 1 Alur Penelitian

3.3 Variabel Penelitian

Penelitian ini menggunakan empat variabel input utama antara lain intensitas cahaya, suhu, dan kelembaban. Serta satu variabal target yaitu daya fotovoltaik seperti yang di jelaskan dalam Tabel 3.1

Tabel 3. 1 Variabel

Variabel	Satuan	Tipe Data
Intenstas Cahaya	Lux	Numerik
Suhu	°C	Numerik
Kelembaban	%	Numerik

WindSpeed	m/s	Numerik
Daya fotovoltaik	W	Numerik

3.4 Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari hasil pencatatan sistem *smart street lighting* yang terletak di kota Malang. Dataset ini terdiri dari 158.300 baris data yang terekam secara berkala selama 4 hari dari tanggal 1 – 4 November 2024 seperti yang di tunjukan pada tabel 3.2. Data ini kemudian digunakan sebagai dasar dalam membangun model *Random Forest.*

Tabel 3. 2 Dataset hasil dari monitoring sistem SSL

Id	date	temp	humi	lux	windSpeed	powerPV
1	01/11/2024 00.00	24,81	81,7	8,81	671,83	0
2	01/11/2024 00.00	24,81	81,8	8,82	671,8	0
3	01/11/2024 00.00	24,81	81,8	8,81	672,11	0
4	01/11/2024 00.00	24,81	81,9	8,83	672,07	0
5	01/11/2024 00.00	24,81	81,9	8,83	672,38	0
6	01/11/2024 00.00	24,81	81,7	8,83	672,37	0
7	01/11/2024 00.00	24,81	81,7	8,83	671,75	0
8	01/11/2024 00.00	24,81	81,6	8,83	671,73	0
9	01/11/2024 00.00	24,81	81,6	8,85	671,37	0
10	01/11/2024 00.00	24,81	81,3	8,85	671,38	0
11	01/11/2024 00.00	24,81	81,3	8,84	670,47	0
12	01/11/2024 00.00	24,81	81,1	8,84	669,87	0
13	01/11/2024 00.00	24,81	80,9	8,83	669,89	0
14	01/11/2024 00.00	24,81	80,9	8,84	669,24	0
15	01/11/2024 00.00	24,81	80,9	8,82	669,3	0
16	01/11/2024 00.00	24,81	80,9	8,83	669,27	0
17	01/11/2024 00.00	24,81	80,9	8,83	669,28	0
18	01/11/2024 00.00	24,81	80,6	8,82	669,33	0
19	01/11/2024 00.00	24,81	80,6	8,82	668,41	0
20	01/11/2024 00.00	24,81	80,4	8,81	668,43	0

3.5 Metode Pengolahan data

3.5.1 Preprocessing Data

1. Normalisasi Data

Terdapat perbedaan dimensi antara data input. Untuk menghilangkan perbedaan ini dan mempercepat operasi model, data dinormalisasi, dan metode normalisasi seperti yang ditunjukkan persamaan 3.1.

$$x_n = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad \textbf{(3.1)}$$

Dimana x_i merupakan data mentah masukan, $x_{min}\, dan\, x_{max}$ adalah nilai minimum dan maksimum dari data asli pada kolom tersebut, dan x_n adalah data yang telah dinormalisasi.

2. Pembagian Dataset

Proporsi pembagian ditentukan melalui studi literatur yang menunjukkan 70:30 optimal untuk dataset berukuran sedang. dengan data pelatihan sebesar 70% (110,810 data) untuk mempelajari pola data, dan data pengujian 30% (47.490 data) untuk evaluasi akhir.

3. Pelatihan Model Awal

Model ini dilatih dengan parameter awal, yaitu jumlah pohon keputusan (ntrees) sebanyak 100 dan jumlah fitur yang diambil secara acak (mtry) yang ditentukan berdasarkan akar kuadrat dari jumlah fitur yang ada. Model ini dilatih dengan menggunakan fungsi *TreeBagger* yang merupakan implementasi dari algoritma *Random Forest* di MATLAB.

4. Optimasi Hyperparameter

Untuk meningkatkan kinerja model, dilakukan optimasi hyperparameter. Proses ini melibatkan pengujian berbagai kombinasi jumlah pohon keputusan (ntrees) dan jumlah fitur (mtry) yang diambil secara acak untuk menemukan kombinasi parameter terbaik.

5. Evaluasi Final

Model akhir dievaluasi menggunakan data pengujian untuk mengukur kinerjanya. Metrik yang sama digunakan untuk mengevaluasi model akhir. Hasil evaluasi ini memberikan gambaran tentang seberapa baik model dapat memprediksi energi yang dihasilkan oleh sistem fotovoltaik.

6. Visualisasi Hasil

Hasil dari model akhir divisualisasikan untuk memberikan pemahaman yang lebih baik mengenai kinerja model. Visualisasi ini mencakup scatter plot antara daya aktual dan daya prediksi, serta histogram dari residual untuk melihat distribusi kesalahan prediksi.

7. Feature Importance

Terakhir, pentingnya fitur dalam model dievaluasi menggunakan OOBPermutedVarDeltaError. Visualisasi dari kontribusi setiap fitur memberikan wawasan tentang faktor-faktor yang paling berpengaruh dalam memprediksi energi yang dihasilkan oleh sistem fotovoltaik.

8. Simpan Model dan Hasil

Setelah semua proses selesai, model akhir dan hasil evaluasi disimpan dalam format file yang dapat diguna kan untuk analisis lebih lanjut atau implementasi di masa depan.

3.6 Indikator Evaluasi Kinerja

Untuk memahami performa model prediksi, diperlukan indikator analisis kesalahan. Penelitian ini menggunakan *Mean Absolute Error* (MAE), *Mean Absolute Percentage Error* (MAPE), *Mean Square Error* (MSE), *Root Mean Square Error* (RMSE) dan *Koefisien Determinasi* (R^2).

3.6.1 *Mean Absolute Error* (MAE)

MAE merupakan nilai rata-rata dari selisih absolut antara hasil prediksi dan nilai aktual, yang digunakan untuk menggambarkan tingkat kesalahan peramalan secara nyata. Nilai MAE dihitung dengan menggunakan persamaan 3.2 (Liu & Sun, 2019).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_n - y_i\right| \quad \textbf{(3.2)}$$

Di mana $\hat{y}_i$ adalah nilai prediksi dan y_i adalah nilai aktual. Rentang nilai MAE berada antara 0 hingga tak hingga positif (+∞). Semakin besar nilai MAE, maka semakin besar pula tingkat kesalahan prediksinya.

3.6.2 *Mean Absolute Percentage Error* (MAPE)

MAPE merupakan ukuran sederhana yang digunakan untuk menilai tingkat akurasi suatu metode peramalan. Oleh karena itu, MAPE sering dianggap sebagai indikator evaluasi yang paling adil dan penting dalam menilai tingkat presisi suatu model. Dengan membagi setiap kesalahan terhadap nilai aktual dan menyatakannya dalam bentuk persentase, MAPE dihitung menggunakan persamaan 3.3 (Liu & Sun, 2019).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \quad \textbf{(3.3)}$$

3.6.3 *Root Mean Square Error (RMSE)*

RMSE merupakan salah satu indikator umum yang digunakan untuk mengevaluasi tingkat akurasi suatu model peramalan. RMSE merupakan akar kuadrat dari MSE dan memiliki sensitivitas yang lebih tinggi terhadap kesalahan yang besar dibandingkan MAPE. RMSE dihitung menggunakan persamaan 3.4 (Liu & Sun, 2019).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \quad \textbf{(3.4)}$$

Indikator evaluasi ini digunakan sebagai ukuran akurasi dari prediksi model pembelajaran mesin. Semakin besar nilainya, maka semakin rendah kinerja model tersebut, dan semakin besar pula deviasi antara nilai prediksi dan nilai sebenarnya. Terdapat perbedaan tertentu antara keduanya, yaitu MAE (Mean Absolute Error) lebih tahan terhadap nilai *outliers* dan lebih mencerminkan kinerja prediksi secara rata-rata. Sedangkan RMSE lebih sensitif terhadap kemampuan model dalam

menyesuaikan data yang menyimpang (*abnormal*), sehingga keduanya saling melengkapi (Liu & Sun, 2019).

3.6.4 *Mean Square Error* (MSE),

Mean Squared Error (MSE) merupakan nilai rata-rata dari kuadrat selisih antara nilai yang diprediksi dengan nilai aktual. Dalam metode ini, perbedaan antara prediksi dan data sebenarnya terlebih dahulu dikuadratkan, kemudian dirata-ratakan. MSE sering digunakan sebagai indikator untuk mengevaluasi seberapa baik performa suatu model prediksi. Nilai MSE yang lebih kecil, atau mendekati nol, menunjukkan bahwa hasil prediksi model tersebut semakin akurat dan mendekati nilai sebenarnya. Oleh karena itu, model dengan nilai MSE rendah dianggap layak untuk digunakan dalam melakukan prediksi pada periode berikutnya. Adapun rumus yang digunakan untuk menghitung MSE menggunakan persamaan 3.5 (Hodson dkk., 2021).

$$RMSE = \frac{1}{n}\sum\left(y - \hat{y}\right)^2 \qquad \textbf{(3.5)}$$

3.6.5 *Koefisien determinasi* (R^2)

Koefisien determinasi (R^2) merupakan suatu ukuran statistik yang digunakan untuk mengevaluasi sejauh mana suatu model regresi mampu menjelaskan variasi dari variabel dependen atau target. Nilai R^2 berkisar antara 0 hingga 1. Semakin mendekati nilai 1, maka model dianggap mampu menjelaskan data dengan sangat baik, karena nilai-nilai observasi terletak dekat dengan garis regresi. Sebaliknya, nilai R^2 yang rendah menunjukkan bahwa observasi tersebar jauh dari garis regresi, yang berarti model kurang mampu merepresentasikan data dengan akurat. Rumus yang digunakan untuk menghitung nilai R^2 menggunakan persamaan 3.6 (Ed-Daoudi dkk., 2023).

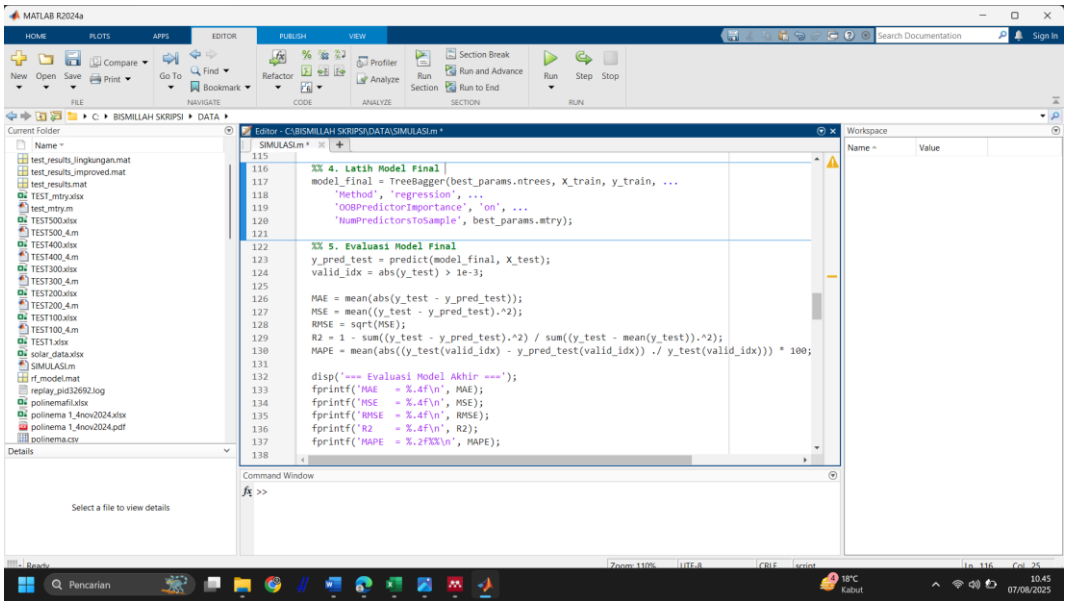
$$R^2 = 1 - \left[\frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \hat{y}_i\right)^2}\right] \qquad \textbf{(3.5)}$$

3.7 Bahan dan Peralatan Penelitian

Penelitian ini menggunakan perangkat laptop dengan processor Intel Celeron N4020, *random acces memory* sebesar 4 GB, *Windows 11,* Selain itu penelitian ini menggunakan *tools* berupa *software* yaitu *Matrix Laboratory* (MATLAB), *Microsoft excel,* dan *Microsoft Word.*

3.8 Pembuatan Model *Random Forest* di Matlab

Dalam MATLAB, algoritma *Random Forest* dapat diterapkan dengan memanfaatkan fungsi *TreeBagger*, yang mendukung implementasi regresi berbasis banyak pohon.



Gambar 3. 2 Tampilan Program *Random Forest*

Beberapa langkah dan perintah penting dalam implementasi *Random Forest* pada MATLAB antara lain:

a. *TreeBagger*: digunakan untuk membuat model *Random Forest* dengan menentukan jumlah pohon (*NumTrees*) dan jumlah prediktor yang dipilih secara acak untuk pemisahan di setiap node (*NumPredictorsToSample* atau *mtry*).

b. *predict*: digunakan untuk memprediksi nilai output (daya keluaran) berdasarkan data input pengujian yang diberikan ke model hasil pelatihan.

c. *OutOfBagError*: mengevaluasi performa model menggunakan data *out-of-bag*, yaitu sampel data yang tidak digunakan selama pelatihan setiap pohon, sebagai bentuk validasi internal.
d. *oobPredict*: menghasilkan prediksi dari data out-of-bag untuk analisis akurasi model.