BAB II
TINJAUAN PUSTAKA

2.1 Kajian Relevan

Lama penyinaran matahari merupakan salah satu parameter klimatologi yang berperan dalam berbagai bidang, seperti pertanian, klimatologi, dan energi surya. Informasi mengenai lama penyinaran matahari dapat digunakan sebagai salah satu parameter pendukung dalam analisis energi surya, khususnya untuk menggambarkan durasi penyinaran matahari pada suatu wilayah. Oleh karena itu, penelitian mengenai prediksi lama penyinaran matahari terus berkembang dengan memanfaatkan berbagai metode statistik, *machine learning*, dan *deep learning*.

Penelitian sebelumnya telah menggunakan berbagai algoritma untuk memprediksi lama penyinaran matahari maupun parameter energi surya lainnya. Beberapa penelitian menggunakan metode statistik seperti regresi linier *dan Singular Spectrum Analysis* (SSA) untuk melakukan prediksi lama penyinaran matahari. Di sisi lain, perkembangan teknologi *machine learning* mendorong penggunaan algoritma seperti *Support Vector Machine* (SVM), *Artificial Neural Network* (ANN), *Multi-Layer Perceptron* (MLP), *Long Short-Term Memory* (LSTM), *Random Forest*, dan XGBoost. Penelitian oleh (Zaenuddin, 2017) menggunakan SVM untuk prediksi lama matahari bersinar di Banjarbaru, sedangkan (Nugroho, 2023) menggunakan model MLP untuk prediksi penyinaran matahari di Kota Bandung. Penelitian lain oleh (Wardani & Andriyani, 2025) menunjukkan bahwa model LSTM memberikan performa terbaik dalam prediksi sunshine duration dibandingkan metode machine learning lainnya.

Selain digunakan untuk prediksi lama penyinaran matahari, algoritma Random Forest dan XGBoost juga banyak diterapkan dalam prediksi parameter energi surya seperti radiasi matahari global. Penelitian oleh (Anwar & Khatib, 2017) menggunakan Random Forest yang dioptimasi dengan Firefly Algorithm untuk prediksi radiasi matahari, sedangkan (Zaim et al., 2023) serta (Obiora & Ali, 2021)

menggunakan XGBoost dalam prediksi radiasi surya dengan hasil performa yang cukup baik. Penelitian (J. Fan et al., 2019) juga menunjukkan bahwa algoritma *machine learning*, termasuk Random Forest dan XGBoost, mampu memberikan performa yang lebih baik dibandingkan model empiris dalam prediksi parameter energi surya.

Berdasarkan tinjauan tersebut, masih terdapat *research gap* dalam penelitian prediksi lama penyinaran matahari. Sebagian besar penelitian terdahulu lebih berfokus pada prediksi radiasi surya, sedangkan penelitian yang secara khusus membandingkan algoritma Random Forest dan XGBoost untuk prediksi lama penyinaran matahari masih relatif terbatas, khususnya di Indonesia. Selain itu, belum banyak penelitian yang mengkaji performa kedua algoritma tersebut pada wilayah tropis dengan potensi energi surya tinggi seperti Kota Kupang. Oleh karena itu, penelitian ini dilakukan untuk membandingkan performa algoritma Random Forest dan XGBoost dalam prediksi lama penyinaran matahari di Kota Kupang menggunakan data meteorologi sebagai variabel prediktor.

Tabel 2. 1 *State of The Art*

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
1.	Wulan Kusuma Wardani dan Ririn Andriyani	2025	Long Short-Term Memory (LSTM), Random Forest, dan Support Vector Machine (SVM).	Input: suhu, kelembapan, curah hujan, kecepatan dan arah angin. Output: sunshine duration.	Data harian Stasiun Meteorologi Radin Inten II Lampung periode 1981–2024 sebanyak 15.887 data.	Model LSTM memberikan performa terbaik dengan RMSE 1,02, MAE 0,77, dan korelasi Pearson 0,84.	Keterkaitan: Sama-sama memprediksi sunshine duration menggunakan data meteorologi. Perbandingan: Penelitian ini menggunakan LSTM, sedangkan penelitian saya membandingkan Random Forest

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							dan XGBoost di Kota Kupang.
2.	Shifa Amelia Rachman dkk.	2025	Penelitian ini menggunakan metode Singular Spectrum Analysis (SSA) dengan teknik R-Forecasting untuk analisis deret waktu.	Variabel tunggal berupa rata-rata lama penyinaran matahari bulanan.	Data bulanan lama penyinaran matahari di Pasuruan periode 2010–2023 sebanyak 168 data.	Model terbaik menggunakan parameter L = 41 dan r = 8 dengan nilai MAPE sebesar 19,53%.	Keterkaitan**:** Sama-sama memprediksi lama penyinaran matahari di Indonesia menggunakan pendekatan time series. Perbandingan: Penelitian tersebut menggunakan metode statistik SSA berbasis univariate, sedangkan

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							penelitian saya menggunakan Random Forest dan XGBoost berbasis machine learning di Kota Kupang.
3.	Zhanhao Zhang dkk.	2026	Kombinasi CatBoost, Transformer, dan model fisik transfer radiasi untuk estimasi SD dan radiasi surya.	Input berupa data satelit, data meteorologi ERA5, ketinggian wilayah, dan radiasi astronomis. Output berupa	Data dari 2.490 stasiun meteorologi dan 170 stasiun radiasi di China periode 2019–2024.	Estimasi SD menghasilkan R = 0,96 dan RMSE 0,11 jam, sedangkan estimasi Rs menghasilkan R = 0,91 dan RMSE 97,31 W/m².	Keterkaitan: Sama-sama memprediksi lama penyinaran matahari menggunakan machine learning. Perbandingan: Penelitian tersebut

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
				sunshine duration (SD) dan radiasi surya (Rs).			menggunakan pendekatan hybrid berbasis satelit dan deep learning skala nasional, sedangkan penelitian saya membandingkan Random Forest dan XGBoost pada wilayah lokal Kota Kupang.
4.	Weiwei Zhu, dkk.	2019	Model empiris berbasis data satelit geostasioner	Input berupa data total cloud amount (CTA),	Data dari 18 stasiun meteorologi di Qinghai, China,	Model menghasilkan performa baik dengan nilai $R^2 \geq 0{,}894$	Keterkaitan: Sama-sama memprediksi lama penyinaran matahari.

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
			menggunakan indeks daytime mean total cloud coverage amount (f_ccom).	waktu matahari terbit-terbenam, dan lokasi geografis. Output berupa sunshine duration harian.	menggunakan data tahun 2016–2018.	dan RMSE ≤ 0,977 jam/hari.	Perbandingan: Penelitian tersebut menggunakan pendekatan empiris berbasis data satelit, sedangkan penelitian saya membandingkan algoritma Random Forest dan XGBoost di Kota Kupang.
5.	Zaenuddin dan Muhammad Rais Wathani	2017	Support Vector Machine (SVM) menggunakan	Variabel utama berupa lama matahari bersinar	Data harian BMKG Banjarbaru tahun 2010–	Model menghasilkan nilai RMSE sebesar 34.038 dan	Keterkaitan: Sama-sama memprediksi lama penyinaran matahari

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
			aplikasi RapidMiner.	rata-rata harian.	2014 sebanyak 1826 data.	dinilai cukup baik untuk prediksi.	menggunakan data historis BMKG. Perbandingan: Penelitian tersebut menggunakan SVM, sedangkan penelitian saya membandingkan Random Forest dan XGBoost di Kota Kupang.
6.	Salma Zaim dkk.	2023	Perbandingan Artificial Neural Network (ANN) dan	Input berupa durasi penyinaran matahari, suhu,	Data meteorologi di Rabat, Maroko tahun 2020–2021 dengan	Model ANN memperoleh performa terbaik dengan R^2	Keterkaitan: Sama-sama menggunakan algoritma machine

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
			XGBoost, dengan Random Forest untuk seleksi fitur.	kelembapan, angin, tekanan udara, dan waktu. Output berupa radiasi matahari global.	interval 30 menit.	0,9805 dan RMSE 45,5 W/m², sedangkan XGBoost juga menunjukkan akurasi tinggi dengan R^2 0,9753.	learning untuk prediksi parameter meteorologi terkait energi surya. Perbandingan: Penelitian tersebut memprediksi radiasi matahari global, sedangkan penelitian saya memprediksi lama penyinaran matahari menggunakan Random Forest

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							dan XGBoost di Kota Kupang.
7.	Junliang Fan dkk.	2019	Perbandingan 12 model empiris dan 12 algoritma machine learning, termasuk XGBoost, Random Forest, SVM, dan ANFIS.	Input berupa rasio durasi penyinaran matahari dan radiasi ekstra-terestrial, dengan output radiasi surya global harian.	Data dari 50 stasiun cuaca di China periode 1966–2015.	Model machine learning menunjukkan performa lebih baik dibanding model empiris. ANFIS menjadi model terbaik, sedangkan XGBoost memiliki efisiensi	Keterkaitan: Sama-sama mengevaluasi algoritma Random Forest dan XGBoost pada bidang meteorologi dan energi surya. Perbandingan: Penelitian tersebut memprediksi radiasi surya, sedangkan penelitian saya memprediksi

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
						komputasi tertinggi.	lama penyinaran matahari di Kota Kupang.
8.	Habiballa A. Mohamed	2020	Regresi linier sederhana untuk memprediksi lama penyinaran matahari dan radiasi surya.	Input berupa lintang, bujur, dan ketinggian wilayah. Output berupa bright sunshine duration dan solar radiation.	Data dari Sudan Meteorological Authority dengan rentang periode 1951–2000 dari beberapa stasiun meteorologi.	Model menunjukkan korelasi tinggi dengan nilai R hingga 0,93 pada kombinasi variabel geografis.	Keterkaitan: Sama-sama memprediksi lama penyinaran matahari. Perbandingan: Penelitian tersebut menggunakan regresi linier berbasis faktor geografis, sedangkan penelitian saya menggunakan Random Forest

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							dan XGBoost di Kota Kupang.
9.	Priambudi Agung Nugroho	2023	Jaringan Syaraf Tiruan (JST) dengan arsitektur Multi-Layer Perceptron (MLP).	Input berupa data historis penyinaran matahari, sedangkan output berupa prediksi lama penyinaran matahari.	Data BMKG dan BPS Kota Bandung periode 2016–2021 sebanyak 72 data.	Model menghasilkan nilai MSE sebesar 0,086182 pada pelatihan dan 0,10921 pada pengujian.	Keterkaitan: Sama-sama memprediksi lama penyinaran matahari menggunakan kecerdasan buatan. Perbandingan: Penelitian tersebut menggunakan model MLP di Kota Bandung, sedangkan penelitian saya membandingkan

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							Random Forest dan XGBoost di Kota Kupang.
10.	Ibrahim Anwar Ibrahim dan Tamer Khatib	2017	Model hybrid Random Forest dan Firefly Algorithm untuk optimasi parameter model.	Input berupa sunshine ratio, suhu, kelembapan, waktu, dan bulan. Output berupa radiasi matahari global per jam.	Data meteorologi per jam dari Stasiun Meteorologi Subang, Malaysia sebanyak 61.320 sampel selama satu tahun.	Model RF-FFA menghasilkan RMSE 18,98%, MAPE 6,38%, dan MBE 2,86%.	Keterkaitan: Sama-sama menggunakan Random Forest untuk prediksi parameter energi surya. Perbandingan: Penelitian tersebut memprediksi radiasi matahari dengan optimasi Firefly Algorithm, sedangkan

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							penelitian saya membandingkan Random Forest dan XGBoost untuk prediksi lama penyinaran matahari di Kota Kupang.
11.	Xin Liu dkk.	2026	Perbandingan RF, XGBoost, LSTM, dan GRU dengan model utama Hybrid RF-XGBoost.	Input berupa suhu udara dalam dan luar ruangan serta intensitas radiasi matahari. Output berupa suhu tanah.	Data rumah kaca surya di China periode Desember 2018–Februari 2019 sebanyak 7.681 data dengan interval 15 menit.	Model Hybrid RF-XGBoost menghasilkan performa terbaik dengan R^2 0,9927 dan RMSE 0,1605.	Keterkaitan: Sama-sama membandingkan performa Random Forest dan XGBoost. Perbandingan: Penelitian tersebut memprediksi suhu tanah pada

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							rumah kaca, sedangkan penelitian saya memprediksi lama penyinaran matahari di Kota Kupang.
12.	Chibuzor N. Obiora, dkk.	2021	XGBoost regression yang dibandingkan dengan Support Vector Machine (SVM).	Input berupa suhu, kelembapan, dan sunshine duration. Output berupa solar irradiance (GHI).	Data per jam kota Johannesburg, Afrika Selatan periode 2015–2019 dengan hingga 22.500 observasi.	Model XGBoost memberikan performa terbaik pada dataset 1 tahun dengan nilai nRMSE 6,63%, MAE 2,45, dan RMSE 28,61.	Keterkaitan: Sama-sama menggunakan XGBoost untuk prediksi parameter energi surya berbasis data meteorologi. Perbandingan: Penelitian tersebut

No	Peneliti	Tahun	Metode	Fitur/Var	Dataset	Akurasi	Keterkaitan
							memprediksi solar irradiance, sedangkan penelitian saya membandingkan Random Forest dan XGBoost untuk prediksi lama penyinaran matahari di Kota Kupang.

Berdasarkan tinjauan literatur, sebagian besar penelitian pada bidang energi surya masih berfokus pada prediksi radiasi surya global (*Global Solar Radiation*/GHI), sedangkan penelitian yang secara khusus memprediksi lama penyinaran matahari (*sunshine duration*) masih relatif terbatas. Beberapa penelitian telah menggunakan metode statistik, *machine learning*, maupun *deep learning* seperti SSA, SVM, MLP, LSTM, Random Forest, dan XGBoost dengan hasil performa yang cukup baik. Namun, sebagian besar penelitian menggunakan *sunshine duration* sebagai variabel masukan untuk memprediksi radiasi surya, bukan sebagai variabel target utama. Selain itu, penelitian yang secara khusus membandingkan performa algoritma Random Forest dan XGBoost dalam prediksi lama penyinaran matahari masih jarang ditemukan, terutama pada wilayah Indonesia.

Kontribusi penelitian ini terletak pada penggunaan pendekatan *machine learning* berbasis *ensemble learning* untuk memprediksi lama penyinaran matahari di Kota Kupang. Penelitian ini menggunakan beberapa parameter meteorologi sebagai variabel prediktor dan membandingkan secara langsung performa algoritma Random Forest dan XGBoost untuk mengetahui algoritma dengan performa terbaik dalam memprediksi lama penyinaran matahari.

Penelitian ini diharapkan dapat mengisi kesenjangan penelitian terkait prediksi *sunshine duration* menggunakan algoritma *ensemble learning*, serta memberikan informasi pendukung dalam analisis potensi energi surya di wilayah tropis, khususnya Kota Kupang.

2.2 Landasan Teori

Landasan teori memuat berbagai konsep dan teori yang digunakan sebagai dasar dalam pelaksanaan penelitian. Pembahasan meliputi konsep energi surya dan lama penyinaran matahari, *machine learning*, regresi, *ensemble learning*, algoritma *Random Forest* dan *Extreme Gradient Boosting* (XGBoost), metrik evaluasi model regresi, metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM),

serta penggunaan Python dan *flowchart* dalam proses penelitian. Teori-teori tersebut disusun berdasarkan berbagai referensi ilmiah yang relevan dan digunakan sebagai landasan dalam membangun, mengevaluasi, serta membandingkan model prediksi lama penyinaran matahari di Kota Kupang.

2.2.1 Energi Surya

Energi surya merupakan sumber energi yang melimpah dan ramah lingkungan yang dapat dikonversi menjadi energi panas maupun energi listrik untuk berbagai kebutuhan manusia. Potensi matahari sebagai sumber energi yang tidak terbatas menjadikannya salah satu alternatif utama dalam diversifikasi energi global serta berperan penting dalam mengurangi ketergantungan terhadap bahan bakar fosil dan dampak negatif terhadap lingkungan (Dwisari et al., 2023). Pemanfaatan energi surya saat ini berkembang pesat karena termasuk energi terbarukan yang tersedia secara berkelanjutan dan tidak menghasilkan emisi gas rumah kaca secara langsung selama proses pembangkitan energi (Kadang & Windarta, 2021).

Energi surya dapat dimanfaatkan melalui teknologi fotovoltaik (Photovoltaic/PV) untuk menghasilkan energi listrik maupun teknologi termal untuk menghasilkan energi panas. Salah satu bentuk pemanfaatannya adalah melalui Pembangkit Listrik Tenaga Surya (PLTS), yaitu sistem pembangkit yang mengubah radiasi matahari menjadi energi listrik menggunakan panel surya. Pengembangan PLTS terus meningkat seiring meningkatnya kebutuhan energi bersih dan upaya transisi menuju energi terbarukan (Kadang & Windarta, 2021).

Indonesia memiliki potensi energi surya yang sangat besar karena berada di wilayah tropis dan dilalui garis khatulistiwa sehingga menerima intensitas penyinaran matahari yang tinggi sepanjang tahun. (Kementerian ESDM, 2022) menyebutkan bahwa potensi energi surya di Indonesia mencapai lebih dari 3.000 GW, namun pemanfaatannya masih relatif kecil dibandingkan potensi yang tersedia. Oleh karena itu, pemerintah terus mendorong pengembangan PLTS, khususnya PLTS Atap,

sebagai bagian dari upaya mencapai target bauran energi baru terbarukan nasional (Kementerian ESDM, 2022).

Kinerja Pembangkit Listrik Tenaga Surya (PLTS) dipengaruhi oleh beberapa parameter meteorologi seperti intensitas radiasi matahari, suhu udara, kelembapan, dan lama penyinaran matahari. Lama penyinaran matahari menjadi indikator penting karena berkaitan dengan jumlah energi surya yang diterima panel surya dalam satu periode waktu. Semakin lama durasi penyinaran matahari, maka energi listrik yang dihasilkan PLTS cenderung semakin besar sehingga informasi mengenai parameter tersebut penting dalam proses perencanaan dan optimalisasi sistem energi surya (Sri Ulina, 2023).

2.2.2 Lama Penyinaran Matahari (*Sunshine Duration*)

Lama penyinaran matahari merupakan durasi waktu di mana wilayah tertentu menerima sinar matahari yang berpengaruh terhadap berbagai aktivitas manusia dan potensi wilayah, seperti pertanian, kesehatan, dan pembangkit listrik tenaga surya (Nugroho, 2023). Parameter ini menunjukkan durasi ketika intensitas radiasi matahari langsung melebihi ambang batas tertentu sehingga dapat terdeteksi oleh alat pengukur meteorologi. Lama penyinaran matahari menjadi salah satu unsur klimatologi penting yang sering digunakan dalam bidang meteorologi, pertanian, hidrologi, serta energi terbarukan karena berkaitan erat dengan jumlah energi surya yang diterima suatu wilayah.

Pengukuran durasi sinar matahari biasanya dilakukan dengan menggunakan alat *Campbell-Stokes Sunshine Recorder* atau sensor radiasi otomatis di stasiun cuaca modern. Data hasil pengukuran biasanya dinyatakan dalam satuan jam per hari atau persentase waktu penyinaran terhadap total durasi penyinaran matahari maksimum dalam satu hari. Di Indonesia, data tentang penyinaran matahari yang sudah lama tersedia diberikan oleh Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) sebagai bagian dari data klimatologi harian dan bulanan yang digunakan untuk keperluan menganalisis cuaca dan iklim (BMKG, 2024c).

Lama penyinaran matahari dipengaruhi oleh beberapa faktor, seperti awan yang menutupi, hujan, udara yang lembap, suhu udara, musim, dan lokasi geografis suatu daerah. Wilayah yang memiliki banyak awan cenderung mendapatkan sinar matahari lebih sedikit karena sinar matahari terhalang sebelum sampai ke permukaan bumi. Selain itu, perubahan iklim setiap hari serta lokasi daerah yang berada di lintang tertentu juga memengaruhi jumlah waktu matahari yang menerpa setiap hari (Xing & Xu, 2019).

Di bidang energi terbarukan, durasi sinar matahari berpengaruh besar terhadap kemampuan menghasilkan energi dari sinar matahari dan efisiensi kerja Pembangkit Listrik Tenaga Surya. Semakin lama sinar matahari menyinari suatu daerah, semakin besar kemungkinan energi listrik yang bisa dihasilkan oleh panel surya. Oleh karena itu, informasi dan prediksi tentang pola penyinaran matahari yang lama sangat penting dalam proses merencanakan, mengoperasikan, dan mengoptimalkan sistem PLTS agar dapat meningkatkan penggunaan energi surya secara lebih efisien (Wardani & Andriyani, 2025).

2.2.3 *Machine Learning*

Machine Learning adalah bagian dari kecerdasan buatan (*Artificial Intelligence*/AI) yang memungkinkan komputer belajar dari data, mengenali pola, dan membuat prediksi atau keputusan tanpa perlu diinstruksikan secara rinci (Mujtaba, 2024). Metode ini berjalan dengan menggunakan algoritma untuk memahami hubungan antar data, sehingga model bisa meningkatkan kemampuannya berdasarkan data yang sudah ada sebelumnya. *Machine learning* digunakan di berbagai bidang seperti kesehatan, ekonomi, cuaca, dan energi karena kemampuannya untuk memproses data dalam jumlah besar serta pola yang rumit (Mujtaba, 2024).

Machine Learning umumnya dibagi menjadi tiga kategori besar, yaitu *supervised learning, unsupervised learning,* dan *reinforcement learning*. *Supervised learning* adalah metode pembelajaran yang menggunakan data yang sudah memiliki

label atau hasil yang dicari, sehingga model bisa belajar bagaimana memahami hubungan antara data yang diberikan dan hasil yang diharapkan. Metode ini sering digunakan untuk mengelompokkan dan memperkirakan sesuatu, contohnya memperkirakan kondisi cuaca, jumlah radiasi matahari, serta durasi sinar matahari yang diterima. *Unsupervised learning* digunakan pada data yang tidak memiliki label untuk menemukan pola atau kelompok tertentu dalam data, seperti pada proses pengelompokan atau pemisahan data. Sementara itu, *reinforcement learning* adalah metode pembelajaran di mana sistem belajar dengan cara melakukan percobaan dan menerima umpan balik berupa hadiah atau hukuman untuk memutuskan tindakan terbaik secara otomatis (Chyan et al., 2024).

Dalam bidang klimatologi dan energi surya, *machine learning* digunakan untuk memperkirakan berbagai parameter cuaca seperti radiasi matahari, suhu udara, kecepatan angin, serta durasi sinar matahari yang terlihat. Penggunaan algoritma *Machine Learning* dianggap mampu menghasilkan prediksi yang lebih tepat dibandingkan metode statistik biasa, karena mampu memahami hubungan yang tidak lurus antar variabel cuaca. Selain itu, metode ini juga bisa digunakan dengan baik di daerah yang tidak memiliki data pengamatan radiasi matahari secara langsung (Zaim et al., 2023).

2.2.4 Regresi

Regresi merupakan salah satu teknik dalam *supervised machine learning* yang bertujuan memodelkan hubungan antara variabel dependen (variabel target) dengan satu atau lebih variabel independen (variabel prediktor) untuk menghasilkan prediksi berupa nilai numerik yang kontinu (Rashidi et al., 2023). Dalam *supervised machine learning*, model dilatih menggunakan data berlabel yang telah diketahui nilai keluarannya, sehingga model dapat mempelajari pola hubungan antara fitur masukan dan nilai target. Berbeda dari klasifikasi yang menghasilkan keluaran berupa kategori diskrit, regresi secara khusus difokuskan pada estimasi nilai kontinu seperti suhu, kelembaban, radiasi surya, maupun durasi penyinaran matahari (Segovia et al., 2023).

Dalam penelitian ini, prediksi lama penyinaran matahari (*sunshine duration*/SS) di Kota Kupang dikategorikan sebagai masalah regresi, karena variabel SS merupakan nilai riil positif dalam satuan jam per hari yang bersifat kontinu.

Tujuan utama model regresi adalah membangun persamaan matematis yang mendefinisikan variabel keluaran (y) sebagai fungsi dari satu atau lebih variabel prediktor (x), sehingga persamaan tersebut dapat digunakan untuk memprediksi keluaran berdasarkan nilai-nilai masukan yang baru (Rashidi et al., 2023). Secara matematis, model regresi berupaya menemukan fungsi f yang memenuhi:

$$\hat{y} \approx f(x) \tag{2.1}$$

Di mana ŷ merupakan nilai prediksi lama penyinaran matahari dan x merupakan kumpulan variabel prediktor yang digunakan dalam model.

Metode regresi banyak diterapkan dalam bidang klimatologi dan energi terbarukan karena mampu memprediksi parameter cuaca dan energi surya secara kuantitatif. Berbagai algoritma regresi seperti *Random Forest Regression*, *Support Vector Regression* (SVR), dan *XGBoost Regression* telah digunakan untuk memprediksi radiasi matahari maupun lama penyinaran matahari dengan tingkat akurasi yang baik (J. Fan et al., 2019). Oleh karena itu, pendekatan regresi dipilih pada penelitian ini untuk membandingkan performa algoritma Random Forest dan XGBoost dalam memprediksi lama penyinaran matahari di Kota Kupang.

2.2.5 *Ensemble Learning*

Ensemble learning merupakan metode dalam *machine learning* yang menggabungkan beberapa model prediksi (*base learner*) untuk menghasilkan performa yang lebih baik dibandingkan model tunggal (*single learner*). Pendekatan ini bertujuan meningkatkan akurasi, stabilitas, dan kemampuan generalisasi model dengan memanfaatkan kombinasi beberapa algoritma sehingga kesalahan prediksi dapat diminimalkan. Metode *ensemble learning* banyak digunakan dalam berbagai bidang prediksi karena mampu menangani hubungan data yang kompleks dan non-linear (Y. Zhang et al., 2022).

Terdapat dua pendekatan utama dalam *ensemble learning*, yaitu:

1. *Bagging (Bootstrap Aggregating)*

   Bagging bekerja dengan melatih beberapa model dasar secara paralel menggunakan subset data yang berbeda, yang diambil secara acak melalui proses *bootstrap*. Hasil akhir diperoleh dengan merata-ratakan prediksi (regresi) atau melalui pemungutan suara (klasifikasi) untuk meningkatkan stabilitas dan mengurangi *overfitting*. Contoh algoritma utamanya adalah *Random Forest* (Z. Fan et al., 2025).

2. *Boosting*

   Boosting adalah metode sekuensial di mana model-model dasar dilatih secara bertahap untuk memperbaiki kesalahan model sebelumnya. Algoritma ini memberikan bobot lebih besar pada data yang salah diklasifikasikan agar menjadi fokus bagi model berikutnya. Prediksi akhir ditentukan melalui pemungutan suara atau rata-rata berbobot untuk menghasilkan satu model yang sangat akurat. Contoh populer meliputi AdaBoost dan XGBoost (Z. Fan et al., 2025).

Dalam konteks penelitian ini, penggunaan *ensemble learning* menjadi relevan karena kemampuannya dalam menangani data dimensi tinggi dan meningkatkan stabilitas prediksi pada variabel-variabel yang memiliki fluktuasi tinggi.

2.2.6 *Random Forest Regressor*

Menurut Breiman (2001), dalam (Schonlau & Zou, 2020), algoritma *Random Forest* merupakan sebuah metode pembelajaran statistik atau pembelajaran mesin berbasis ansambel yang dirancang untuk melakukan tugas prediksi, baik dalam pemodelan klasifikasi dengan hasil kategoris maupun pemodelan regresi dengan hasil kontinu. Algoritma ini bekerja dengan menggabungkan banyak *decision tree* untuk menghasilkan prediksi yang lebih stabil dan akurat dibandingkan penggunaan satu pohon keputusan tunggal (*single decision tree*).

Random Forest termasuk metode *bagging* (*bootstrap aggregating*), yaitu teknik yang membangun sejumlah model dari sampel data berbeda, kemudian menggabungkan hasil prediksi seluruh model tersebut menjadi satu hasil akhir (Shibaprasad Bhattacharya, 2023). Dalam implementasinya, *Random Forest* menciptakan keragaman antar model tersebut melalui tiga mekanisme utama berikut:

1. *Bootstrap Sampling* (Keragaman Data Pelatihan)

   Setiap pohon dilatih menggunakan sampel data acak yang diambil dengan pengembalian (*with replacement*) dari dataset asli. Proses ini menciptakan kombinasi data pelatihan yang unik untuk setiap pohon dan menyisakan sepertiga data sebagai *out-of-bag* (OOB) untuk mendeteksi kesalahan prediksi (Anwar & Khatib, 2017).

2. *Feature Randomness* (Keragaman Struktur Pohon)

   Setelah data disampel, setiap pohon keputusan akan dibangun secara independen. Pada proses pembangunan ini, *Random Forest Regressor* membatasi pemilihan fitur secara acak (*Feature Randomness*) pada setiap simpul pemisahan (*split node*) (Anwar & Khatib, 2017).

   Dalam menentukan titik potong (*threshold*) terbaik pada setiap simpul, algoritma menggunakan kriteria *Sum of Squared Errors* (SSE) dengan meminimalkan total galat kuadrat dari hasil pembagian data (López et al., 2022):

   $$\text{SSE} = \sum_{i \in R_1} (y_i - \overline{y_1})^2 + \sum_{i \in R_2} (y_i - \overline{y_2})^2 \qquad (2.2)$$

   Keterangan:

   - R_1, R_2 = Wilayah (*region*) data hasil pembagian simpul biner
   - y_i = Nilai aktual lama penyinaran matahari pada data ke-i
   - $\overline{y_1}, \overline{y_2}$ = Rata-rata nilai target pada masing-masing wilayah.

3. *Aggregation* (Penyatuan Prediksi)

Setelah semua pohon selesai dilatih secara independen, hasil prediksi akhir untuk tugas regresi diperoleh dengan menghitung nilai rata-rata dari seluruh prediksi pohon di dalam *forest*. Penyatuan ini menghasilkan prediksi final yang lebih stabil, akurat, dan minim varians (Anwar & Khatib, 2017). Secara matematis, prediksi *Random Forest Regressor* dapat dirumuskan sebagai berikut (Aji et al., 2025):

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \tag{2.3}$$

Keterangan:

- $\hat{y}$ = Prediksi akhir
- B = Jumlah pohon (*trees*)
- $h_b(x)$ = Prediksi dari pohon ke-b
- $\sum_{b=1}^{B}$ = Penjumlahan semua prediksi
- $\frac{1}{B}$ = Rata-rata aritmetika

Untuk menghasilkan performa prediksi yang optimal, *Random Forest Regressor* memerlukan pengaturan beberapa *hyperparameter* penting pada proses pembentukan model:

a. *n_estimators:*

   Parameter yang menentukan jumlah *decision tree* yang dibangun dalam model *Random Forest*. Semakin banyak pohon yang digunakan, model cenderung menghasilkan prediksi yang lebih stabil.

b. *max_depth:*

   Parameter yang menentukan kedalaman maksimum setiap *decision tree*. Parameter ini digunakan untuk mengontrol kompleksitas model dan membantu mengurangi risiko *overfitting*.

c. *max_featurers:*

Parameter yang menentukan jumlah fitur yang dipilih secara acak pada setiap proses pembentukan *split node* sehingga meningkatkan keragaman antar pohon.

d. *min_samples_split:*

Parameter yang menentukan jumlah minimum sampel yang diperlukan untuk membagi suatu node pada proses pembentukan pohon keputusan (Muhamad & Matin, 2023).

Random Forest memiliki beberapa kelebihan dalam proses prediksi, di antaranya mampu menangani data dalam jumlah besar, bekerja baik pada data multivariat, mengurangi risiko *overfitting*, serta dapat menentukan tingkat kepentingan variabel (*feature importance*). Selain itu, algoritma ini efektif dalam menangani hubungan data yang bersifat non-linear sehingga banyak digunakan pada prediksi meteorologi dan energi surya (Ardianto, Agus Budi Raharjo, 2022; Haryoko et al., 2026).

Namun demikian, Random Forest juga memiliki beberapa kekurangan, seperti kebutuhan komputasi yang lebih tinggi ketika jumlah pohon semakin banyak serta interpretasi model yang relatif lebih sulit dibandingkan metode regresi sederhana (R. C. Chen et al., 2020).

2.2.7 *Extreme Gradient Boosting (*XGBoost*) Regressor*

Extreme Gradient Boosting (XGBoost) merupakan implementasi dari algoritma gradient boosting yang dikembangkan secara khusus untuk mengoptimalkan kecepatan komputasi dan performa prediktif. Algoritma ini pertama kali diperkenalkan oleh Chen dan Guestrin melalui makalah ilmiah berjudul XGBoost: *A Scalable Tree Boosting System* yang dipresentasikan pada konferensi KDD tahun 2016. Sejak diperkenalkan, XGBoost menjadi salah satu algoritma machine learning yang banyak digunakan dalam kompetisi ilmu data maupun berbagai aplikasi industri karena kemampuannya dalam menangani data tabular berskala besar secara efisien (T. Chen & Guestrin, 2016).

XGBoost bekerja dengan pendekatan *ensemble learning* berbasis *gradient boosting*, yaitu membangun sejumlah pohon keputusan (*decision tree*) secara bertahap. Pada setiap iterasi, pohon baru dibentuk untuk memperbaiki kesalahan prediksi yang dihasilkan oleh pohon sebelumnya, sehingga model secara bertahap mampu menghasilkan prediksi yang semakin akurat dibandingkan penggunaan satu pohon keputusan saja (Idham & Yosrita, 2025).

Menurut (Idham & Yosrita, 2025), XGBoost memiliki kemampuan yang baik dalam menangani data yang mengandung *noise* dan *outlier*. Selain itu, algoritma ini juga mampu memodelkan hubungan yang kompleks dan nonlinier antarvariabel, sehingga sesuai untuk diterapkan pada data deret waktu (*time series*) yang umumnya memiliki pola dinamis dan fluktuatif.

Dalam penelitian ini, XGBoost digunakan sebagai model regresi untuk memprediksi lama penyinaran matahari di Kota Kupang. Kemampuannya dalam memodelkan hubungan nonlinier antarvariabel meteorologi serta mekanisme regularisasi yang dimilikinya untuk mengurangi risiko *overfitting* menjadikan algoritma ini sebagai salah satu metode yang potensial dalam menghasilkan prediksi nilai kontinu (Zaim et al., 2023).

a. Fungsi Prediksi Aditif

   XGBoost menggunakan konsep model aditif, yaitu prediksi akhir diperoleh dari penjumlahan seluruh *decision tree* yang dibangun secara bertahap. Secara matematis dirumuskan sebagai berikut (T. Chen & Guestrin, 2016):

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{2.4}$$

   Keterangan:

   - $\hat{y}_i$: nilai prediksi data ke-i
   - K: jumlah total pohon keputusan

- $f_k(x_i)$: hasil prediksi pohon ke-k
- $\mathcal{F}$: himpunan fungsi pohon regresi

b. Fungsi Objektif

Pelatihan XGBoost bertujuan meminimalkan fungsi objektif yang terdiri dari *loss function* dan regularisasi. Fungsi ini digunakan untuk mengukur kesalahan prediksi sekaligus mengontrol kompleksitas model (T. Chen & Guestrin, 2016).

$$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.5}$$

Keterangan:

- $l(y_i, \hat{y}_i)$: fungsi kerugian (*loss function*)
- $\Omega(f_k)$: fungsi regularisasi pada pohon ke-k

Untuk kasus regresi, fungsi kerugian yang umum digunakan adalah *Mean Squared Error* (MSE):

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \tag{2.6}$$

c. Regularisasi

Salah satu keunggulan utama XGBoost adalah adanya regularisasi eksplisit yang digunakan untuk mengurangi kompleksitas model dan mencegah *overfitting*. Fungsi regularisasi dirumuskan sebagai berikut (T. Chen & Guestrin, 2016):

$$\Omega(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} w_{kj}^2 \tag{2.7}$$

Keterangan:

- T_k : jumlah daun (*leaves*) pada pohon ke-k

- w_{kj}: bobot pada daun ke-j dari pohon ke-k
- γ: parameter penalti terhadap jumlah daun pada pohon
- λ: parameter regularisasi L2
- $\sum_{j=1}^{T_k} w_{kj}^2$: jumlah kuadrat bobot seluruh daun pada pohon ke-k

d. Ekspansi Taylor Orde Kedua

XGBoost menggunakan pendekatan ekspansi Taylor orde kedua untuk mengoptimalkan fungsi objektif. Pada iterasi ke-t, prediksi model diperbarui sebagai berikut:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2.8}$$

Fungsi objektif pada iterasi ke-t dituliskan sebagai:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{2.9}$$

Untuk mengoptimalkan fungsi objektif ini secara efisien, XGBoost menggunakan pendekatan ekspansi deret Taylor orde kedua. Dengan mengasumsikan $\varepsilon = f_t(x_i)$ sebagai besaran kecil yang ditambahkan, ekspansi Taylor orde kedua dari $l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$ di sekitar $\hat{y}_i^{(t-1)}$ adalah (T. Chen & Guestrin, 2016):

$$l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) \approx l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \tag{2.10}$$

Keterangan:

- g_i: gradien orde pertama
- h_i: gradien orde kedua (*Hessian*)

Gradien dan Hessian dirumuskan sebagai berikut:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{2.11}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{2.12}$$

Dengan menghilangkan suku-suku yang konstan terhadap f_t, fungsi objektif yang disederhanakan menjadi:

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{2.13}$$

Penggunaan ekspansi Taylor orde kedua ini memberikan keunggulan penting: informasi kelengkungan (*curvature*) dari fungsi kerugian dimanfaatkan secara penuh, sehingga optimisasi menjadi lebih presisi dan konvergen lebih cepat dibandingkan metode *gradient boosting* tradisional yang hanya menggunakan gradien orde pertama (T. Chen & Guestrin, 2016).

Dengan mendefinisikan $I_j = \{i \mid q(x_i) = j\}$ sebagai himpunan indeks sampel yang berada pada daun ke-*j*, dan mensubstitusikan definisi regularisasi, fungsi objektif menjadi:

$$Obj^{(t)} = \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{2.14}$$

Bobot optimal w_j* untuk setiap daun *j* diperoleh dengan menurunkan fungsi objektif terhadap w_j dan menyamakan hasilnya dengan nol, sehingga diperoleh:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{2.15}$$

Dan nilai fungsi objektif optimal yang dicapai dengan bobot tersebut adalah:

$$Obj^* = -\frac{1}{2}\sum_{j=1}^{T}\left[\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda}\right] + \gamma T \tag{2.16}$$

Nilai Obj* ini digunakan sebagai skor kualitas pohon (*tree quality score*). Semakin kecil nilainya, semakin baik pohon tersebut dalam meminimalkan fungsi objektif.

e. Perhitungan Gain untuk Split

Pada proses pembentukan pohon, XGBoost menentukan pemisahan (*split*) terbaik menggunakan nilai *Gain*. Semakin besar nilai *Gain*, maka semakin baik kualitas pemisahan node tersebut (T. Chen & Guestrin, 2016).

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{2.17}$$

Keterangan:

- G_L, G_R: jumlah gradien pada node kiri dan kanan
- H_L, H_R: jumlah Hessian pada node kiri dan kanan
- λ: parameter regularisasi
- γ: penalti penambahan daun

Meskipun XGBoost dikembangkan dari *Gradient Boosting Machine* (GBM) yang diperkenalkan oleh (Friedman, 2001), XGBoost memiliki sejumlah pengembangan yang membuat performanya lebih unggul dalam hal akurasi, efisiensi komputasi, dan kemampuan generalisasi. Berbeda dengan GBM tradisional yang belum memiliki regularisasi eksplisit, XGBoost menambahkan regularisasi L1 dan L2 serta penalti jumlah daun (*gamma*) ke dalam fungsi objektif sehingga mampu mengurangi risiko *overfitting* dan meningkatkan stabilitas model (P. Zhang et al., 2022).

Dari sisi optimisasi, GBM tradisional hanya memanfaatkan gradien orde pertama, sedangkan XGBoost menggunakan gradien orde pertama dan kedua (*Hessian*) melalui ekspansi Taylor sehingga proses optimasi menjadi lebih akurat dan

konvergen lebih cepat (T. Chen & Guestrin, 2016). Selain itu, XGBoost mendukung komputasi paralel, *column subsampling*, serta mekanisme penanganan *missing values* secara otomatis sehingga lebih efisien dan praktis untuk digunakan pada dataset berukuran besar maupun kompleks (Zaim et al., 2023).

Performa XGBoost dipengaruhi oleh beberapa hyperparameter penting, antara lain:

- *learning_rate* (eta)
  Mengatur kontribusi setiap pohon baru terhadap model akhir.
- *n_estimators*
  Menentukan jumlah total decision tree yang dibangun.
- *max_depth*
  Menentukan kedalaman maksimum setiap pohon keputusan.
- *Subsample*
  Menentukan proporsi data pelatihan yang digunakan pada setiap iterasi.
- *colsample_bytree*
  Menentukan proporsi fitur yang dipilih secara acak pada setiap pohon.
- *Gamma*
  Menentukan minimum penurunan loss agar node dapat dibagi.
- *lambda dan alpha*
  Parameter regularisasi L2 dan L1 untuk mengurangi kompleksitas model (T. Chen & Guestrin, 2016).

Dari perspektif kelebihan, XGBoost unggul dalam hal akurasi prediksi yang tinggi, kemampuan menangani berbagai tipe *loss function*, resistensi terhadap *overfitting* berkat regularisasi eksplisit, efisiensi komputasi dengan dukungan paralel dan GPU, kemampuan menangani *missing values* secara otomatis, serta menyediakan informasi *feature importance* yang dapat diinterpretasikan (Rathee et al., 2026; Wang & Ni, n.d.; P. Zhang et al., 2022). Di sisi lain, XGBoost memiliki keterbatasan berupa

banyaknya hiperparameter yang perlu dikalibrasi, yang membutuhkan proses *tuning* yang lebih kompleks dan intensif secara komputasi. Selain itu, seperti kebanyakan metode *ensemble*, XGBoost kurang dapat diinterpretasikan secara langsung dibandingkan model sederhana seperti regresi linier (Abdurrahman et al., 2022).

2.2.8 Metrik Evaluasi

Metrik evaluasi merupakan suatu ukuran matematis yang digunakan untuk menilai seberapa dekat hasil prediksi model dengan nilai aktual. Evaluasi model regresi bertujuan untuk mengukur *goodness of fit* atau tingkat kecocokan model terhadap data, sehingga dapat diketahui seberapa baik performa model dalam melakukan prediksi. Penggunaan metrik evaluasi juga membantu dalam membandingkan performa beberapa model regresi secara objektif (Plevris, 2022). Dalam penelitian ini, evaluasi performa model dilakukan menggunakan tiga metrik, yaitu *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan *R-Squared* (R^2).

1. *Mean Absolute Error* (MAE)

   Mean Absolute Error (MAE) merupakan salah satu metrik evaluasi regresi yang digunakan untuk menghitung rata-rata selisih absolut antara nilai aktual dan nilai prediksi model. Metrik ini memberikan gambaran tingkat kesalahan prediksi tanpa memperhatikan arah kesalahan, sehingga hasil evaluasi lebih mudah diinterpretasikan. Nilai MAE yang semakin kecil menunjukkan bahwa model memiliki tingkat kesalahan yang rendah dan kemampuan prediksi yang semakin baik. Selain itu, MAE relatif tidak terlalu sensitif terhadap *outlier* dibandingkan metrik berbasis kuadrat error. Secara matematis, MAE dirumuskan sebagai berikut (Wardani & Andriyani, 2025):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.18}$$

   Keterangan:

- n : jumlah total data
- y_i: nilai aktual hasil observasi
- $\hat{y}_i$: nilai hasil prediksi model

2. *Root Mean Squared Error* (RMSE)

   Selain MAE, penelitian ini juga menggunakan *Root Mean Squared Error* (RMSE) sebagai metrik evaluasi model regresi. RMSE digunakan untuk mengukur akar kuadrat dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Metrik ini memiliki sensitivitas yang tinggi terhadap kesalahan prediksi yang besar karena proses perhitungannya menggunakan kuadrat error. Oleh karena itu, RMSE lebih sensitif terhadap keberadaan kesalahan prediksi yang besar atau ekstrem. Semakin kecil nilai RMSE, maka kemampuan model dalam meminimalkan kesalahan prediksi semakin baik. Secara matematis, RMSE dapat dituliskan sebagai berikut (Wardani & Andriyani, 2025):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.19}$$

   Keterangan:

   - y_i: nilai aktual hasil observasi
   - $\hat{y}_i$: nilai hasil prediksi model
   - n: jumlah total data

3. Koefisien Determinasi (R^2)

   Koefisien determinasi atau *R-Squared* (R^2) merupakan metrik evaluasi yang digunakan untuk mengukur kemampuan model dalam menjelaskan variasi data target. Nilai R^2 menunjukkan seberapa besar proporsi variasi variabel dependen yang dapat dijelaskan oleh variabel independen dalam model regresi. Nilai R^2 berada pada rentang 0 hingga 1. Semakin mendekati nilai 1, maka model dianggap memiliki kemampuan prediksi

yang semakin baik. Sebaliknya, nilai yang mendekati 0 menunjukkan bahwa model belum mampu menjelaskan variasi data dengan baik. Rumus koefisien determinasi dapat dituliskan sebagai berikut (Hallan & Fajri, 2025):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.20}$$

Keterangan:

- y_i: nilai aktual pada data ke-i
- $\hat{y}_i$: nilai prediksi pada data ke-i
- $\bar{y}$: rata-rata nilai aktual
- n: jumlah data pengamatan

2.2.9 CRISP-DM (*Cross-Industry Standard Process for Data Mining*)

Metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM) merupakan standardisasi pemrosesan *data mining* yang digunakan secara luas baik dalam riset akademis maupun industri karena sifatnya yang terstruktur, efisien, dan independen terhadap domain aplikasi tertentu (Hasanah et al., 2021). Metodologi ini menyediakan kerangka kerja sistematis yang bersifat berulang (*iterative*) guna mentransformasikan data mentah menjadi pengetahuan prediktif yang andal. Secara operasional, metodologi CRISP-DM terdiri atas enam tahapan utama yang saling berkesinambungan, yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*.

1. Pemahaman Bisnis (*Business Understanding*): Tahapan awal dalam metodologi CRISP-DM yang bertujuan memahami permasalahan penelitian atau kebutuhan bisnis, menetapkan tujuan yang ingin dicapai, serta menerjemahkannya ke dalam permasalahan analisis data yang dapat diselesaikan menggunakan teknik data mining atau *machine learning* (Siregar et al., 2017).

2. Pemahaman Data (*Data Understanding*): Tahap ini meliputi proses pengumpulan data, eksplorasi karakteristik data, identifikasi kualitas data, serta analisis awal untuk memahami kondisi dataset sebelum dilakukan proses pengolahan lebih lanjut (Afandi et al., 2024).
3. Persiapan Data (*Data Preparation*): Tahap persiapan data mencakup proses pemilihan variabel, pembersihan data, penanganan nilai hilang (*missing values*), transformasi data, serta penyusunan dataset agar siap digunakan pada proses pemodelan (Jatnika et al., 2025).
4. Pemodelingan (*Modeling*): Tahap pemodelan merupakan proses membangun model menggunakan algoritma *machine learning* yang dipilih. Pada tahap ini dilakukan proses pelatihan model, pengaturan parameter, dan pengujian beberapa konfigurasi untuk memperoleh model dengan performa terbaik (Afandi et al., 2024).
5. Evaluasi (*Evaluation*): Tahap evaluasi bertujuan menilai kualitas model yang telah dibangun berdasarkan metrik evaluasi yang sesuai serta memastikan model telah memenuhi tujuan penelitian atau kebutuhan bisnis sebelum digunakan lebih lanjut (Hasanah et al., 2021).
6. Penyebaran (*Deployment*): Tahap deployment merupakan proses penerapan hasil pemodelan ke dalam lingkungan nyata. Implementasinya dapat berupa publikasi model dalam bentuk aplikasi atau sistem sehingga pengguna dapat melakukan prediksi/klasifikasi secara langsung berdasarkan data masukan yang tersedia (Siswipraptini et al., 2023).

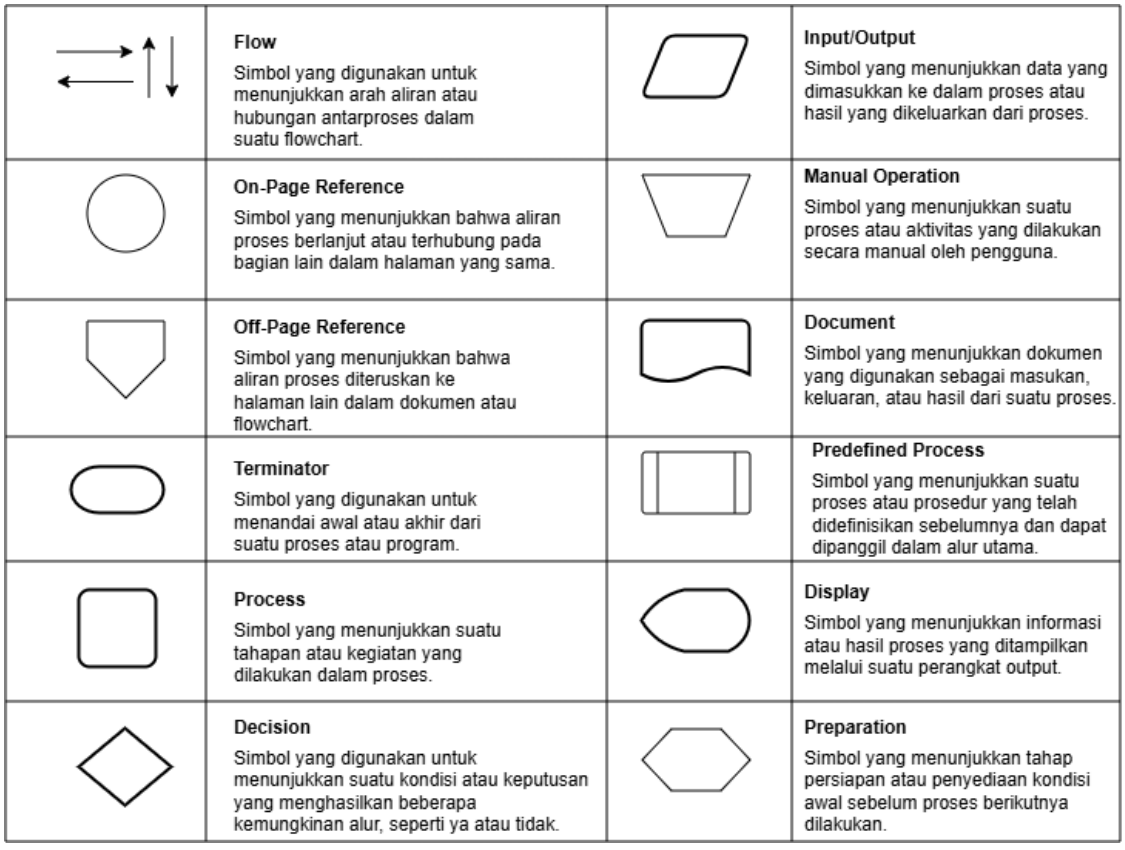
2.2.10 Python

Python merupakan bahasa pemrograman tingkat tinggi yang banyak digunakan dalam pengembangan perangkat lunak, *data science*, dan *machine learning* (Raschka et al., 2020). Python memiliki sintaks yang sederhana sehingga mudah dipahami serta didukung berbagai library seperti Pandas, NumPy, Scikit-learn, dan XGBoost yang membantu proses pengolahan data dan pembangunan model

prediksi. Karena kemudahan dan fleksibilitasnya, Python menjadi salah satu bahasa pemrograman yang populer dalam penelitian berbasis data mining dan machine learning (Raschka et al., 2020).

Pada penelitian ini, Python digunakan sebagai bahasa pemrograman utama dalam proses pengolahan data dan pemodelan *machine learning*. Python dimanfaatkan untuk melakukan preprocessing data, visualisasi data, pembagian dataset training dan testing, pembangunan model Random Forest dan XGBoost, serta evaluasi hasil prediksi. Penggunaan Python memudahkan proses analisis data karena tersedia berbagai library yang mendukung implementasi algoritma machine learning secara efisien dan akurat (Syahreza et al., 2024).

2.2.11 Flowchart



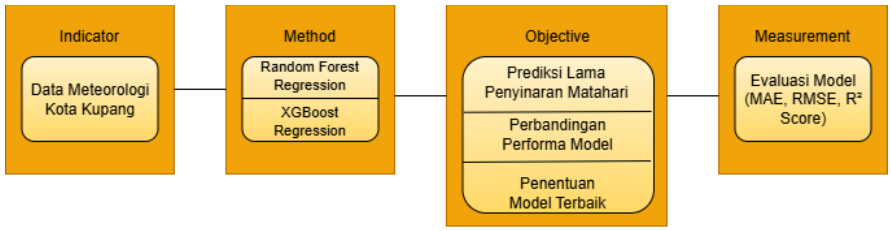
Gambar 2. 1 Simbol-simbol Flowchart

Flowchart merupakan diagram yang digunakan untuk menggambarkan urutan proses atau alur kerja secara visual dan sistematis. Flowchart membantu menjelaskan tahapan-tahapan suatu proses menggunakan simbol-simbol tertentu yang saling terhubung sehingga alur proses menjadi lebih mudah dipahami (Diaraya, 2017).

Dalam penelitian *machine learning*, flowchart digunakan untuk memvisualisasikan tahapan penelitian mulai dari pengumpulan data, preprocessing data, pembagian dataset, pembangunan model Random Forest dan XGBoost, hingga evaluasi hasil prediksi lama penyinaran matahari di Kota Kupang. Penggunaan flowchart membantu peneliti menyusun proses penelitian secara terstruktur dan memudahkan pembaca memahami alur penelitian yang dilakukan (Bostr, 2022).

2.3 Kerangka Pemikiran

Kerangka pemikiran merupakan gambaran konseptual yang menunjukkan hubungan antar variabel serta alur penelitian yang digunakan untuk menyelesaikan permasalahan penelitian. Kerangka pemikiran membantu peneliti dalam menyusun tahapan penelitian secara sistematis mulai dari pengumpulan data, pengolahan data, pemodelan, hingga evaluasi hasil penelitian. Dalam penelitian ini, kerangka pemikiran digunakan untuk menggambarkan proses prediksi lama penyinaran matahari di Kota Kupang menggunakan metode Random Forest dan XGBoost.



Gambar 2. 2 Kerangka Pemikiran

1. *Indicator* (Data Input)

   Indikator dalam penelitian ini merupakan variabel independen yang digunakan sebagai data masukan dalam proses pemodelan *machine learning* untuk memprediksi lama penyinaran matahari di Kota Kupang. Variabel yang digunakan

berasal dari data meteorologi yang berkaitan dengan kondisi cuaca dan penyinaran matahari.

Variabel independen yang digunakan meliputi Tn (suhu minimum (°C)), Tx (suhu maksimum (°C)), Tavg (suhu rata-rata (°C)), RH_avg (kelembapan rata-rata (%)), RR (curah hujan (mm)), dan ff_avg (kecepatan angin rata-rata (m/s)). Parameter-parameter meteorologi tersebut digunakan sebagai prediktor karena memiliki keterkaitan dengan kondisi atmosfer yang dapat memengaruhi lama penyinaran matahari. Misalnya, curah hujan dan kelembapan udara yang tinggi cenderung berkaitan dengan kondisi berawan sehingga dapat mengurangi durasi penyinaran matahari. Sebaliknya, kondisi cuaca yang lebih cerah dan curah hujan yang rendah cenderung berkaitan dengan durasi penyinaran matahari yang lebih panjang.

Variabel dependen dalam penelitian ini adalah *ss* (lama penyinaran matahari (jam)). Lama penyinaran matahari merupakan durasi waktu matahari bersinar pada suatu wilayah dalam periode tertentu dan menjadi target yang diprediksi dalam penelitian ini.

2. *Method* (Metode)

   Penelitian ini menggunakan pendekatan *machine learning* berbasis *ensemble learning* untuk memprediksi lama penyinaran matahari dan membandingkan performa dua algoritma, yaitu Random Forest dan XGBoost.

   a. Random Forest

      Random Forest merupakan metode *ensemble* berbasis *bagging* yang membangun banyak *decision tree* dan menggabungkan hasil prediksi dari setiap *tree*. Metode ini mampu menangani hubungan nonlinier antarvariabel dan dapat mengurangi risiko *overfitting* melalui penggabungan banyak *tree*, sehingga sesuai untuk digunakan dalam pemodelan data meteorologi.

   b. XGBoost

      XGBoost merupakan metode *ensemble* berbasis *boosting* yang membangun model secara bertahap, di mana setiap *tree* berikutnya berupaya memperbaiki kesalahan dari model sebelumnya. Algoritma ini menggunakan mekanisme

gradient boosting dan regularisasi untuk menghasilkan model prediksi yang efektif.

3. *Objective* (Tujuan)

   Tujuan utama penelitian ini adalah membangun model prediksi lama penyinaran matahari di Kota Kupang menggunakan algoritma Random Forest dan XGBoost serta membandingkan performa kedua algoritma tersebut untuk menentukan model dengan performa terbaik. Penelitian ini juga bertujuan untuk mengetahui kemampuan kedua model dalam memprediksi lama penyinaran matahari berdasarkan parameter meteorologi yang digunakan. Hasil prediksi diharapkan dapat menjadi salah satu informasi pendukung dalam analisis energi surya di Kota Kupang.

4. *Measurement* (Pengukuran)

   Evaluasi performa model dilakukan menggunakan beberapa metrik regresi untuk mengetahui tingkat akurasi hasil prediksi.

   a. *Mean Absolute Error* (MAE)

      MAE digunakan untuk menghitung rata-rata selisih absolut antara data aktual dan hasil prediksi. Semakin kecil nilai MAE, maka semakin baik tingkat akurasi model.

   b. *Root Mean Squared Error* (RMSE)

      RMSE digunakan untuk mengukur akar dari rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. RMSE lebih sensitif terhadap kesalahan prediksi yang besar karena error dikuadratkan dalam proses perhitungannya. Semakin kecil nilai RMSE, maka semakin rendah tingkat kesalahan prediksi model.

   c. Koefisien Determinasi (R^2)

      R^2 digunakan untuk mengukur proporsi variasi nilai aktual variabel target yang dapat dijelaskan oleh model berdasarkan variabel prediktor yang digunakan. Nilai R^2 yang semakin mendekati 1 menunjukkan kemampuan model dalam menjelaskan variasi data target yang semakin baik.