A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Felipe Luiz Bauer Steudel |
| *Student Number* | 2024153 |
| *Module Title* | Strategic Thinking |
| *Assessment Title* | CA3 |
| *Assessment Due Date* | 18th May 2025 |
| *Date of Submission* | 18th May 2025 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Sumário

[Introduction 4](#_Toc198310805)

[Motivation 4](#_Toc198310806)

[Research Problem 5](#_Toc198310807)

[Research Questions 5](#_Toc198310808)

[Research Hypothesis (Ho, Ha) 5](#_Toc198310809)

[Research Objectives 5](#_Toc198310810)

[Literature Review / Related works 6](#_Toc198310811)

[Technologies you are using 6](#_Toc198310812)

[Summary of Related Work 6](#_Toc198310813)

[Methodology 7](#_Toc198310814)

[Dataset information 7](#_Toc198310815)

[Variable Identification: Generation data 8](#_Toc198310816)

[Variable Identification: Temperature and Solar Irradiation data 8](#_Toc198310817)

[Merging Dataframes 8](#_Toc198310818)

[DataFrame Information 8](#_Toc198310819)

[Outlier Detection 9](#_Toc198310820)

[Descriptive Statistics 10](#_Toc198310821)

[Unique Values in the Dataset 11](#_Toc198310822)

[Data analysis and Preprocessing 11](#_Toc198310823)

[Preprocessing 11](#_Toc198310824)

[Variable Reduction 11](#_Toc198310825)

[Automatic Feature Selection 11](#_Toc198310826)

[Key Insights from Feature Scores 12](#_Toc198310827)

[Exploratory Data Analysis (EDA) 12](#_Toc198310828)

[Scatter Plot and Mean DC Power 12](#_Toc198310829)

[Correlation Heatmap 13](#_Toc198310830)

[Pair Plot of Key Variables 14](#_Toc198310831)

[DC Power vs. Irradiation 14](#_Toc198310832)

[AC\_POWER vs. DC\_POWER Over Time 15](#_Toc198310833)

[Histograms 16](#_Toc198310834)

[Scatter Plot and Mean Daily Yield 17](#_Toc198310835)

[Scatter Plot and Mean Module Temperature 17](#_Toc198310836)

[Model Training 17](#_Toc198310837)

[Data Overview 17](#_Toc198310838)

[Preserving Temporal Order 18](#_Toc198310839)

[Time Series Split 18](#_Toc198310840)

[Feature Considerations 18](#_Toc198310841)

[Cross-Validation Results 18](#_Toc198310842)

[Linear Regression and Random Forest 18](#_Toc198310843)

[ARIMA Modeling 19](#_Toc198310844)

[Model Evaluation and Discussion 19](#_Toc198310845)

[Comparative Summary of Model Performance 19](#_Toc198310846)

[Overfitting Discussion 20](#_Toc198310847)

[Final Remarks 20](#_Toc198310848)

[Result and Discussion 20](#_Toc198310849)

[Actual vs. Predicted Values 20](#_Toc198310850)

[Linear Regression – Actual vs. Predicted Plot 20](#_Toc198310851)

[Random Forest – Actual vs. Predicted Plot 21](#_Toc198310852)

[ARIMA – Actual vs. Predicted Plot 22](#_Toc198310853)

[Conclusion 22](#_Toc198310854)

[Research Problem Recap 22](#_Toc198310855)

[Aim and Solution 22](#_Toc198310856)

[Results 22](#_Toc198310857)

[Contribution 22](#_Toc198310858)

[Future Work 23](#_Toc198310859)

[References 24](#_Toc198310860)

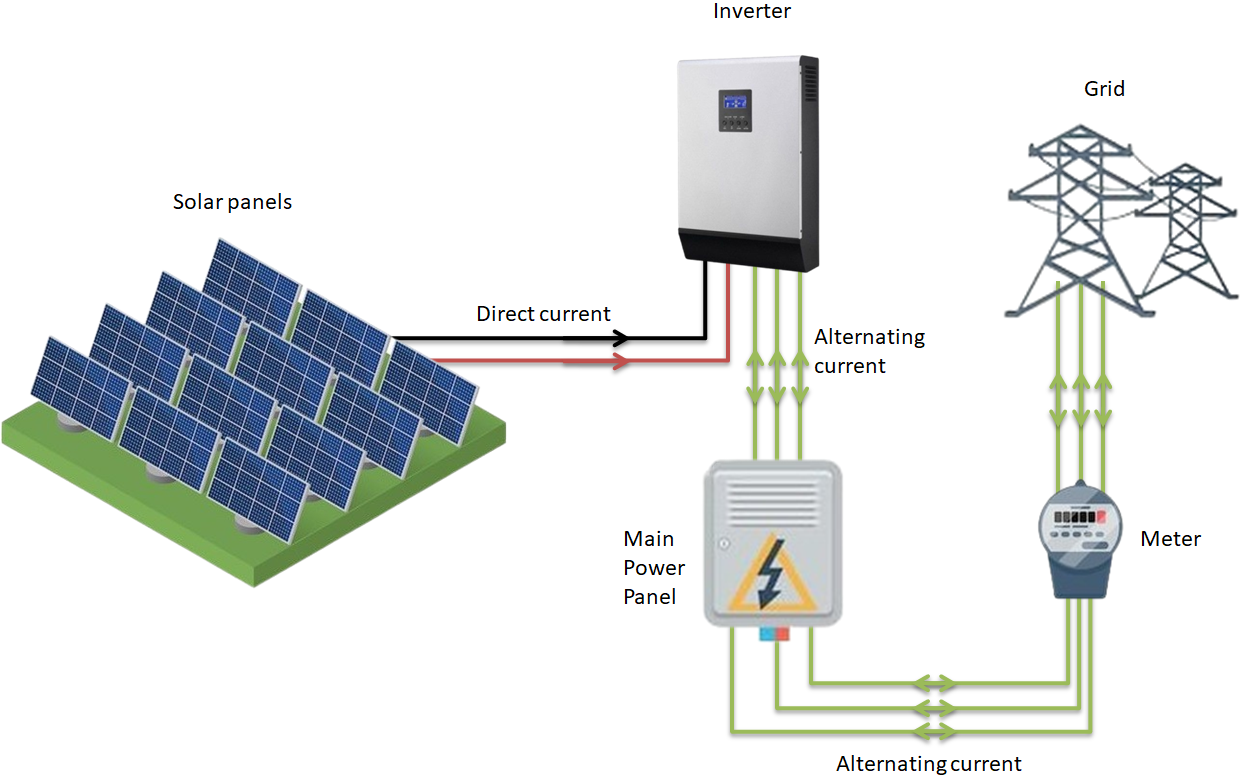
Predictive Analysis in Solar Energy using Machine Learning

# 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.

**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 and statsmodels)

## 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.

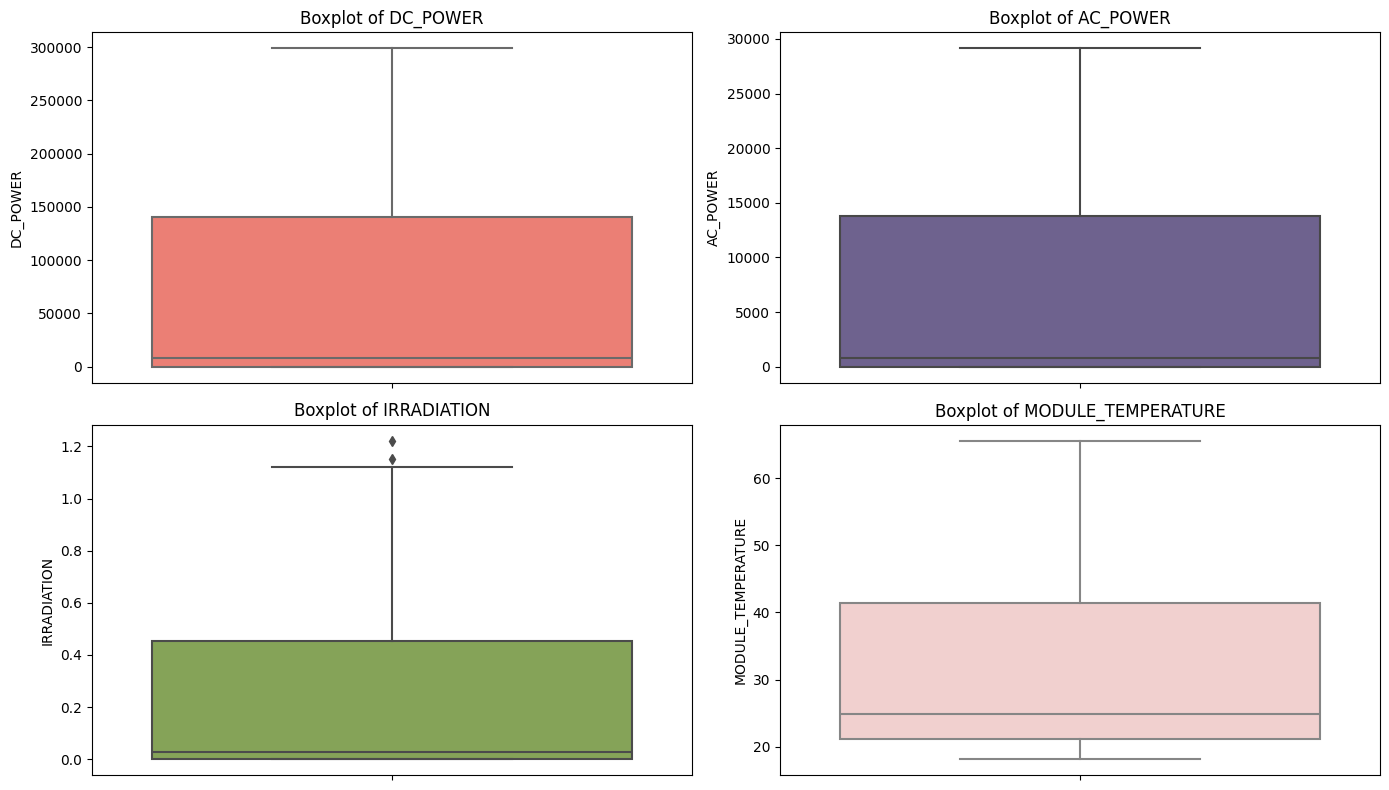
## DataFrame Information

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.

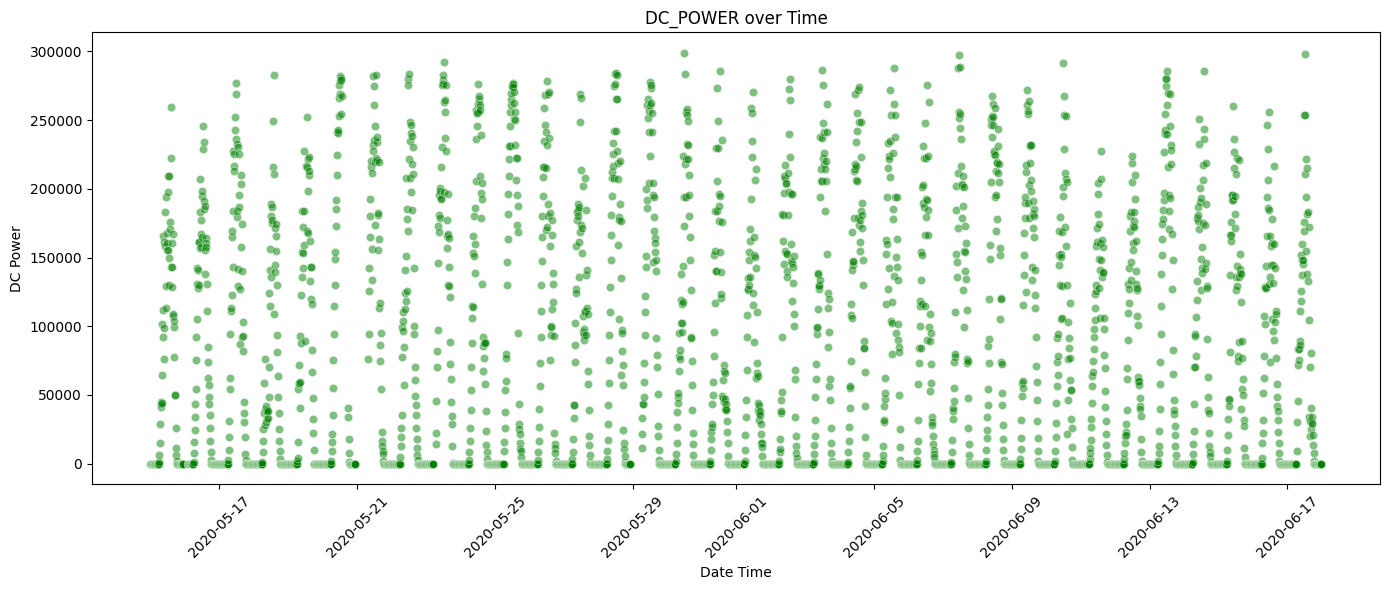
|  |  |  |  |
| --- | --- | --- | --- |
| # | Column | Non-Null Count | Dtype |
| 0 | DATE\_TIME | 3157 non-null | datetime64[ns] |
| 1 | PLANT\_ID | 3157 non-null | int64 |
| 2 | SOURCE\_KEY | 3157 non-null | object |
| 3 | AMBIENT\_TEMPERATURE | 3157 non-null | float64 |
| 4 | MODULE\_TEMPERATURE | 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.

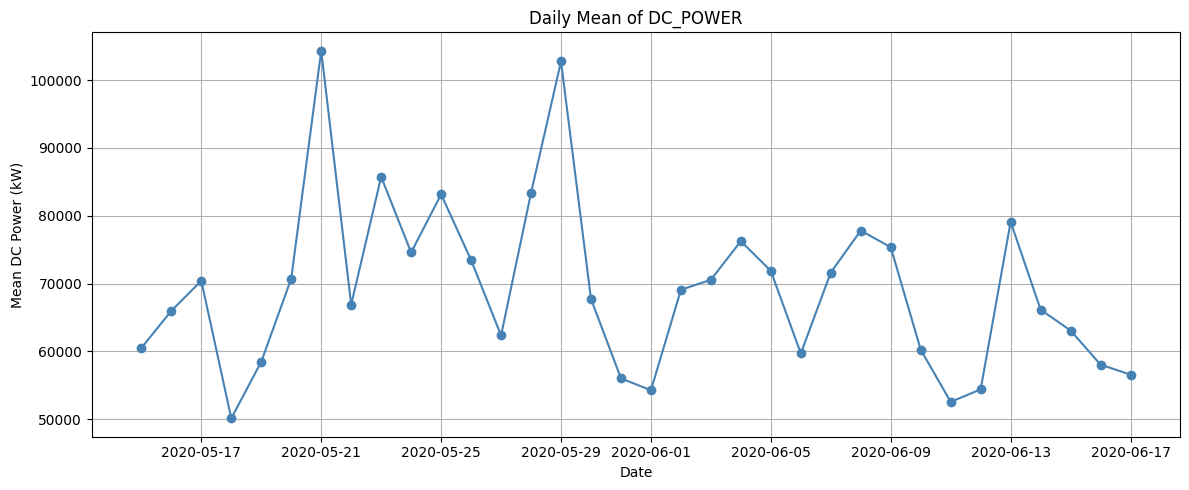


## Descriptive Statistics

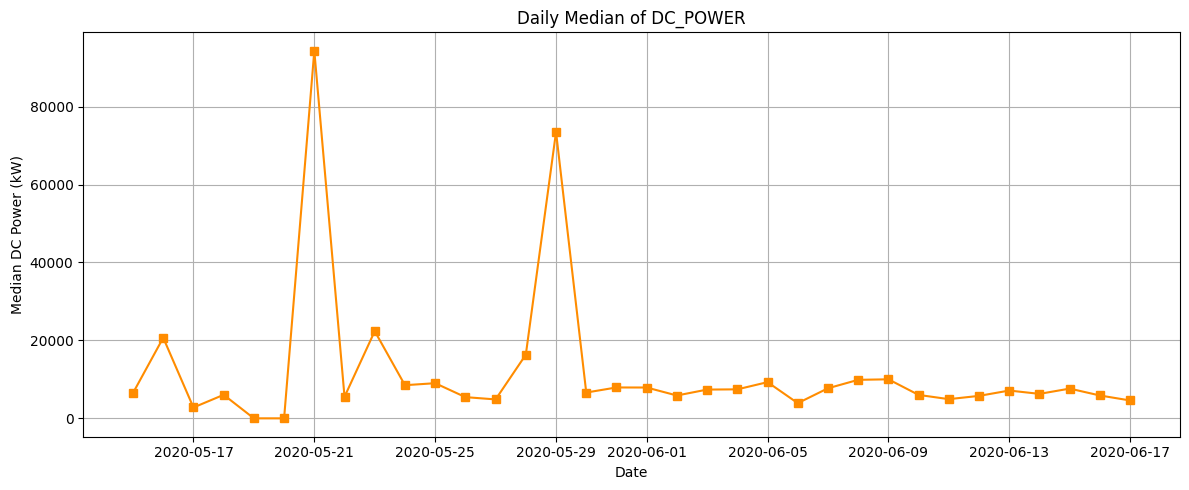
Descriptive statistics were applied to evaluate the central tendency and dispersion of the main variables in the dataset. The mean, median, minimum, and maximum values were calculated for key numeric variables, including DC\_POWER, AC\_POWER, IRRADIATION, AMBIENT\_TEMPERATURE, and MODULE\_TEMPERATURE.

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.

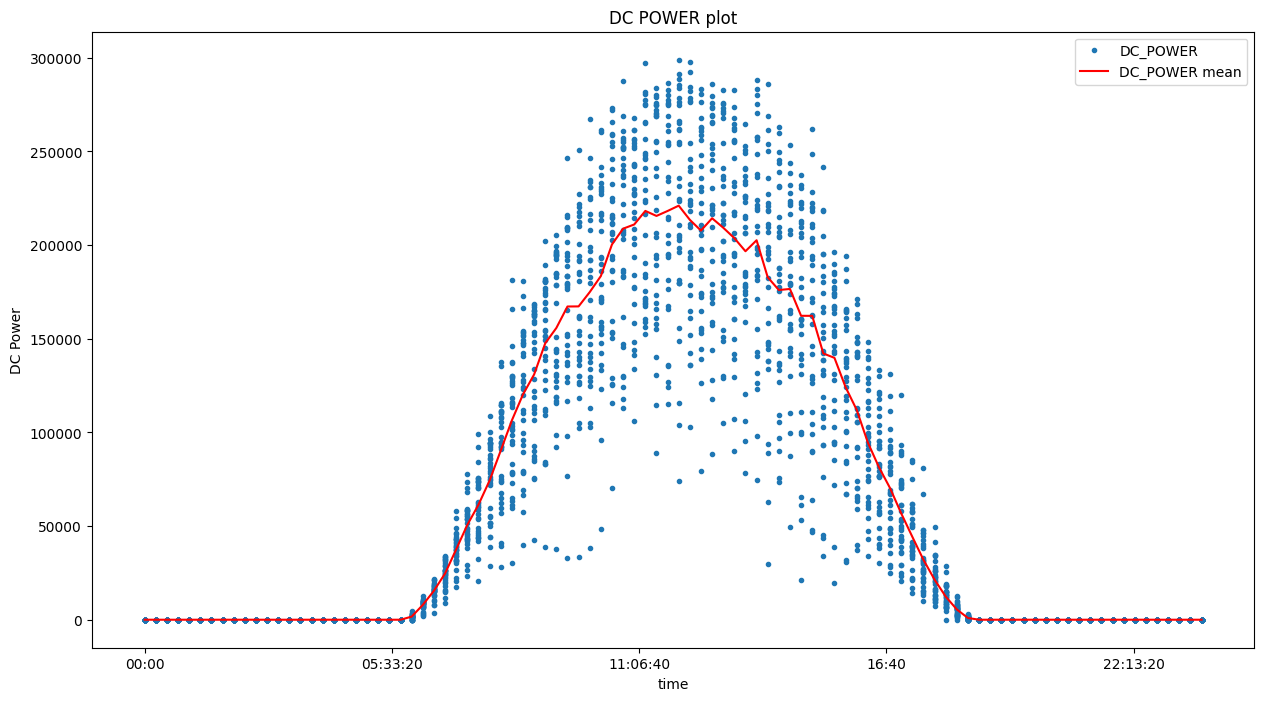
AMBIENT\_TEMPERATURE: Potentially correlated with MODULE\_TEMPERATURE, but less directly impactful on solar panel performance.

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)

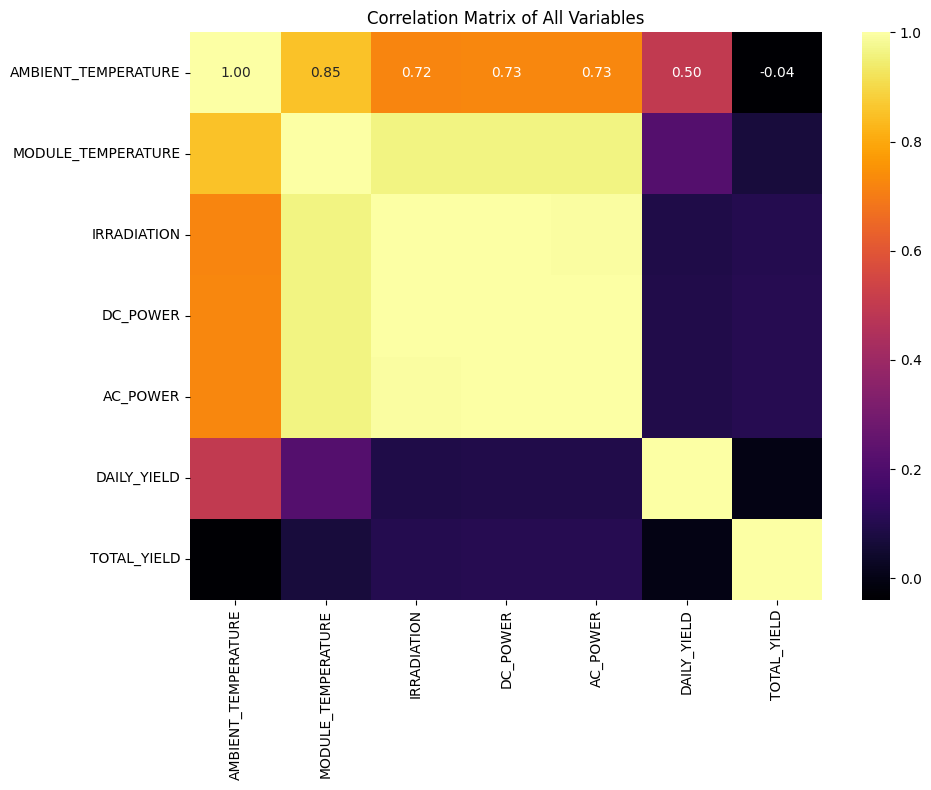
### Scatter Plot and Mean DC Power

A scatter plot was created to display DC\_POWER values across different hours of the day. This visualization is useful for identifying daily variability, unexpected peaks, anomalies, or potential operational failures. Additionally, a line representing the mean DC\_POWER grouped by hour was overlaid, allowing the identification of daily seasonality patterns and overall production trends.



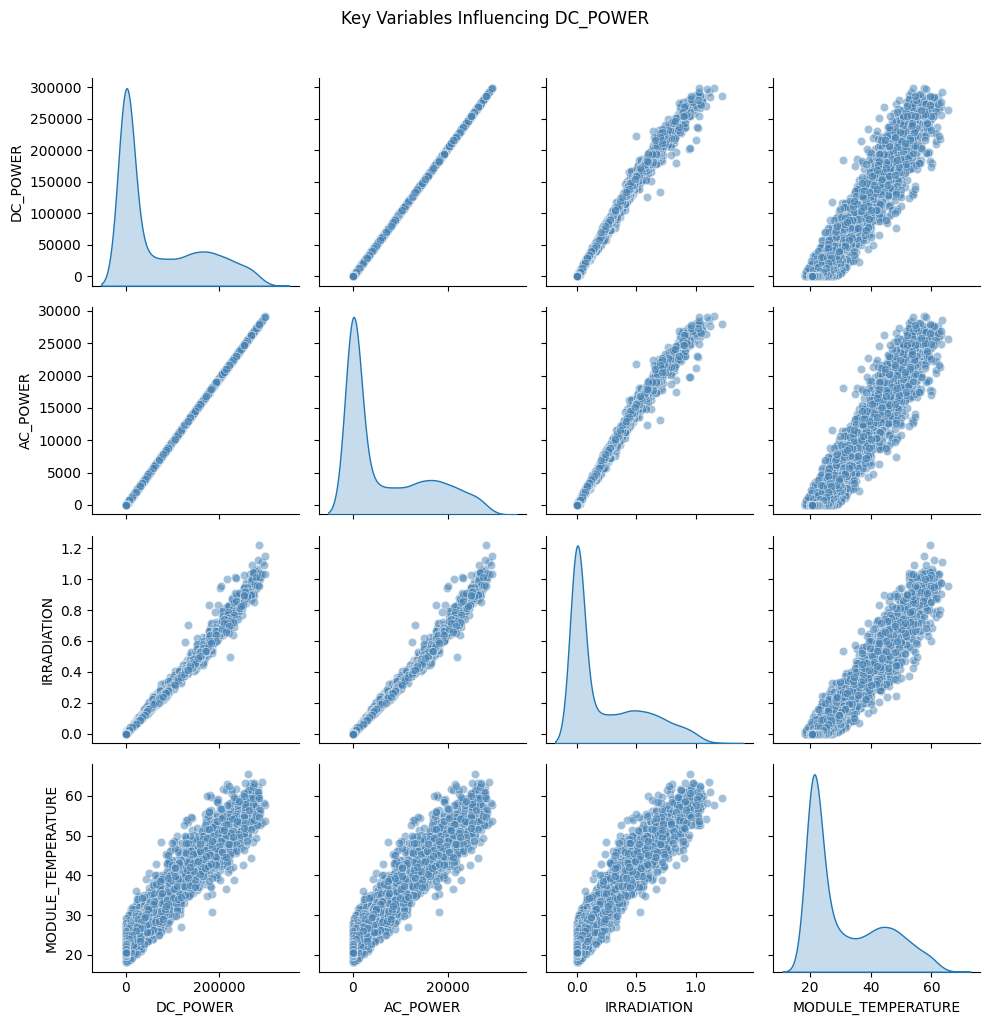
### Correlation Heatmap

A correlation matrix heatmap was used to investigate the relationships between variables in the dataset. This technique highlights which variables are more strongly correlated with DC\_POWER, assisting in feature selection and understanding variable interactions.



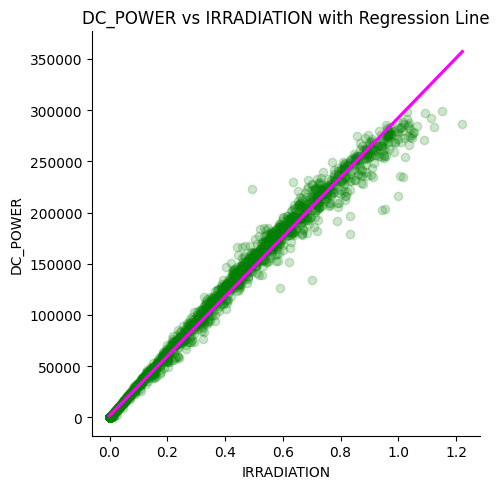
### Pair Plot of Key Variables

A scatter matrix (pair plot) was generated to explore bivariate relationships between key variables. This visual aid provides insight into linear or nonlinear relationships, clusters, and outliers, enhancing the understanding of how variables behave in combination.



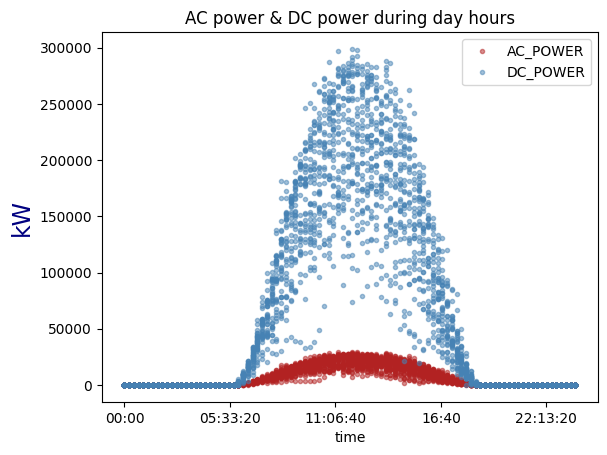
### DC Power vs. Irradiation

A dedicated scatter plot was created to examine the direct relationship between DC\_POWER and IRRADIATION, since solar radiation is a fundamental driver of energy production. A positive correlation is expected and visually confirmed.



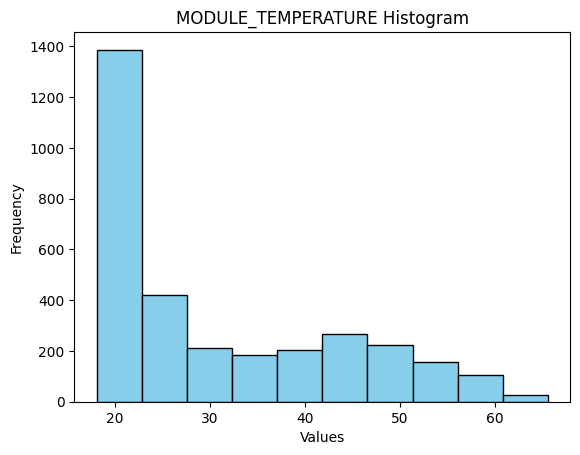
### AC\_POWER vs. DC\_POWER Over Time

A comparison of AC\_POWER and DC\_POWER values across the hours of the day was plotted. This helps analyze conversion efficiency, identify losses, and assess whether both types of power follow similar patterns during generation hours.

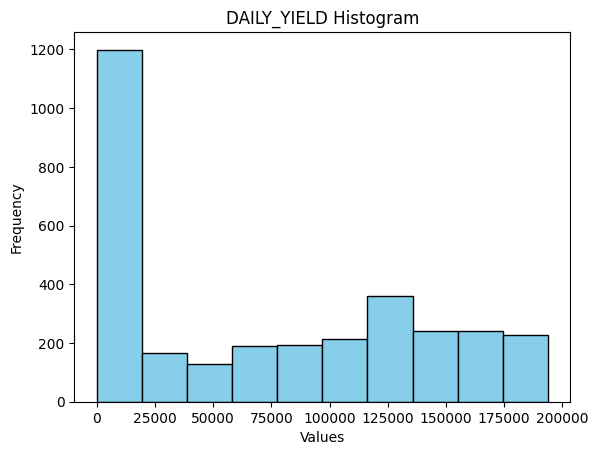


### Histograms

**Module Temperature:** The distribution of module temperatures was plotted to assess the range and frequency of values, detect abnormal readings, and understand environmental conditions during energy generation.

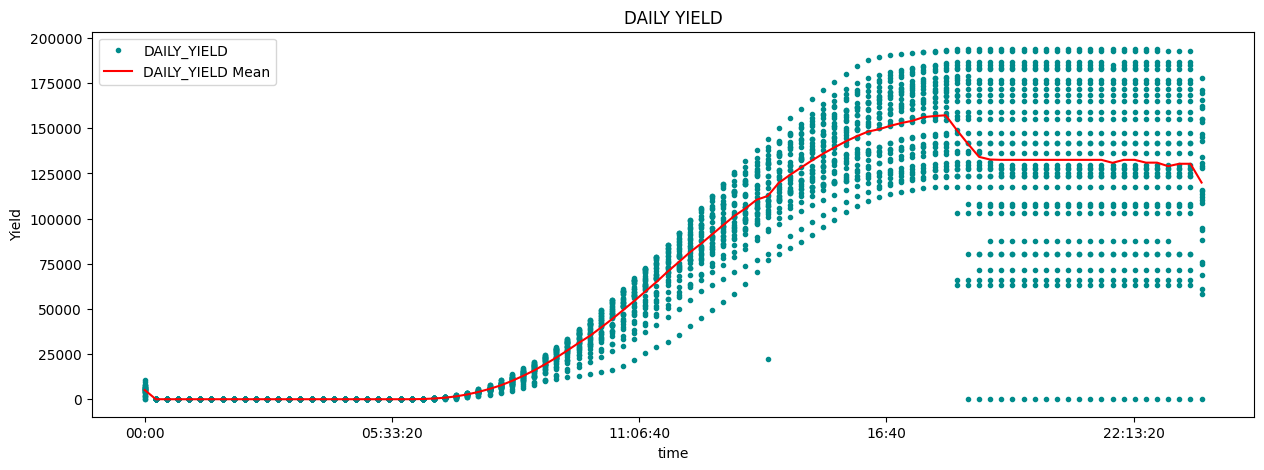


**Daily Yield:** A histogram of DAILY\_YIELD was produced to understand its spread and identify potential anomalies or consistency in daily energy output.



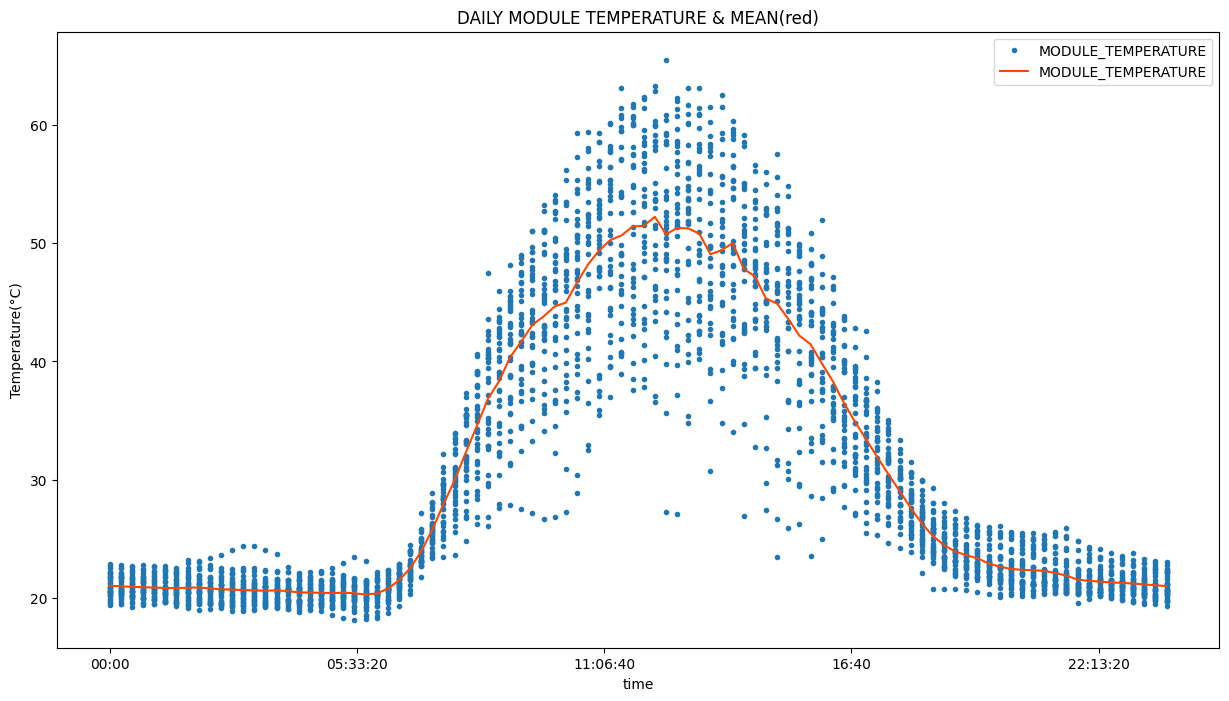
### Scatter Plot and Mean Daily Yield

A scatter plot of DAILY\_YIELD against time was used to observe trends, growth, or variability in daily production. The addition of a line showing the mean daily yield aids in recognizing long-term trends or seasonality effects.



### Scatter Plot and Mean Module Temperature

This visualization displays MODULE\_TEMPERATURE values throughout the day, highlighting variability, peaks, and potential sensor malfunctions. A line showing the mean temperature per hour was included to clarify daily thermal patterns and operational conditions for the solar modules.



# Model Training

## Data Overview

The dataset used for model training is stored in the power\_sensor DataFrame, which contains merged information on power generation and temperature readings from a solar power plant. This data serves as the foundation for all machine learning models developed in this project.

## Preserving Temporal Order

For machine learning models such as Linear Regression and Random Forest, it is essential to maintain the chronological order of the data due to its time series nature. Standard train/test splits that randomly shuffle data would compromise the integrity of model evaluation by introducing data leakage, where future information could inadvertently inform past predictions.

## Time Series Split

To respect the temporal dependencies, a Time Series Split strategy was employed. This method divides the dataset into multiple sequential train/test sets. In each fold, the training set includes all data that precedes the test set, simulating real-world forecasting where future data is never seen during training.

This validation strategy was applied consistently across all three modeling approaches:

* **Linear Regression**
* **Random Forest**
* **ARIMA**

The goal is to provide a fair and realistic comparison of model performance, particularly in forecasting DC\_POWER for photovoltaic systems.

### Feature Considerations

The variable AC\_POWER was deliberately excluded from the feature set when predicting DC\_POWER, as both are highly correlated and represent different stages of the same process. Including AC\_POWER would risk data leakage. Instead, only environmental features (e.g., irradiation, ambient and module temperature) were used to inform the models.

## Cross-Validation Results

### Linear Regression and Random Forest

|  |  |  |
| --- | --- | --- |
| Fold | Linear Regression R² | Linear Regression RMSE |
| 1 | 0.9911 | 9291.01 |
| 2 | 0.9915 | 8203.70 |
| 3 | 0.9915 | 7993.96 |
| 4 | 0.9865 | 9910.21 |
| 5 | 0.9921 | 7406.09 |

Linear Regression consistently achieved high R² values (~0.99) and relatively low RMSE across all folds. Although it performs well, its linear assumptions limit its ability to capture complex patterns.

|  |  |  |
| --- | --- | --- |
| Fold | Random Forest R² | Random Forest RMSE |
| 1 | 0.9940 | 7657.51 |
| 2 | 0.9932 | 7365.99 |
| 3 | 0.9934 | 7025.59 |
| 4 | 0.9906 | 8260.53 |
| 5 | 0.9953 | 5723.81 |

Random Forest outperformed Linear Regression on every fold, achieving higher R² (up to 0.9953) and lower RMSE, especially in fold 5. Its ability to model non-linear relationships and variable interactions makes it the strongest performer among the ML models.

## ARIMA Modeling

An ARIMA(2,0,3) model was trained using only the DC\_POWER time series, without additional exogenous variables.

**ARIMA RMSE:** 74,770.96

This result indicates that the ARIMA model performed significantly worse than the machine learning models. The likely reason is that ARIMA, in this configuration, does not incorporate external environmental variables such as irradiation or temperature, which are crucial to understanding energy generation patterns. Additionally, ARIMA may not adequately capture complex, nonlinear relationships found in the data.Model evaluation

## Model Evaluation and Discussion

### Comparative Summary of Model Performance

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Average R² | Average RMSE | Strengths | Weaknesses |
| Random Forest | ~0.9933 | ~7206.29 | Captures non-linear relationships, low RMSE, high accuracy | Higher computational cost, harder to interpret |
| Linear Regression | ~0.9905 | ~8561.39 | Simple, interpretable, fast training | Limited in modeling complex or non-linear patterns |
| ARIMA (2,0,3) | — | 74,770.96 | Captures temporal structure in single-variable series | Very poor accuracy in this case, no exogenous variables used |

Random Forest is clearly the best-performing model in terms of both predictive power (R²) and error minimization (RMSE). It consistently performs well across all folds, capturing complex interactions between environmental variables and energy generation.

Linear Regression, while simpler and easier to interpret, performs only slightly worse, making it a viable alternative for less complex deployment scenarios.

ARIMA drastically underperforms, suggesting that it is not suitable in this context, likely due to its reliance on a single variable and lack of support for exogenous factors.

### Overfitting Discussion

**Random Forest:**

* No clear signs of overfitting were observed across the folds. The model maintained consistently high R² values and low RMSE on all validation sets.
* This stability suggests strong generalization performance.
* However, given Random Forest's flexibility, there’s always a risk of overfitting if the number of trees or depth is too high, especially on small or noisy datasets. These risks were mitigated by using Time Series Cross-Validation and not training on future data.

**Linear Regression:**

* Overfitting is not a concern for this model due to its linear nature and fewer degrees of freedom.
* The slightly lower performance is expected and acceptable, especially considering the model's simplicity.

**ARIMA:**

* ARIMA is unlikely to overfit in its current configuration since it doesn’t model complex relationships, but its underfitting is evident — it fails to capture the influence of external variables, resulting in poor accuracy.

## Final Remarks

* Random Forest balances powerful modeling capability with robust validation performance, making it the most suitable model for forecasting DC\_POWER based on environmental variables.
* Linear Regression can be considered in scenarios where interpretability or computational cost are key concerns.
* ARIMA is not recommended unless exogenous variables are incorporated (e.g., ARIMAX), or if the forecasting task is reformulated to focus on univariate time series prediction.

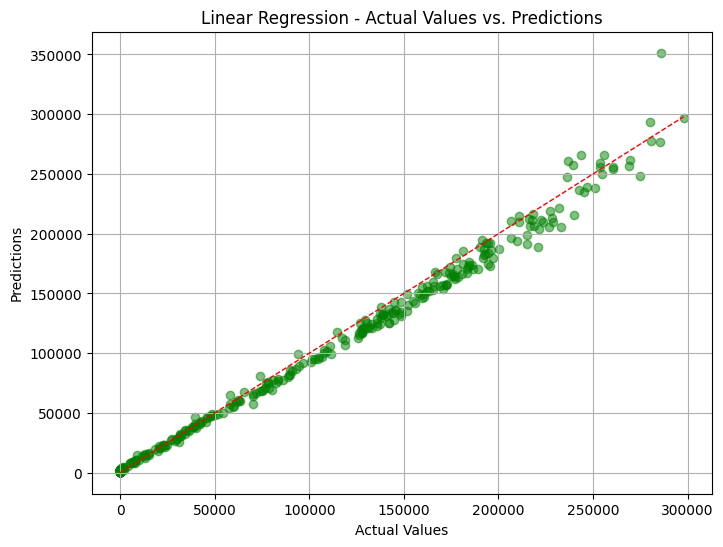
## Result and Discussion

### Actual vs. Predicted Values

To visually assess model performance, Actual vs. Predicted plots were generated for each model. These plots allow us to evaluate how closely the predicted values align with the actual values of DC\_POWER.

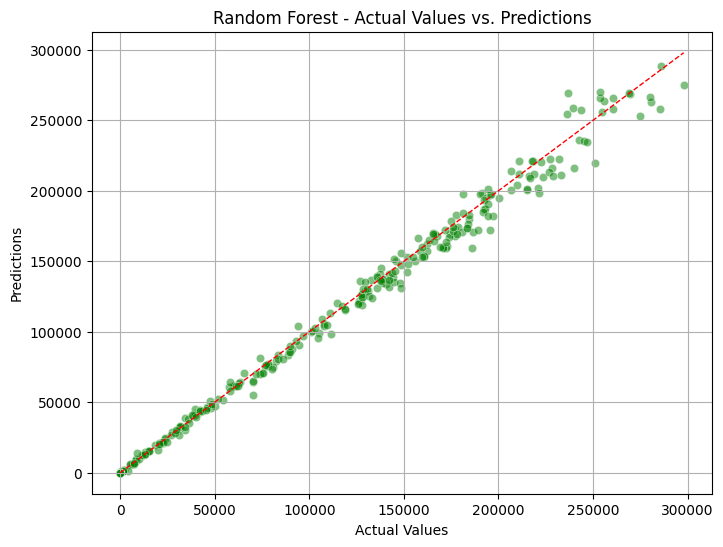
### Linear Regression – Actual vs. Predicted Plot

The scatter plot for Linear Regression shows a strong linear alignment along the diagonal, indicating good prediction accuracy. However, some minor deviations from the diagonal suggest limitations in capturing more complex patterns or non-linear behavior in the data.



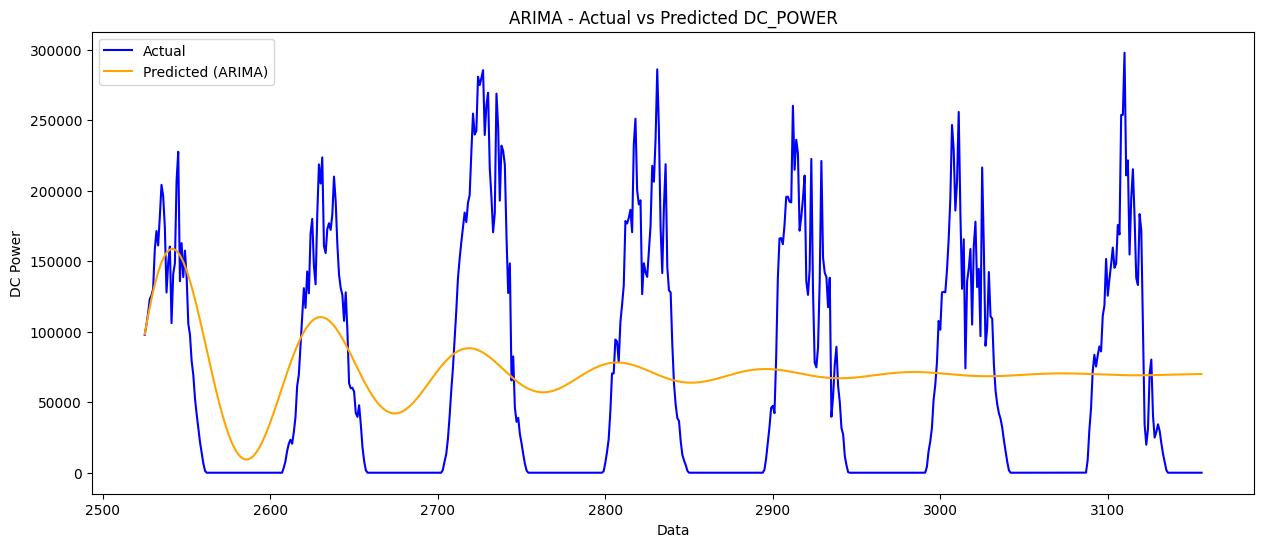
### Random Forest – Actual vs. Predicted Plot

The Random Forest model displays excellent alignment between predicted and actual values, with points tightly clustered along the diagonal.



### ARIMA – Actual vs. Predicted Plot

The ARIMA model’s plot exhibits significant scatter and deviation from the ideal diagonal line, indicating poor prediction performance.



# Conclusion

## Research Problem Recap

The primary challenge addressed in this study was the lack of efficient monitoring and forecasting in solar power generation systems. This issue often results in decreased energy output and increased operational costs, primarily due to undetected anomalies and suboptimal performance conditions.

## Aim and Solution

The goal of this project was to develop machine learning-based models capable of accurately forecasting solar energy generation and detecting performance anomalies. These models aim to support predictive maintenance and enable data-driven decision-making to improve operational efficiency.

## Results

Among the models evaluated, Random Forest consistently delivered the best performance, with high R² values and low RMSE across all validation folds. While Linear Regression also performed well, it showed limitations in capturing non-linear relationships. The ARIMA model, although commonly used in time series analysis, was less effective due to the absence of exogenous variables and its univariate nature.

## Contribution

This study demonstrates that predictive maintenance using relatively simple ML models (e.g., Random Forest) can significantly enhance energy production efficiency. These models can be integrated into solar monitoring systems to provide real-time insights, reduce downtime, and optimize energy yield.

## Future Work

* To further improve the robustness and accuracy of forecasting models, the following directions are proposed:
* Extend the dataset to include multi-seasonal and long-term operational data.
* Integrate real-time meteorological forecasts to enhance environmental context in predictions.
* Explore deep learning architectures, such as LSTM (Long Short-Term Memory) networks, which are well-suited for capturing sequential dependencies in time series data.

# References

Alsheikh, M.I.A., Awaysheh, F.M. and Alshehri, M.D., Machine Learning Schemes for Anomaly Detection in Solar Power Plants.

Berghout, T., Benbouzid, M., Bentrcia, T., Ma, X., Djurović, S. and Mouss, L.H., Machine Learning-Based Condition Monitoring for PV Systems: State of the Art and Future Prospects.

Betti, A., Lo Trovato, M.L., Leonardi, F.S., Leotta, G., Ruffini, F. and Lanzetta, C., Predictive Maintenance in Photovoltaic Plants with a Big Data Approach.

Iqbal, M., Decision Trees and Random Forest - Week 3, Study Material, CCT College Dublin, Dublin.

Iqbal, M., Linear Regression - Week 4. Study Material, CCT College Dublin, Dublin.

Iqbal, M., ARIMA and SARIMA Time Series Models - Week 7. Study Material, CCT College Dublin, Dublin.

McQuaid,D., Colour Theory, CCT College Dublin, Dublin.

McQuaid, D., Data Visualization Python Tutorial, CCT College Dublin, Dublin.

McQuaid, D., How to Deal with Missing Data in Python. Study Material, CCT College Dublin, Dublin.

McQuaid, D., Simple tips for prettier and customised plots in Seaborn (Python), CCT College Dublin, Dublin.

McQuaid, D., What is Exploratory Data Analysis. Study Material, CCT College Dublin, Dublin.

Refaee, E.A., Using Machine Learning for Performance Classification and Early Fault Detection in Solar Systems. Edited by M.F. Nadeem.

Steudel, F., 2024. Project Proposal CA1. Unpublished coursework for the module *Strategic Thinking*.. CCT College Dublin, Higher Diploma in Data Analytics for Business.

Steudel, F., 2024. Project Proposal CA2. Unpublished coursework for the module *Strategic Thinking*.. CCT College Dublin, Higher Diploma in Data Analytics for Business.

Vyas, S., Goyal, Y., Bhatt, N., Bhuwania, S., Patel, H., Mishra, S. and Tripathi, B., Forecasting Solar Power Generation on the Basis of Predictive and Corrective Maintenance Activities.