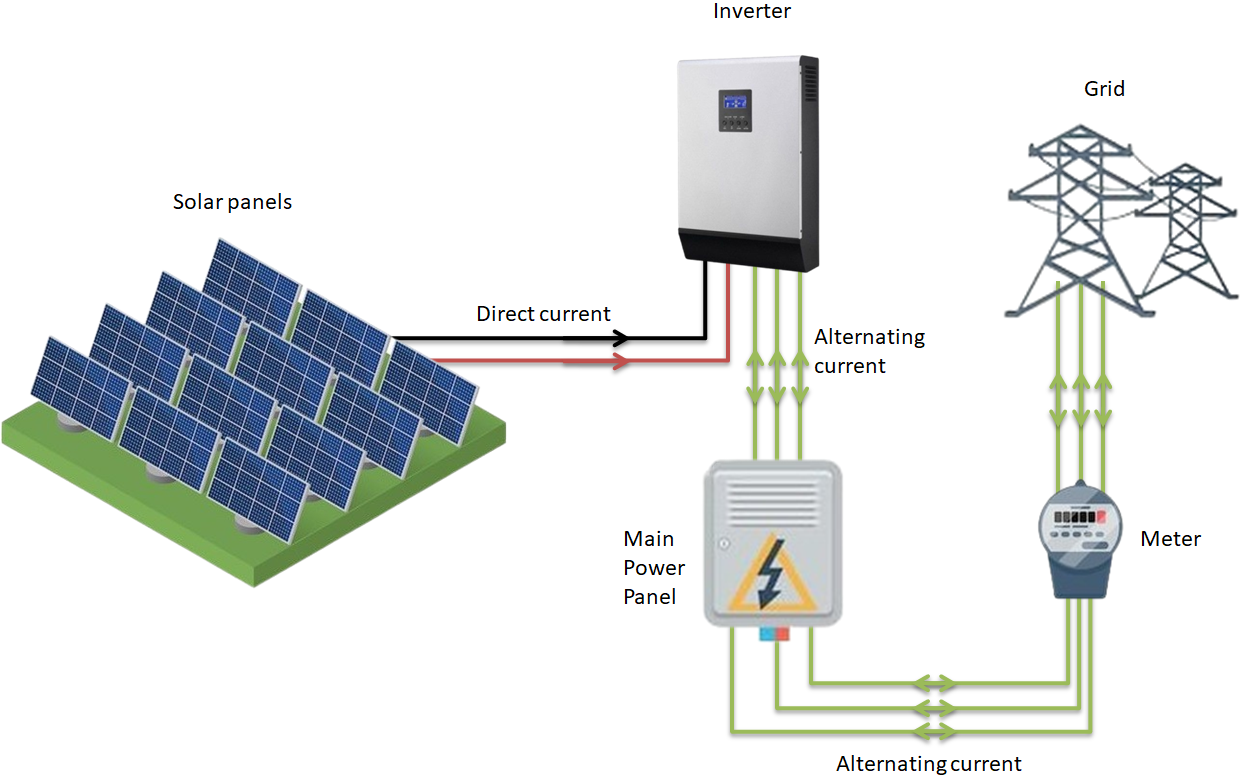
# Introduction

## Motivation

Photovoltaic solar energy has been expanding significantly around the world, due to its generation being considered renewable and sustainable. Often, this source is seen as a way to diversify the energy matrix of countries, besides generating low environmental and social impacts. Despite this, solar energy still represents a very small portion of the global energy matrix, but this index is expected to increase in the coming years due to the high investment in so-called green energies and concern about climate change due to global warming.

Basically, a solar generation system consists of interconnected solar panels connected to an inverter, which in turn is connected to a main power panel. The solar panels are responsible for converting sunlight into Direct Current (DC) energy. The inverter converts the current to Alternating Current (AC) and matches its frequency with the utility grid frequency, transforming the energy into a usable form for storage in batteries or direct supply to the grid. The energy going into the grid passes through a bidirectional meter that measures the amount in kWh of supplied energy.



Solar energy is not always available; after all, we have nights and cloudy, rainy days. This requires efficient ways to transport and store energy, as well as requiring higher efficiency from photovoltaic cells. Nevertheless, the generation system presents its advantages: It generates less environmental impact since it is an energy generator that does not produce greenhouse gases; It has an infinite source of energy since it uses the sun to produce it; and has low operation and maintenance costs compared to the costs of other forms of energy generation.

Although Solar Plants are considered to have low maintenance costs, there is still a need for improvement in the operation and maintenance performance since a failure or defect can significantly reduce generation or even render the plant unable to generate energy.

A good way to ensure the efficiency of the solar system as a whole would be to understand how photovoltaic solar panels behave in different external temperatures and the levels of solar irradiation that fall on the panels, as well as providing swift maintenance to detect generation failures or decreases in energy efficiency in the components that make up a photovoltaic solar power plant.

## Research Problem

The lack of effective monitoring is one of the causes that make photovoltaic solar plants less efficient, leading to many unnecessary maintenance tasks, decreased equipment lifespan, and thus increasing the operation cost of the plant, as well as energy generation losses due to equipment downtime that required maintenance.

The increase in maintenance costs and the loss of energy generation directly impact the monetary gains that the plant is capable of generating, whether in the form of energy savings, where energy generation must exceed consumption, or when the generated energy is sold to the end customer, where reduced generation indicates a lower supply of electrical energy.

## Research Questions

**Q1**: How do temperature and irradiation levels impact the efficiency of photovoltaic solar panels?

**Q2**: Can failures or maintenance needs be predicted based on historical generation and temperature data?

**Q3**: Which machine learning models are most effective in predicting energy production and detecting anomalies in photovoltaic plants?

## Research Hypothesis (Ho, Ha)

**Null Hypothesis (H₀)**: There is no significant correlation between ambient/module temperature, irradiation, and the energy generation efficiency of solar panels.

**Alternative Hypothesis (Hₐ)**: There is a significant correlation between ambient/module temperature, irradiation, and the energy generation efficiency of solar panels, enabling accurate forecasting models.

## Research Objectives

**Monitoring and Predictive Maintenance**: The objective is to identify anomalies that may indicate the need for maintenance or cleaning of solar panels, using temperature and irradiation sensor data, ensuring that these panels operate in better conditions, increasing their performance and energy yield sustainably.

**Enhancement of Efficiency**: Through equipment performance data, it will be possible to identify equipment with inferior performance. The goal is to enhance the overall efficiency of the plant, reduce equipment downtime, and extend the lifespan of solar panels and inverters that make up the photovoltaic generation system.

**Effective Energy Generation Management**: The objective is to develop models that can accurately predict energy production by analyzing energy generation and solar irradiation metrics. This capability can reduce losses, enabling effective energy generation management and ensuring better electricity supply.

# Literature Review / Related works

## Technologies you are using

**Data Source**: The source of the datasets is Kaggle, which provides public datasets for analysis. Two datasets generated from photovoltaic energy generation and temperature sensor data from a solar plant in India will be used. These datasets contain 34 days of generation data from the year 2020.

**Employed Technique:** Aggregation into Time Intervals, Merging Dataframes, Exploratory Data Analysis (EDA), Data Preprocessing, Training and Test Sets, Cross Validation, Modelings.

**Machine Learning Models:** Linear Regression, Random Forest and ARIMA, was made to balance interpretability, flexibility, and time-series forecasting capability. Linear Regression offers a simple, explainable baseline model that helps identify linear relationships between irradiation, temperature, and energy output. Random Forest brings robustness and the ability to capture non-linear interactions among features, making it ideal for more complex patterns in the data. Finally, ARIMA is a classical time series model well-suited for capturing temporal dependencies and trends in the energy generation data, especially when seasonality and autocorrelation are present. Together, these models allow for both cross-sectional and temporal analyses, ensuring a comprehensive evaluation of forecasting performance.

**Language and Libraries:** Python (Pandas, Numpy, Scikit-Learn, Seaborn, Matplotlib)

## Summary of Related Work

The reviewed studies on preventive maintenance in solar power plants show the growing application of machine learning (ML) in predictive maintenance by contrasting traditional approaches with data-driven strategies. Betti et al. propose a monitoring system that forecasts inverter breakdowns up to seven days in advance with a 95% sensitivity using SCADA and techniques like neural networks and unsupervised clustering. Alsheikh et al.'s examination of many machine learning models (AE-LSTM, Facebook-Prophet, and Isolation Forest) for anomaly detection in photovoltaic components highlights the importance of specialized models for decision-making. Vyas et al. examine how maintenance operations impact solar power generation prediction by using machine learning to refine forecasts based on historical issues and meteorological variables. Using data gathered from 26 solar power facilities, Refaee uses supervised models for performance classification and early failure identification, with 98.85% accuracy using the J48 algorithm. In their state-of-the-art review of condition monitoring for solar systems, Berghout et al. include typical failures, conventional detection methods, and machine learning developments including deep learning and transfer learning. All things considered, these studies show how ML-based techniques are better than conventional ones, allowing for quicker and more precise remedial operations, lower operating costs, and optimized energy production.

# Methodology

**Objective**: To develop an ML-based prediction model to improve solar power plant maintenance.

**Archtectural Diagram**

## Dataset information

This project's analysis of temperature sensor and photovoltaic power generating data from an Indian solar plant is its goal. The datasets are on Kaggle and include the generated power, ambient temperature, module temperature, and irradiation.

There will be two datasets used, one from temperature sensor data from two solar plants in India and the other from photovoltaic energy generation. 34 days' worth of generation data from 2020 are included in these databases. The analysis will be facilitated and made simpler by using data from just one of the photovoltaic plants.

**Dataset names**: Plant\_1\_Generation\_Data and Plant\_1\_Weather\_Data.

The project's objective is to discover how energy generation is related to environmental factors and then create a power generation prediction model using the collected data.

### Variable Identification: Generation data

* DATE\_TIME - Date and time for each observation. Observations recorded at 15 minute intervals.
* PLANT\_ID - Plant ID number.
* INVERTER - Inverter id.
* DC\_POWER - Amount of DC power generated by the Inverter in this 15 minute interval (kW).
* AC\_POWER - Amount of AC power generated by the Inverter in this 15 minute interval (kW).
* DAILY\_YIELD - Daily yield is a cumulative sum of power generated on that day, till that point in time.
* TOTAL\_YIELD - This is the total yield for the inverter till that point in time.

### Variable Identification: Temperature and Solar Irradiation data

* DATE\_TIME - Date and time for each observation. Observations recorded at 15 minute intervals.
* Plant ID - this will be common for the entire file.
* SENSOR - Stands for the sensor panel id.
* AMBIENT\_TEMPERATURE - This is the ambient temperature at the plant.
* MODULE\_TEMPERATURE - There is a module (solar panel) attached to the sensor panel. This is the temperature reading for that module.
* IRRADIATION - Amount of irradiation for the 15 minute interval.

The dependent variable for all regression models is DC\_POWER, which measures the direct current output generated by the solar panels every 15 minutes. DC\_POWER was selected because it reflects the system's raw energy generation performance, making it a more direct and sensitive indicator of anomalies caused by environmental factors.

### Merging Dataframes

Merging the Plant\_1\_Generation\_Data and Plant\_1\_Weather\_Data using the 'DATE\_TIME' column as the key. This results in a new DataFrame called power\_sensor that contains all columns from both DataFrames, but combined based on the timestamp.

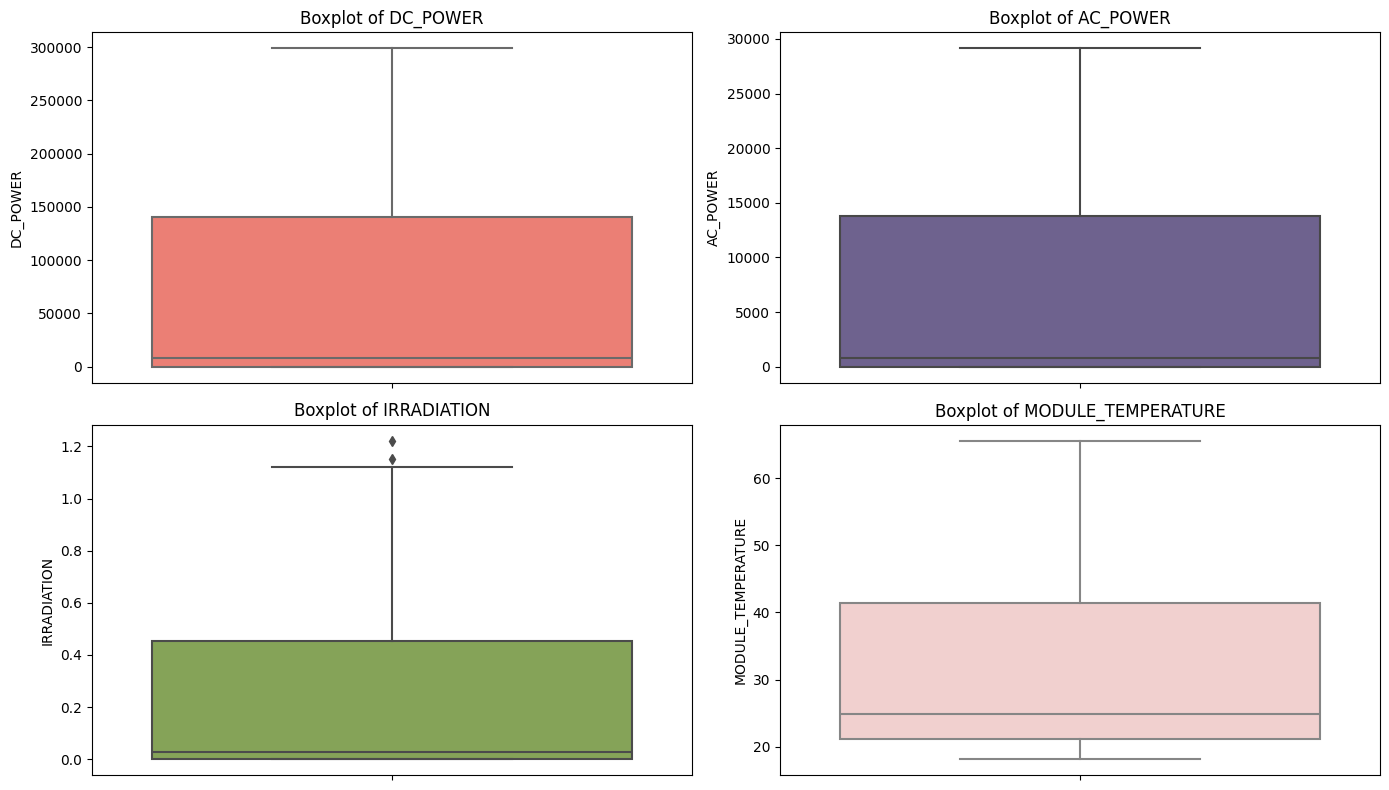
### DataFrame Information

The dataset used in this study refers to solar energy generation and associated environmental parameters, collected from sensors installed in a photovoltaic plant. The main variables analyzed include DC\_POWER, AC\_POWER, IRRADIATION, MODULE\_TEMPERATURE, among others.

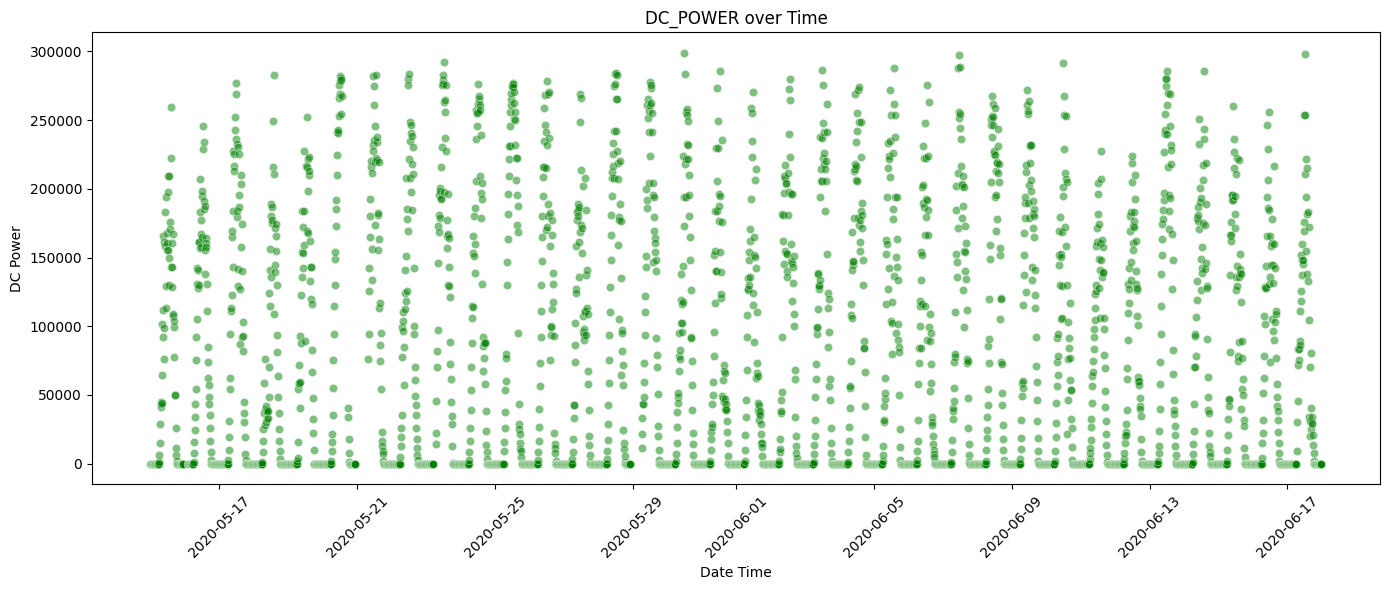
|  |  |  |  |
| --- | --- | --- | --- |
| # | Column | Non-Null Count | Dtype |
| 0 | DATE\_TIME | 3157 non-null | datetime64[ns] |
| 1 | PLANT\_ID | 3157 non-null | int64 |
| 2 | SOURCE\_KEY | 3157 non-null | object |
| 3 | AMBIENT\_TEMPERATURE | 3157 non-null | float64 |
| 4 | MODULE\_TEMPERATURE | 3157 non-null | float64 |
| 5 | IRRADIATION | 3157 non-null | float64 |
| 6 | DC\_POWER | 3157 non-null | float64 |
| 7 | AC\_POWER | 3157 non-null | float64 |
| 8 | DAILY\_YIELD | 3157 non-null | float64 |
| 9 | TOTAL\_YIELD | 3157 non-null | float64 |

### Outlier Detection

Outlier detection was performed using Boxplot graphs for the main numerical variables: DC\_POWER, AC\_POWER, IRRADIATION, and MODULE\_TEMPERATURE. This visual analysis enabled the identification of data points outside the interquartile range (IQR), which may indicate sensor failures, measurement errors, or rare events.



Additionally, a scatter plot over time was used to identify outliers distributed throughout the time series. This approach is helpful for detecting seasonal or isolated anomalies that might not be evident in aggregated views.

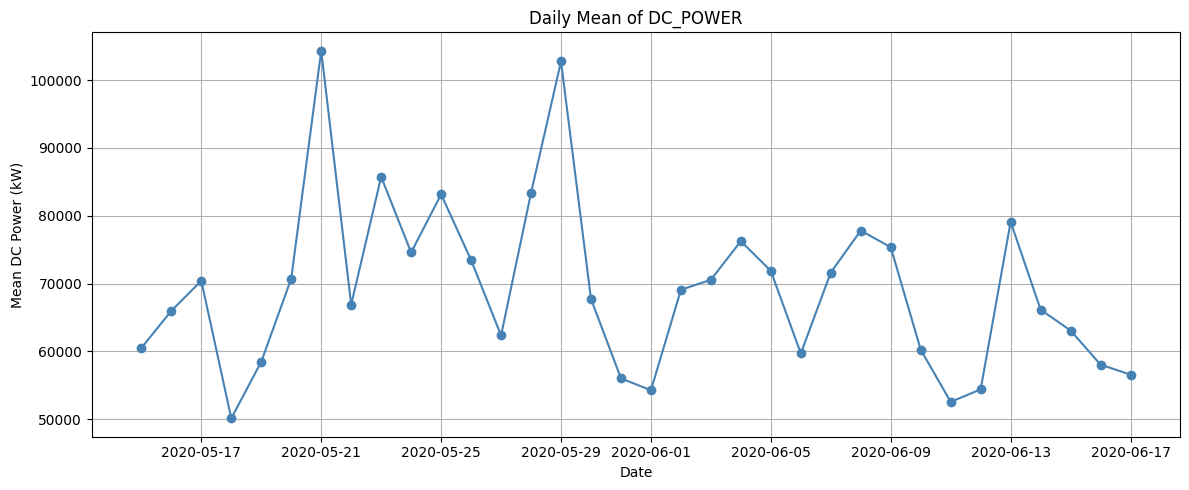


### Descriptive Statistics

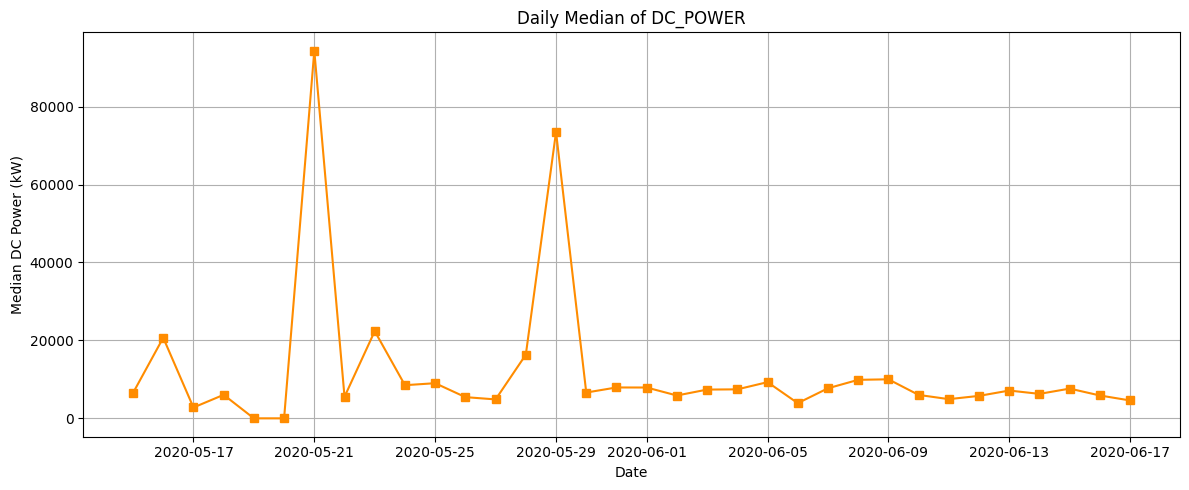
Descriptive statistics were applied to evaluate the central tendency and dispersion of the main variables in the dataset. The mean, median, minimum, and maximum values were calculated for key numeric variables, including DC\_POWER, AC\_POWER, IRRADIATION, AMBIENT\_TEMPERATURE, and MODULE\_TEMPERATURE.

Additionally, line plots were used to visualize the temporal evolution of:

* The daily mean of DC\_POWER, highlighting overall generation patterns and potential seasonal trends.



* The daily median of DC\_POWER, which helps reduce the influence of outliers and better reflect the typical performance on each day.



### Unique Values in the Dataset

An analysis of unique values per column revealed the following:

|  |  |
| --- | --- |
| Variable | Unique Values |
| DATE\_TIME | 3,157 |
| PLANT\_ID | 1 |
| SOURCE\_KEY | 1 |
| AMBIENT\_TEMPERATURE | 3,157 |
| MODULE\_TEMPERATURE | 3,157 |
| IRRADIATION | 1,755 |
| DC\_POWER | 1,688 |
| AC\_POWER | 1,688 |
| DAILY\_YIELD | 1,777 |
| TOTAL\_YIELD | 1,734 |
| DATE | 34 |

# Data analysis and Preprocessing

## Preprocessing

### Variable Reduction

In the preprocessing step, variables with constant or near-constant values were identified. Specifically, PLANT\_ID and SOURCE\_KEY had only one unique value each, meaning they do not contribute meaningful variance to the model. As such, these columns were dropped from the dataset to reduce dimensionality and improve model efficiency.

### Automatic Feature Selection

The most relevant features for predicting DC\_POWER were selected using the SelectKBest method with the f\_regression statistical test. This test evaluates the linear correlation between each independent feature and the target variable, assigning a score that reflects the predictive importance of each feature.

|  |  |  |
| --- | --- | --- |
| Index | Feature | Score (f\_regression) |
| 3 | AC\_POWER | 562,372,700 |
| 2 | IRRADIATION | 387,838.4 |
| 1 | MODULE\_TEMPERATURE | 38,035.4 |
| 0 | AMBIENT\_TEMPERATURE | 3,509.7 |
| 5 | TOTAL\_YIELD | 36.1 |
| 4 | DAILY\_YIELD | 27.2 |
| 19 | DATE\_2020-05-29 | 10.9 |

### Key Insights from Feature Scores

AC\_POWER (562 million): Extremely correlated with DC\_POWER, as expected, since both represent electrical power at different stages of the system. However, this strong correlation may lead to collinearity, making it unsuitable as a predictive feature when DC\_POWER is the target.

IRRADIATION (387 thousand): Highly important, as solar radiation directly influences energy production in photovoltaic modules.

MODULE\_TEMPERATURE (38 thousand): Also very relevant. Higher module temperatures typically reduce energy efficiency, which aligns with technical expectations.

AMBIENT\_TEMPERATURE: Potentially correlated with MODULE\_TEMPERATURE, but less directly impactful on solar panel performance.

TOTAL\_YIELD and DAILY\_YIELD: These are cumulative metrics. Individually, they may offer limited predictive power, but could contribute value when combined with other variables in more complex modeling approaches.

## Exploratory Data Analysis (EDA)

* Correlation analysis (heatmaps, scatter plots).
* Time series visualization of AC/DC power, irradiation, temperatures.

In this project, we adopt the **Time Series Split** strategy for model validation due to the temporal nature of the data. Traditional random train/test splits are not suitable for time series data, as they risk introducing data leakage by allowing future information to influence model training. To ensure the integrity of the evaluation and simulate real-world forecasting scenarios, we apply a time-aware validation strategy that respects the chronological order of observations.

The Time Series Split divides the dataset into multiple sequential train/test sets, where each training set includes all data prior to the test set. This approach mimics the process of making predictions on unseen future data based only on past observations.

By applying this method across all machine learning models—Linear Regression, Random Forest, and ARIMA—we ensure fair comparison and realistic performance estimates, particularly in the context of forecasting photovoltaic energy generation.

## Model Training

### 3.3.1 Linear Regression

### 3.3.2 Random Forest

### 3.3.3 ARIMA

## 3.4 Model evaluation

Metrics:

* Coefficient of Determination (R²).
* Mean Absolute Error (MAE).
* Mean Squared Error (MSE).
* Root Mean Squared Error (RMSE).

Model Comparison:

* Compare models using the above metrics.
* Discuss overfitting signs, if any.

# Result and Discussion

## 4.1 Results from each model

Table with all models' performance (R², MAE, MSE).

Line plots: predicted vs actual energy generation.

## 4.2 Classification from ML algorithms

## 4.3 Regression Result from algorithms

Detailed discussion of each model’s results.

Identify best-performing model and factors influencing the accuracy (e.g., impact of irradiation vs temperature).

Paragraph (Discuss your result and compare with research hypothesis)

# Conclusion

Research Problem Recap: Lack of efficient monitoring leading to decreased generation and increased costs.

Aim and Solution: Develop ML-based forecasting and anomaly detection models.

Results: Neural Networks/Random Forest performed best (depending on your actual results).

Contribution: Showed that predictive maintenance using simple ML models significantly helps in optimizing energy production.

Future Work:

* Extend dataset with multi-seasonal data.
* Integrate real-time meteorological forecasting.
* Explore deep learning architectures like LSTM for sequential data.