# Solar Irradiance Prediction Using Satellite data and Machine Learning

Internship Project | February – June 2025

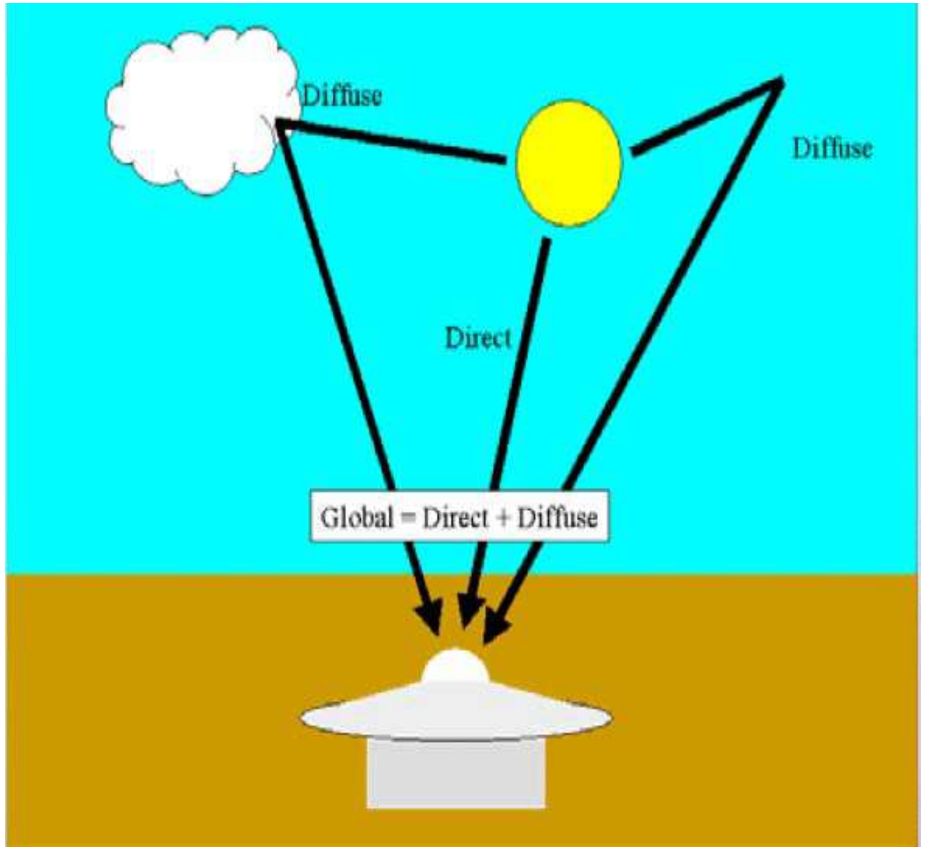
Gaurav Lute

# Outline

- Introduction
- Project Overview
- Problem Statement
- Dataset Description
- Project Workflow
- Exploratory Data Analysis
- Feature Engineering
- Models and Results
- Evaluation Table

# Introduction

- Solar irradiance is the power per unit area received from the sun in the form of electromagnetic radiation, measured in watts per square meter($\frac{W}{m^2}$).
- It represents the amount of solar solar energy hitting a surface at given time.
- Types of Solar irradiance:
  - **GHI(Global Horizontal irradiance) :-** Total Solar radiation received on horizontal surface including both direct and difuse sunlight.
  - **Direct Normal irradiance (DNI) :-** Radiation coming from directly from the sun, measured perpendicular to its rays.
  - **Diffuse Horizontal Irradiance (DHI) :-** Sunlight scattered by the atmosphere (cloud,



# Project Overview

### 📌 Objective

- Forecast Global Horizontal Irradiance (GHI) using satellite data and machine learning.

### Dataset

- NASA POWER (Feb 2016 – Feb 2025),
- Pune, India.

### Techniques Used

- Classical ML
- Time Series Models.

### Goal

- Predict GHI for **March 2025** to support short-term solar energy planning.

### Application

- Useful for solar panel optimization,
- Grid integration.
- Renewable energy forecasting.

# Problem Statement

### Problem 1: Parametric Estimation

- **Objective:**
  Estimate GHI using a simple parametric model based on known physical and atmospheric relationships.
- **Approach:**
  Use meteorological features (temperature, humidity, wind, etc.) to Random forest, Gradient Boosting etc.
- **Purpose:**
  Understand feature influence and provide quick estimates using ML.

### Problem 2: ML-Based Forecasting

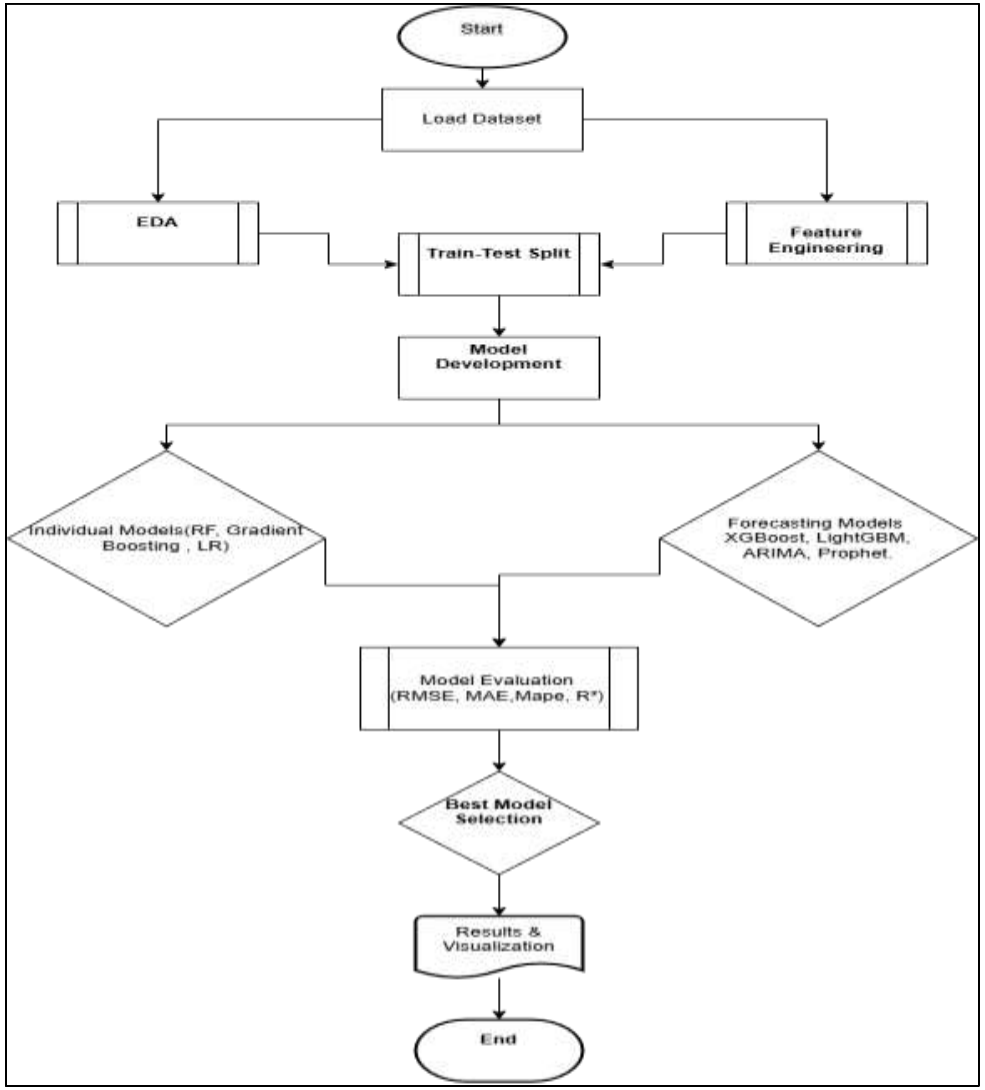
- **Objective:**
  Forecast solar irradiance values for **March 2025** using historical satellite data.
- **Approach:**
  Use ML and time series techniques like XGBoost, LightGBM, ARIMA, Prophet.
- **Purpose:**
  Build a data-driven model that learns temporal and atmospheric patterns to give accurate short-term GHI forecasts.

# Dataset Description

- **Source:**
  - NASA POWER (Prediction Of Worldwide Energy Resources)
- **Location:**
  - Pune, India
  - Latitude: **18.52°N,** Longitude: **73.86°E**
- **Time Period:**
  - February 28, 2016 – February 28, 2025
- **Frequency:**
  - **Daily resolution** (365 observations per year)
- **Target Variable:**
  - **ALLSKY SFC SW DWN** Global Horizontal Irradiance (GHI)
- **Input Features (Examples):**
  - ☀ **Solar:** DNI, DHI radiation
  - **Temperature:** T2M, Dew Point
  - **Humidity:** RH2M, QV2M
  - **Wind:** Speed and Direction at 2m/10m
  - **Time:** Day of Year (DOY), Lag Features

# Project Workflow

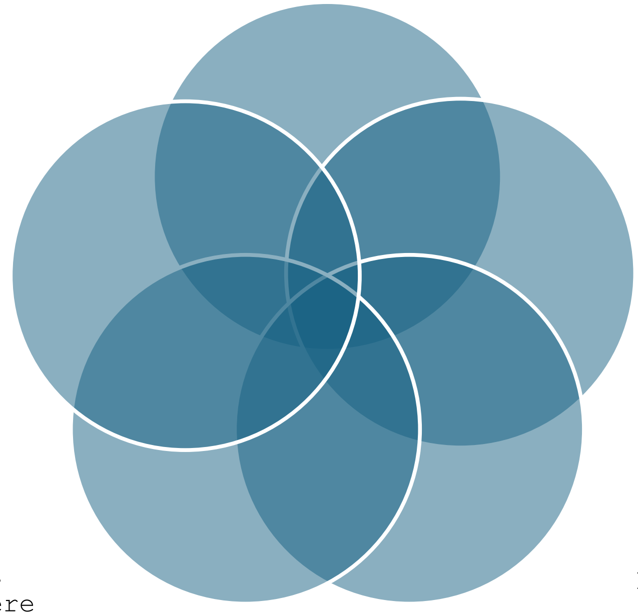
- **Objective**
  Predict **solar irradiance** using satellite data.
  Build reliable models for energy planning and forecasting.
- **Data Source**
  NASA POWER (2016–2025) and others like NSRDB, GHI, temperature, humidity, cloud cover, wind speed, etc.
- **Methodology Highlights**
  Applied **Exploratory Data Analysis (EDA)** to understand patterns.
  Used **feature engineering** for time-based and domain-specific variables.
  Applied **ML models (RF, XGBoost, LR)** and **forecasting models (ARIMA, Prophet)**.
- **Model Evaluation Metrics**
  RMSE, MAE, MAPE, $R^2$ to assess model accuracy.
  Cross-validation or walk-forward validation (if time series).
- **Outcome**
  Best-performing model identified.
  Visualized predictions vs actual values.
  Insights generated for solar energy planning.



# Exploratory Data Analysis

✓ **Seasonal Patterns:**
GHI shows a clear **annual cycle** with peak values during summer and dips during monsoon/winter.

✓ **Correlation Analysis:**
Features like T2M, RH2M, and cloud-related variables showed **moderate to strong correlation** with GHI.



✓ **Feature Relationships:**
Temperature and solar radiation show **non-linear patterns;** humidity and precipitation show **negative correlation** with GHI.

✓ **Wind Direction Encoding:**
Circular features (wind direction) were transformed using **sine and cosine** encoding.

✓ **Outliers & Distribution:**
Few extreme values found in precipitation, wind speed, and solar radiation handled using **log transformation** and **boxplot analysis.**

✓ **Seasonal Patterns:** GHI shows a clear **annual cycle** with peak values during summer and dips during monsoon/winter.

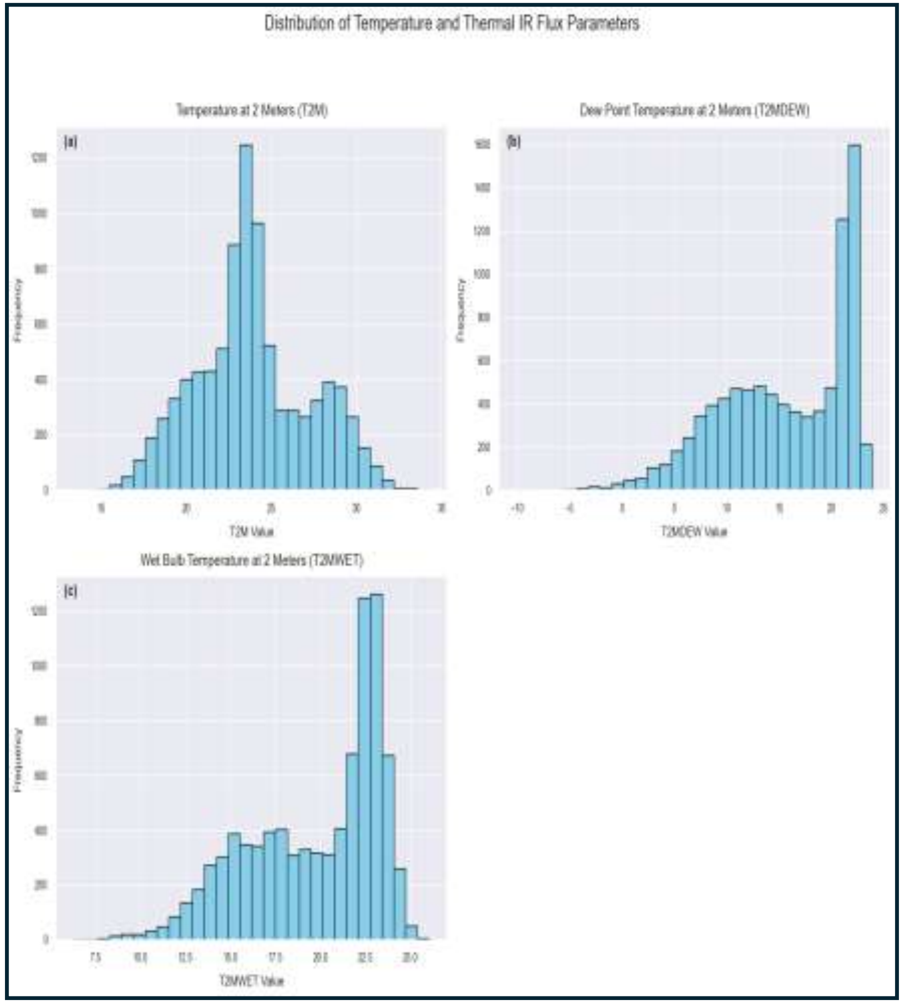
✓ **Feature Relationships:** Temperature and solar radiation show **non-linear patterns;** humidity and precipitation show **negative correlation** with GHI.

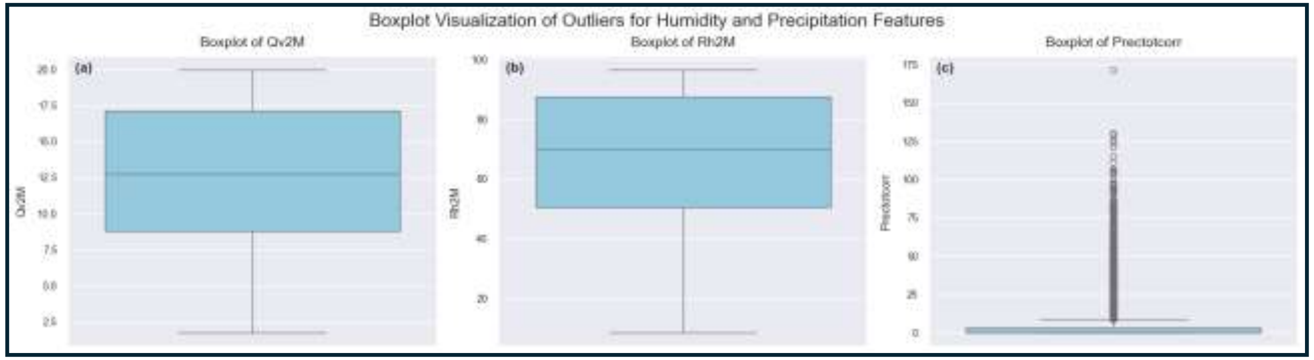
✓ **Outliers & Distribution:** Few extreme values found in precipitation, wind speed, and solar radiation handled using **log transformation** and **boxplot analysis.**

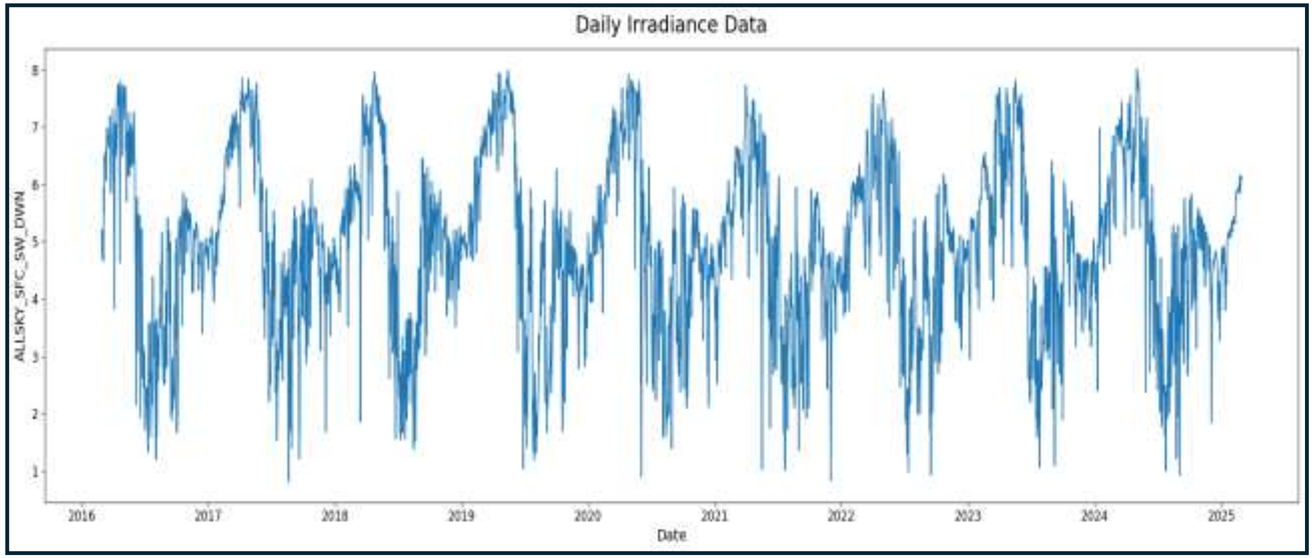
✓ **Wind Direction Encoding:** Circular features (wind direction) were transformed using **sine and cosine** encoding.

✓ **Correlation Analysis:** Features like T2M, RH2M, and cloud-related variables showed **moderate to strong correlation** with GHI.

# EDA for ML Model



# Feature Engineering

## Classical ML Model

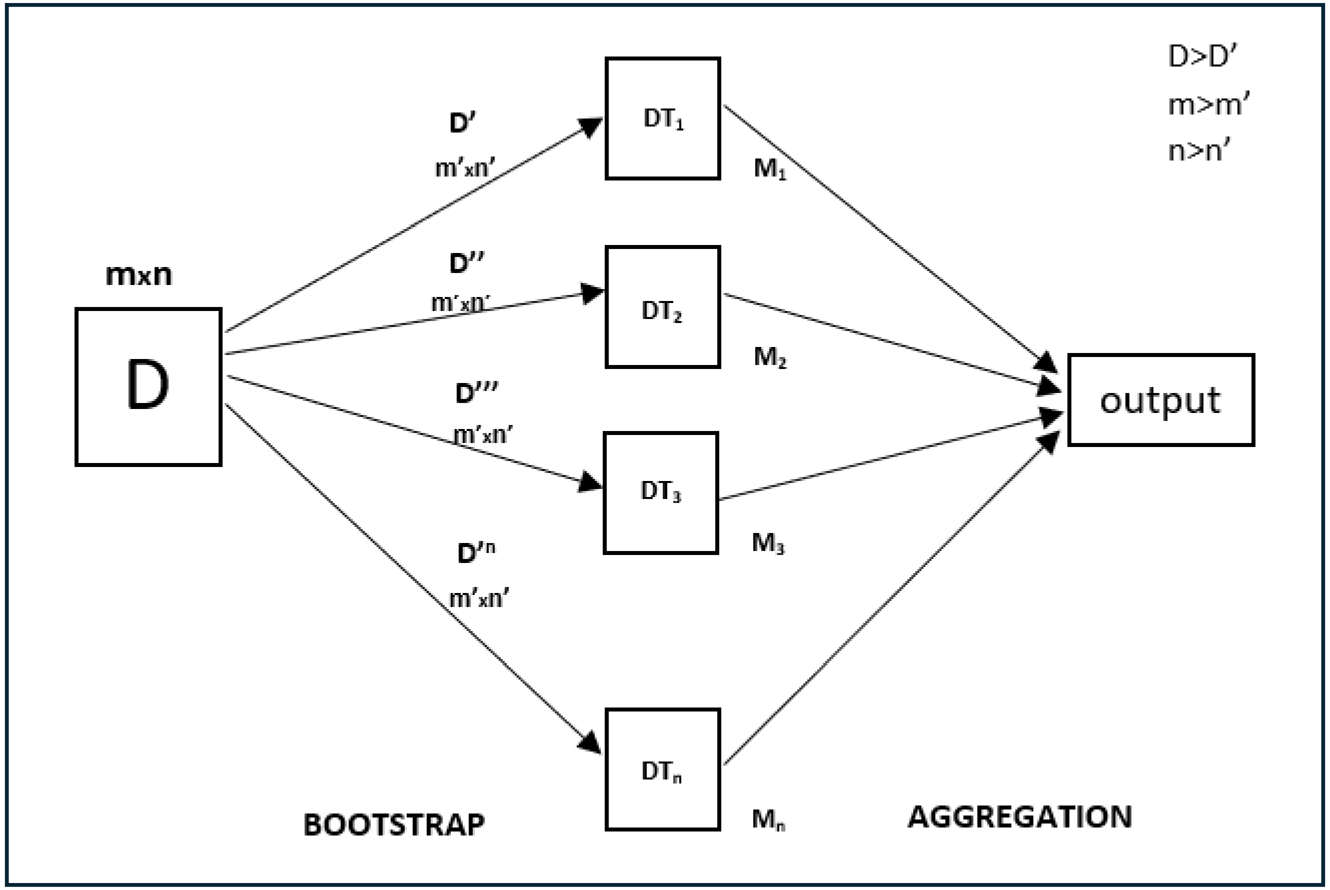
- Logarithmic transformation in PRECTCORR columns(Highly Right skew).
- Transform Wind direction Feature.
- Remove Highly Uncorrelated Feature.

## Time Series Prediction Mode

- Developed New Lags Features.
- Lag_1 : GHI value from the previous day.
- Lag_2 : GHI value from two days prior.
- Lag_7 : GHI value from the same day one week earlier.
- Lag_365 : GHI value from the same day one year earlier.

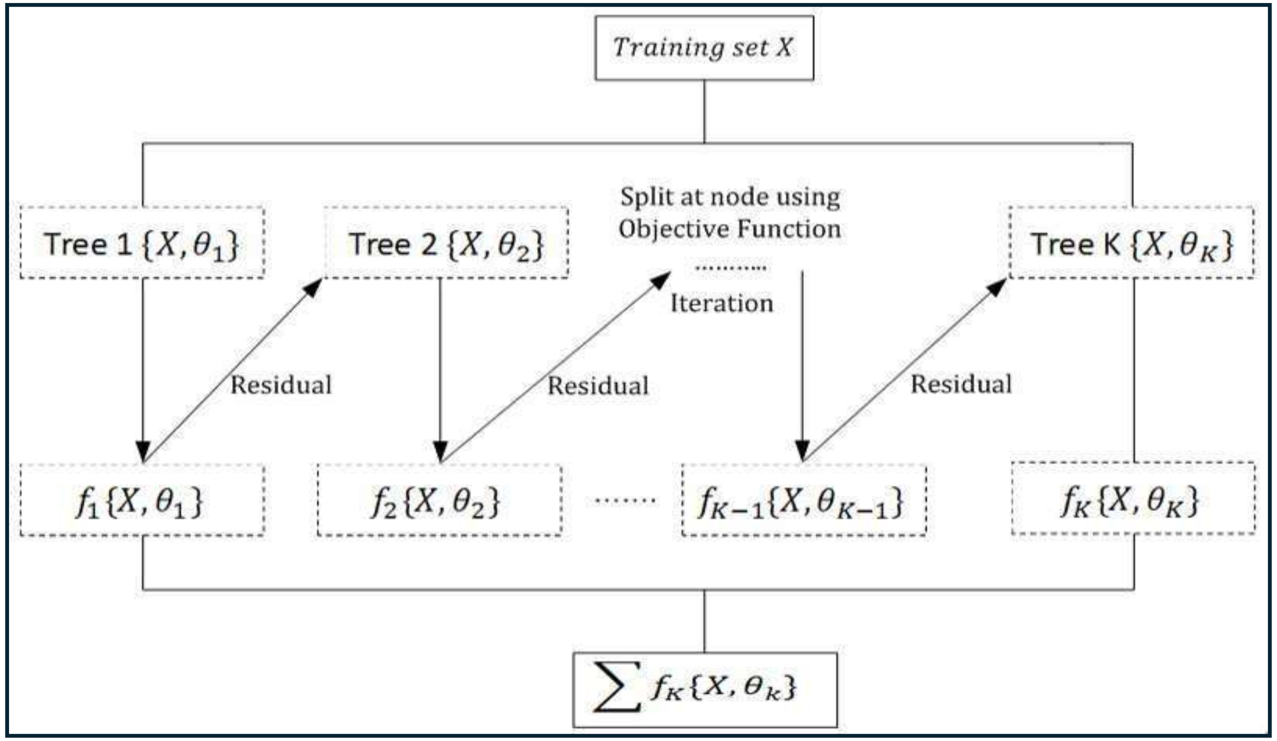
# Classical Machine Learning Models

- Random Forest Model
  - A **tree-based ensemble learning model.**
  - Handles **non-linear relationships** very well.
- Result
  - Average MSE: 0.39
  - Average R2 Score: 0.98

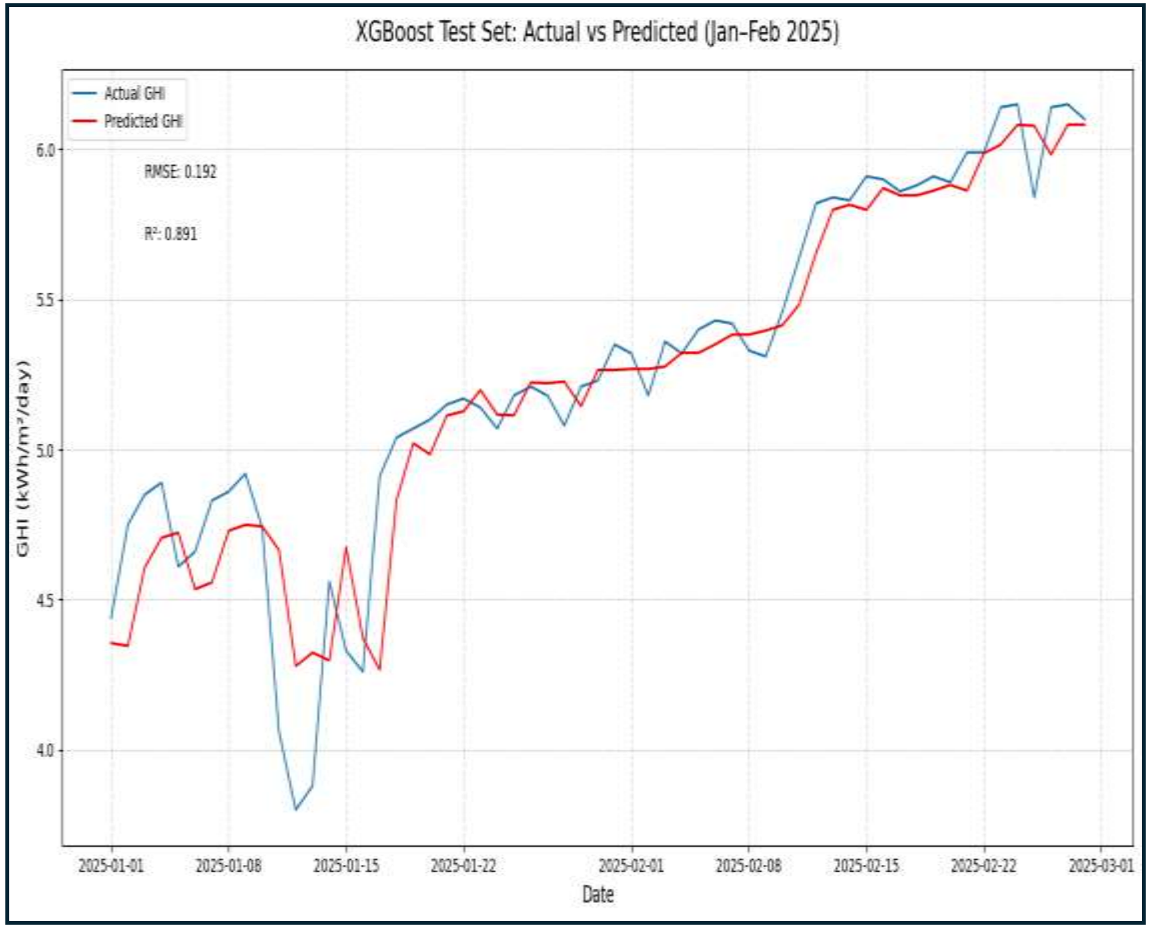


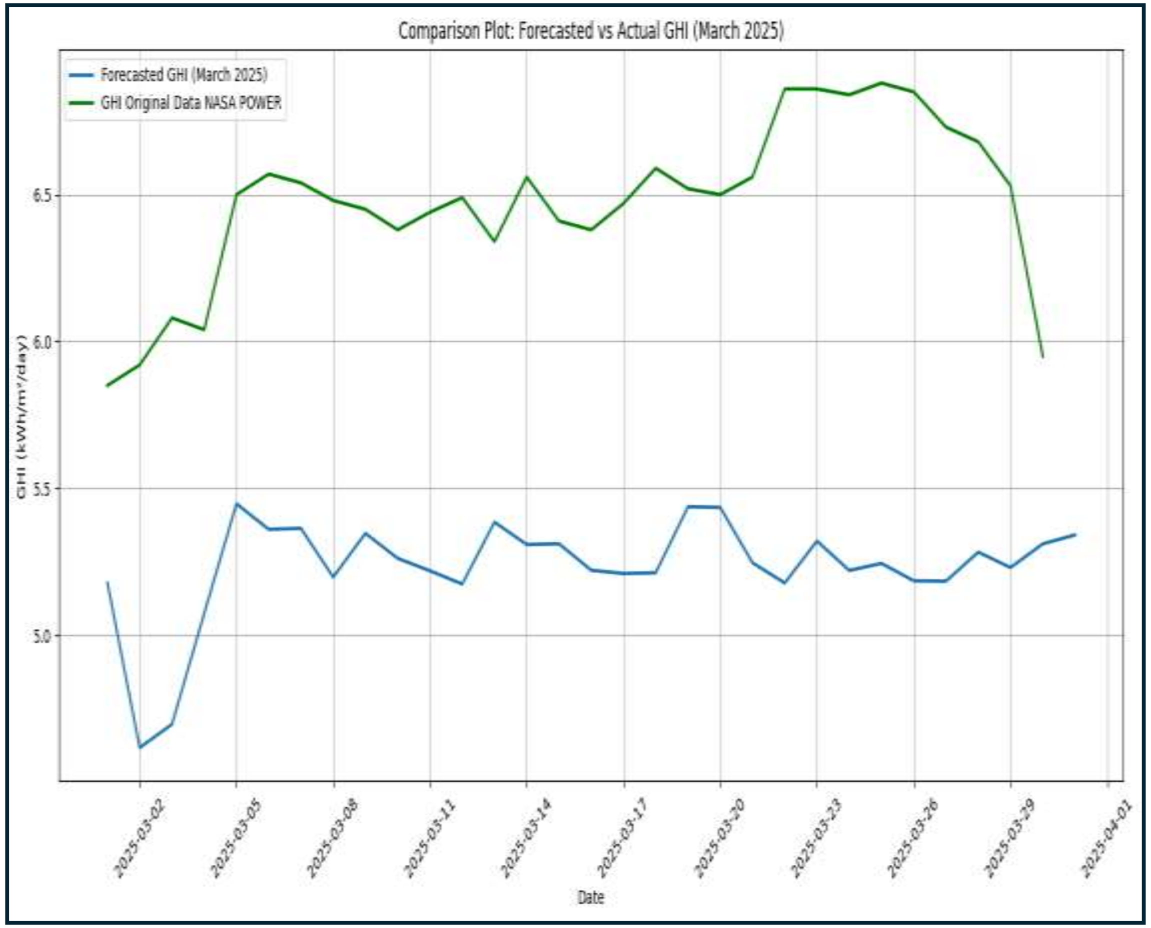
# XGBoost Model with Lag variables.

- **XGBoost** stands for **Extreme Gradient Boosting**.
- Builds trees **sequentially**, each one trying to **correct the errors** of the previous model.
- Feature Engineering - Developed some lag features.
- Results
  - RMSE :- 0.192
  - R2 Score :- 0.891



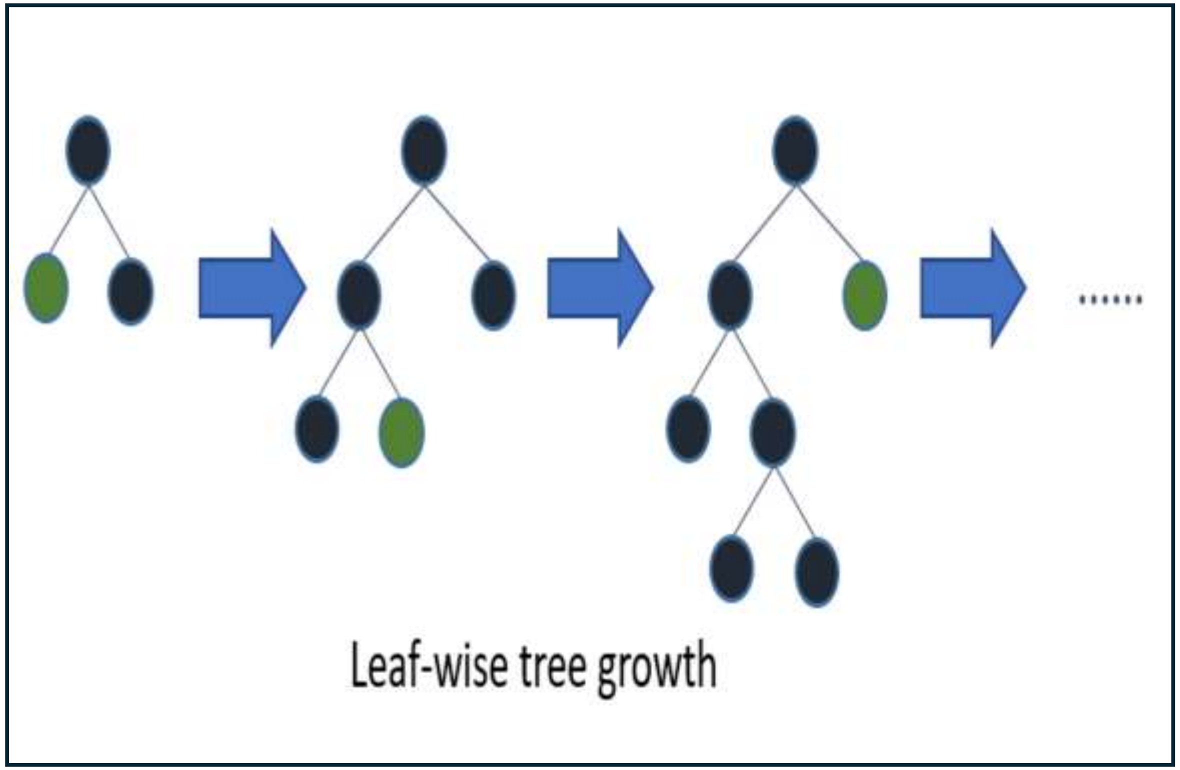
# XGBoost Models with Lag variables.



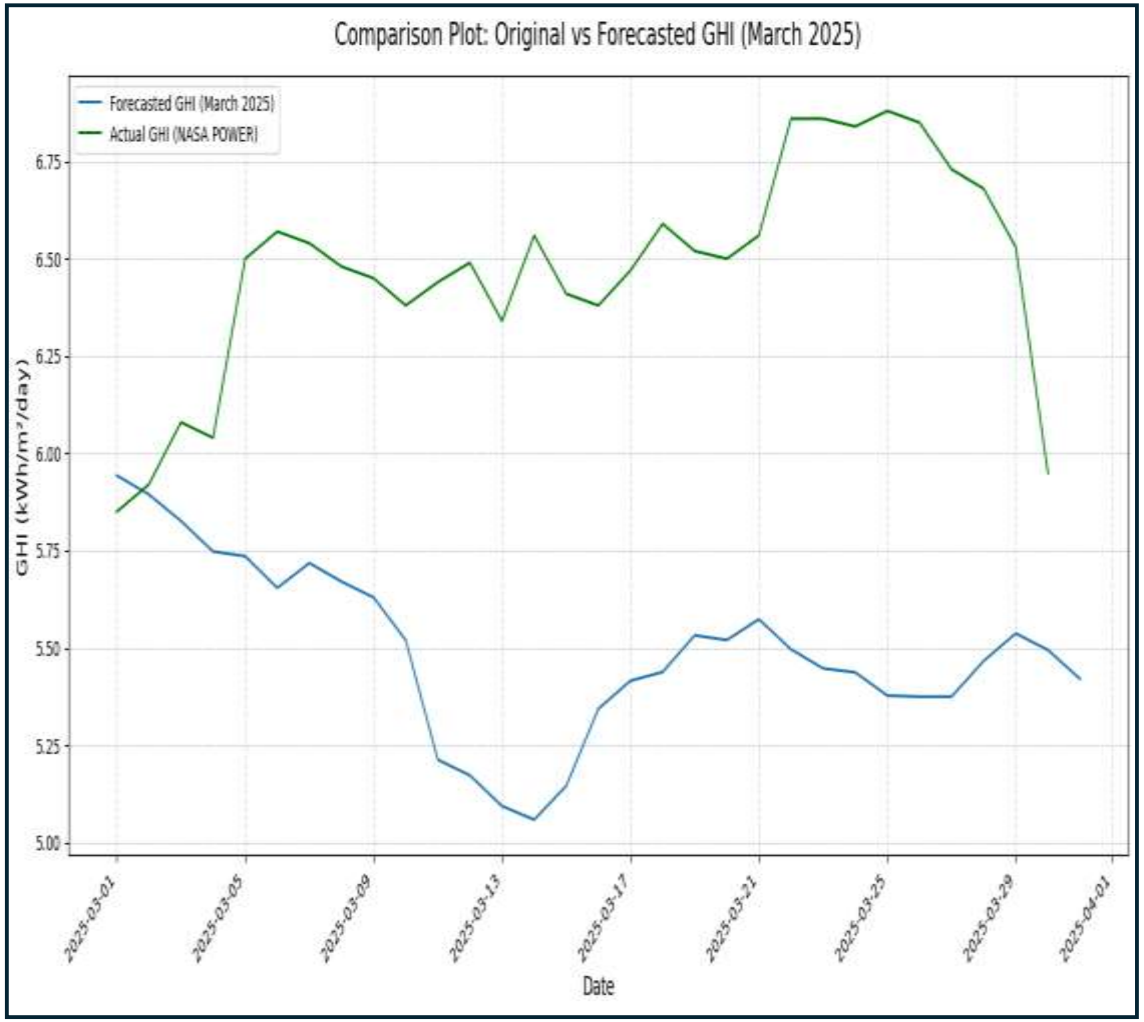



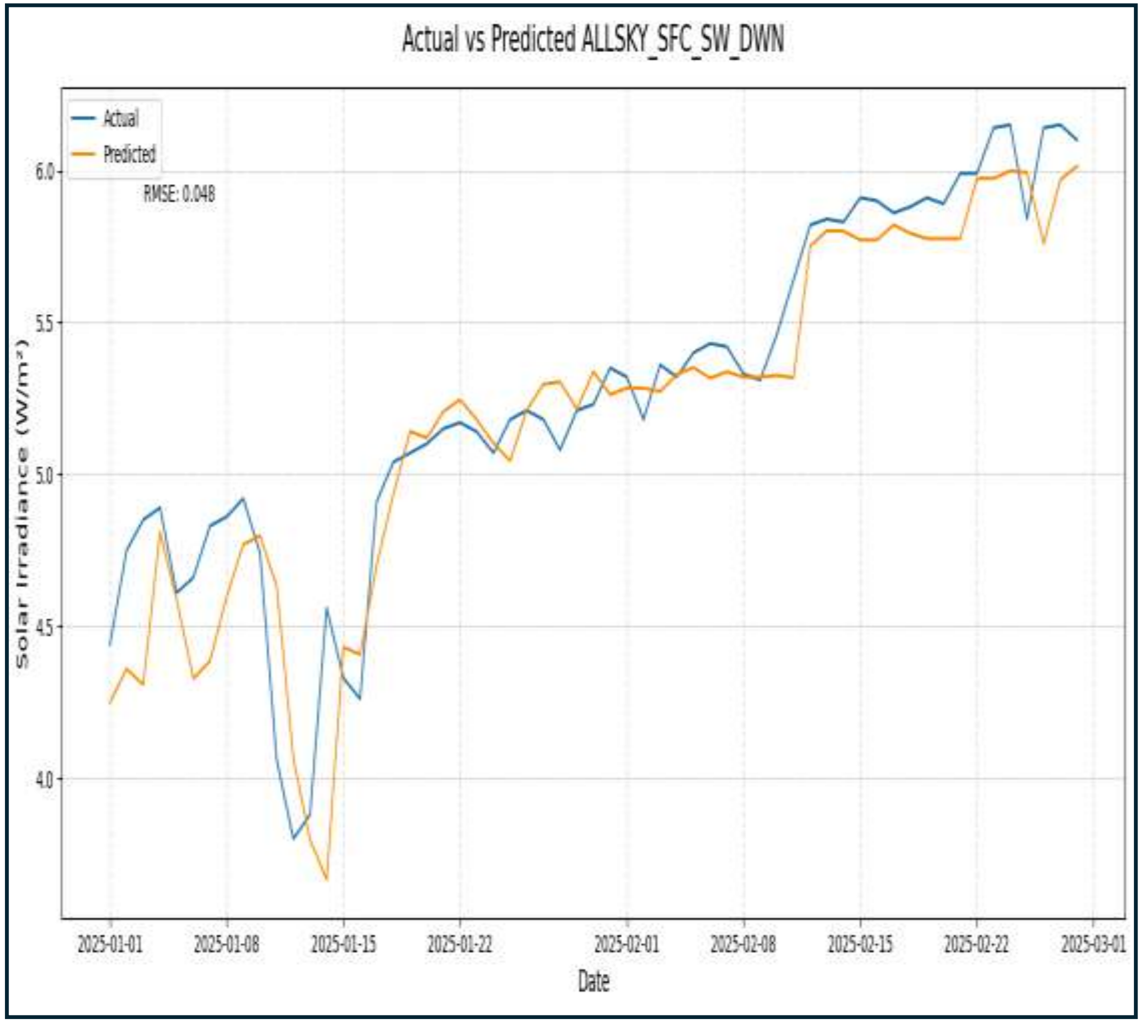
# Light-GBM (Light Gradient Boosting Machine) model

- **LightGBM (Light Gradient Boosting Machine)** is fast.
- Like XG-Boost, it builds decision trees **sequentially** to minimize errors.
- But unlike traditional boosting:
  - It grows tree **leaf-wise** (best leaf first), not level-wise.
  - Uses **histogram-based binning** for faster split finding.
  - Works best when data is **large and sparse**.
- Result
  - RMSE:- 0.05
  - R 2 Score :- 0.856



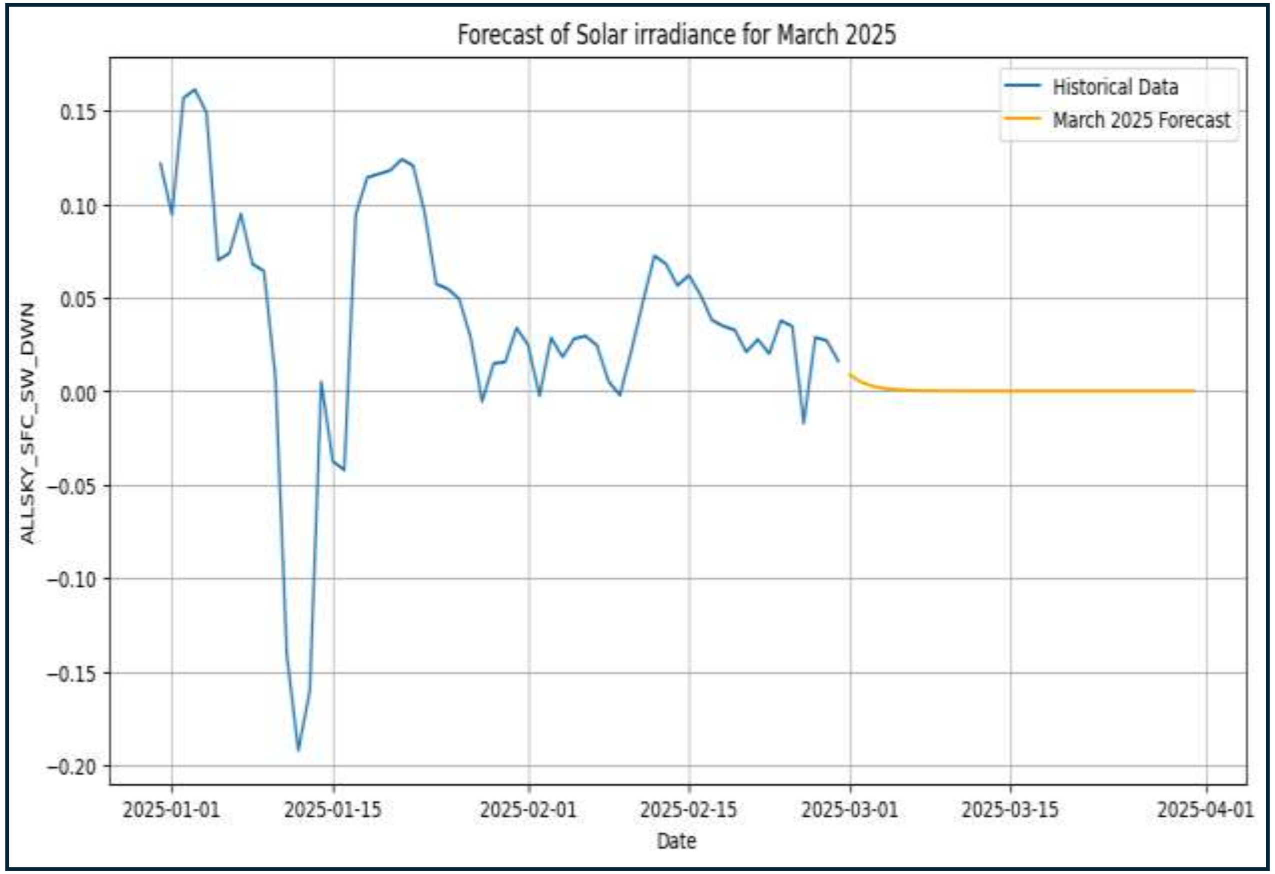
# Light-GBM (Light Gradient Boosting Machine) model



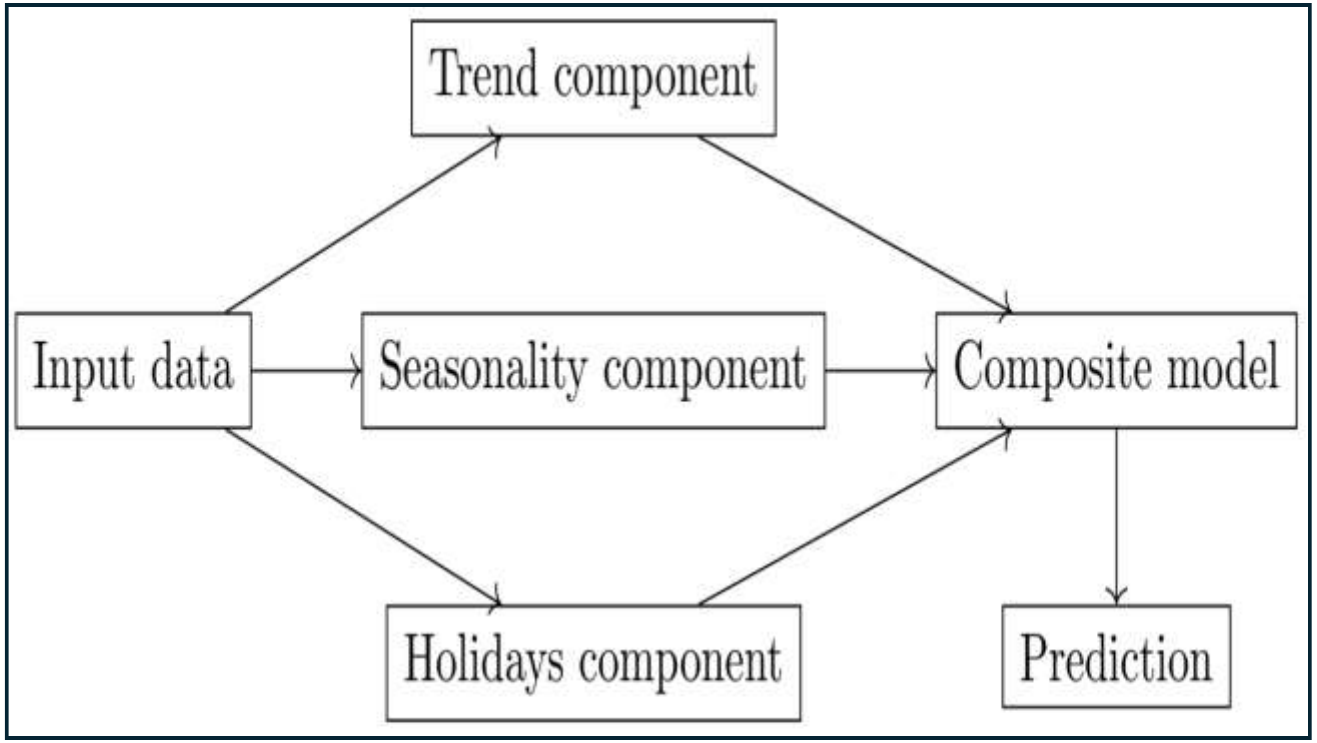
# ARIMA (Autoregressive Integrated Moving Average)

- **ARIMA** is a **classical time series forecasting model.**
- It combines **Auto-Regression (AR)**, **Integration (I)**, and **Moving Average (MA)** to model temporal data.
- **Component Explanation:**
  - **AR (Auto-Regression):** Predicts using **past values** (lags of the target)
  - **I (Integration):** Applies **differencing** to make the time series **stationary**
  - **MA (Moving Average):** Uses **past forecast errors** to improve prediction.
- Results
  - RMSE :- 0.035
  - MAE :- 0.035

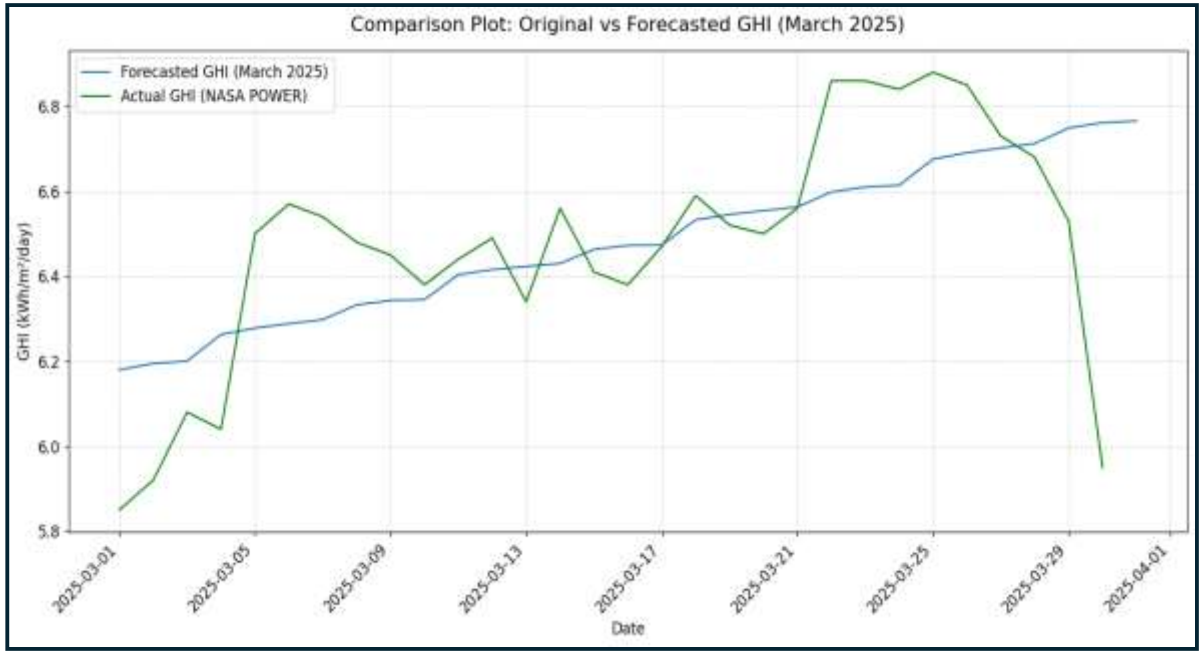


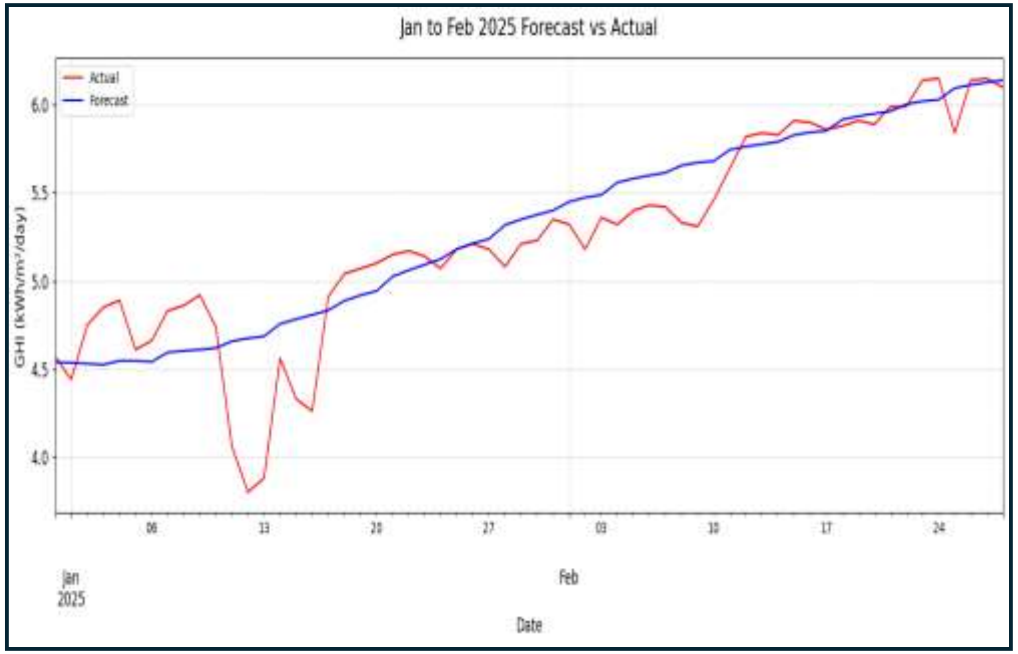
# Facebook Prophet Model

- **Prophet** is an **open-source time series forecasting tool** developed by **Facebook.**
- Ideal for **daily, weekly, and monthly time series** with **seasonality and trends.**
- Automatically detects and models:
  - **Trend** (piecewise linear or saturating)
  - **Seasonality**
  - **Holidays/events** (Optional)
- Results
  - MAPE(Mean absolute percentage error) :- 0.03
  - Mean absolute error :- 0.17



# Facebook Prophet Model

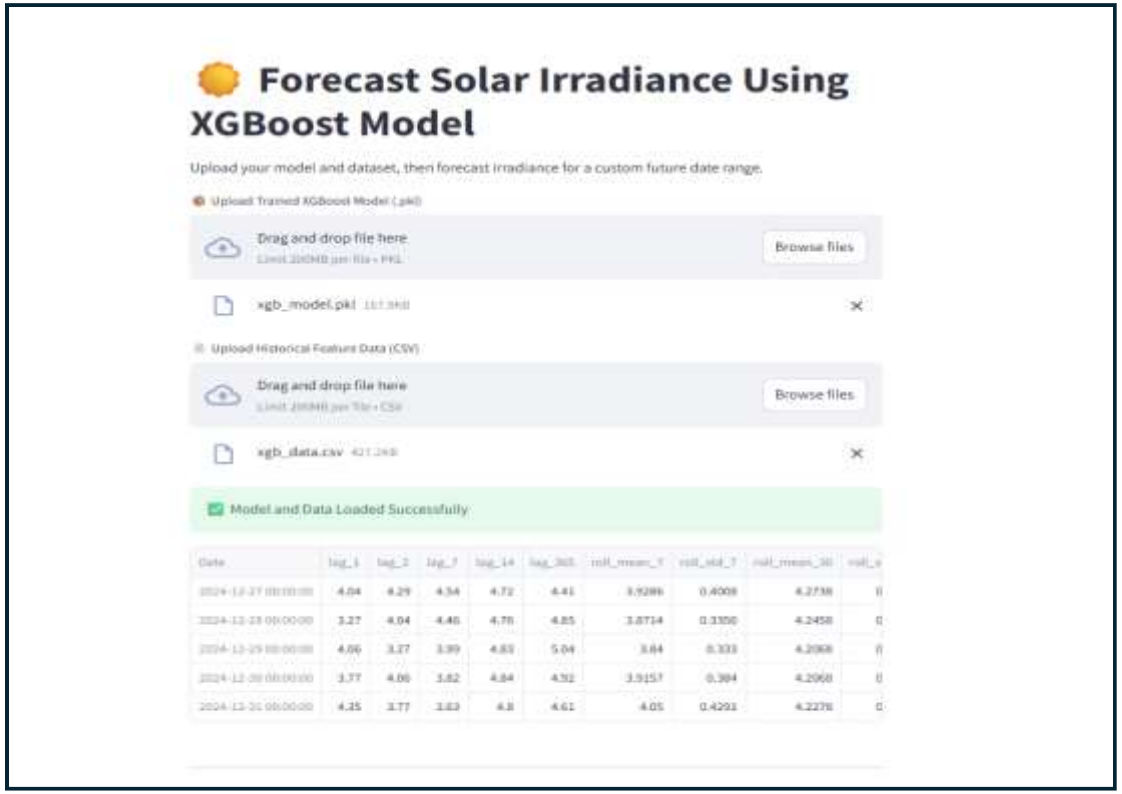


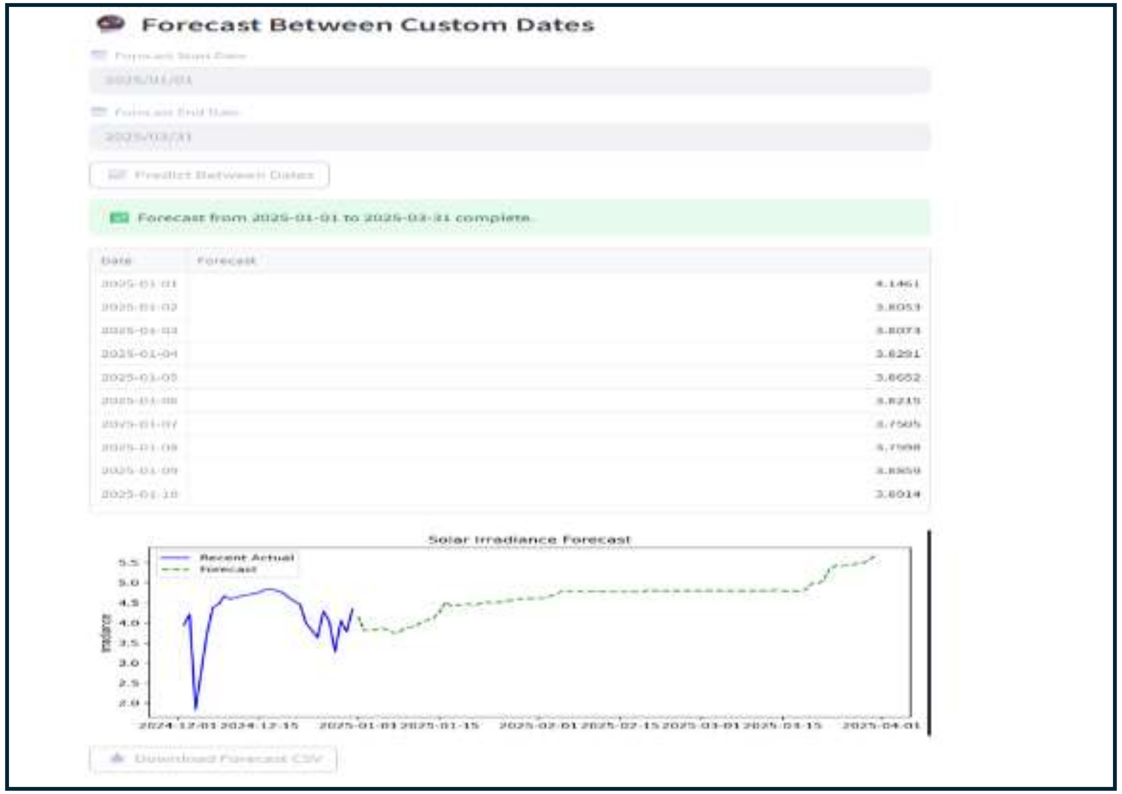



# Evaluation Table

| Model | RMSE ↓ | $R^2$ Score ↑ | Key Strength |
|---|---|---|---|
| XGBoost | 0.05 | 0.89 | Best overall accuracy & robust with lag features |
| Random Forest | 0.39 | 0.98 | Strong on non-linear patterns |
| Gradient Boosting | 0.43 | 0.98 | Good balance, slightly slower |
| LightGBM | 0.05 | 0.856 | Fastest training, low memory usage |
| ARIMA | RMSE :- 0.035 | MAE :- 0.035 | Simple, interpretable, but univariate |
| Facebook Prophet | MAPE = 0.03 | 0.81 | Trend-aware & easy, but less flexible |

# Simple Streamlit UI interface



Thank you